International Journal of Computational Geometry & Applications
© World Scientific Publishing Company

Measuring the Error in Approximating the Sub-level Set Topology of Sampled Scalar Data

KENES BEKETAYEV
Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA
National Laboratory Astana, 53 Kabanbay Batyr Ave, Astana, 010000, Kazakhstan

DAMIR YELISSIZOV
Department of Mathematics, University of California, 405 Hilgard Ave, Los Angeles, CA 90095, USA
Kazakh-British Technical University, 59 Tole Bi St, Almaty, 050000, Kazakhstan

DMITRIY MOROZOV
Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA

GUNTHER H. WEBER
Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA

BERND HAMANN
Department of Computer Science, University of California, 1 Shields Ave, Davis, California 95616, USA

Received May 10, 2017
Revised Oct 16, 2017
Communicated by Kenes Beketayev (kenes.b@gmail.com)

ABSTRACT

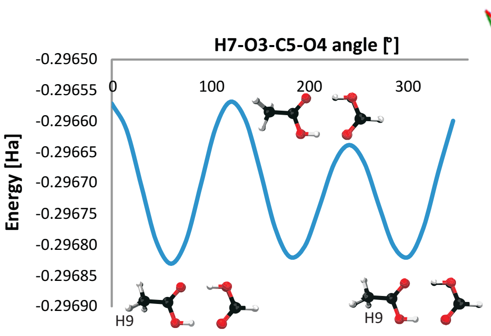
This paper studies the influence of the definition of neighborhoods and methods used for creating point connectivity on topological analysis of scalar functions. It is assumed that a scalar function is known only at a finite set of points with associated function values. In order to utilize topological approaches to analyze the scalar-valued point set, it is necessary to choose point neighborhoods and, usually, point connectivity to meaningfully determine critical-point behavior for the point set. Two distances are used to measure the difference in topology when different point neighborhoods and means to define connectivity are used: (i) the bottleneck distance for persistence diagrams and (ii) the distance between merge trees. Usually, these distances define how different scalar functions are with respect to their topology. These measures, when properly adapted to point sets coupled with a definition of neighborhood and connectivity, make it possible to understand how topological characteristics depend on connectivity. Noise is another aspect considered. Five types of neighborhoods and connectivity are discussed: (i) the Delaunay triangulation; (ii) the relative neighborhood graph; (iii) the Gabriel graph; (iv) the k-nearest-neighbor (kNN) neighborhood; and (v) the Vietoris-Rips complex. It is discussed in detail how topological characterizations depend on the chosen connectivity.

Keywords: Sub-level set topology; error quantification; topological structures.

1. Introduction

Scalar functions account for a significant portion of the scientific data generated today. They usually carry information about the behavior of a system, making them a frequent target of interest in many areas of research (e.g., chemistry, physics, climate simulation).

The topological characterization of a (scalar) function is particularly important for



(a) (b)

Fig. 1: (a) Rotation of the methyl group of the molecule of dimer of formic and acetic acid produces the potential energy function, i.e., the real function. (b) Although the first approximated function deviates from the real function, it still preserves the correct number of minima — three — thus bearing no error on the count of minima. However, the second approximated function contains only one minimum, leading to an error of $2/3$.

structure-driven comparison and simplification of functions. When only point-sampled versions of functions are known, the definitions used for point connectivity/neighborhood and distances between topological structures and critical point behavior are crucial. These definitions determine how close two functions, i.e., their topological structures, are when comparing them; or, when simplifying a function's representation, these definitions determine in what sequence simplification steps must be performed to minimally destroy a function's topology. Thus, we were motivated to obtain a deeper understanding of the effects of commonly used distance measures/connectivity on topological characterizations of and algorithms applied to scalar functions in such a discrete setting.

We consider the following scenario to illustrate the problem. In chemistry, a potential energy function of a molecule is an important analytical tool. For example, in Figure 1, we rotate the methyl group in the dimer of formic and acetic acid (DFA). The minima of this function correspond to the stable states of the DFA molecule. Hence, they can be used to better understand the structure of the molecule and its evolution during chemical reactions. However, we are only able to measure or to simulate a discrete set of samples of this function. Figure 1(b) shows two sets of samples and their approximated functions; we try to extract information about the original minima.

If we are interested only in the correct number of minima (i.e., stable states), we can quantify the error as the ratio of the missing to the correct minima. The real function has three minima. Thus for the first approximated function the error would be measured as $0/3 = 0$. However, the second approximated function misses two minima, leading to the error of $2/3 = 0.67$. The latter non-zero error reflects the fact that it missed some of the topological information we are interested in, in this particular case, the number of minima. Note that we are considering the error only in terms of the number of minima, not positions

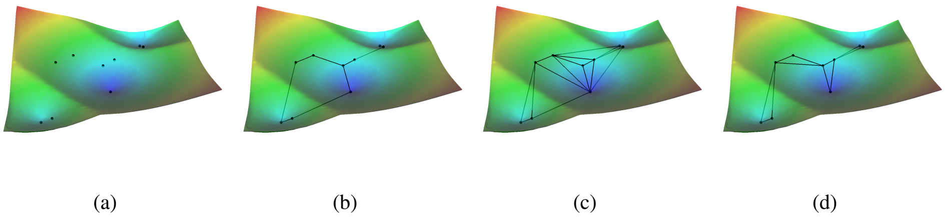


Fig. 2: (a) Given a set of 9 samples of a two-dimensional function, we can construct various meshes with potentially different approximation error. For example, following meshes are generated by (b) Gabriel graph, (c) Delaunay triangulation, (d) K-Nearest-Neighbors, with $K = 2$, lowest value for which the mesh becomes connected.

or function values.

If we look at the sources of error in this example, we note that the density and the distribution of samples have separate effects on the error. Indeed, even with a large number of samples, a bad distribution can cause a large error. On the other hand, a too small number of points would not sample the domain well enough, regardless of their distribution. We limit ourselves to uniformly random distributions of samples, hence, focusing mainly on the *sampling density error*.

Another factor that causes the error is how we connect the samples and approximate the function in between. In the previous, one-dimensional example, we intuitively connected samples along the only axis and used linear interpolation in between the samples. However, with functions on multi-dimensional domains, the strategy of how to connect samples becomes less clear. Consider a two-dimensional function in Figure 2. Given a set of samples, we can connect them into different meshes, each of which might lead to a different error. For the purpose of this work, we fix the interpolation scheme to be linear (1D interpolation along the edges of the mesh for any dimensionality) and focus on the *mesh error*.

So far, we considered the minimum as a topological feature of interest. Minima, or, more generally, extrema, were the focus in an evaluation study by Correa and Lindstrom [1], where the authors developed an error measure based upon F-measure. However, computing error in terms of F-measure requires special considerations to enable the exact matching of extrema, limiting the extend to which it can be used. Takahashi et al. [2] considered more generally the critical points, and defined a topological error criterion for them. We go beyond the discrete points and focus on topological features related to *sub-level sets* of the function, quantifying the error in their approximation. This approach leads to more comprehensive error measures, which are the main focus of this work.

Given a compact subset X of the Euclidean space R^d and a threshold c, the sub-level set of a function $f : X \to R$ is the set of all points in the domain of the function whose value does not exceed c; formally, it is the set $f^{-1}(-\infty, c] = \{(x_1, \ldots, x_d) \in X \mid f(x_1, \ldots, x_d) \leq c\}$. We consider two topological approaches to describing the evolution of sub-level sets. The first is based on a *merge tree*, which tracks the evolution of components of sub-level sets by recording their birth and merge events as the threshold value c increases. The second is a *persistence diagram*, which records the lifespan of components of sub-level sets. In particular, it records all pairs (b,d) such that a component born in $f^{-1}(-\infty,b]$ dies in $f^{-1}(-\infty,d]$ by merging with an older component.

To quantify the error for these topological structures, we need a notion of distance between a measure computed from the real function and the one from the approximation. In case of persistence diagrams, a natural choice is the *bottleneck distance*, introduced by Cohen-Steiner, Edelsbrunner, and Harer [3]. We use it as our first error measure. In case of merge trees, we choose the *distance between merge trees*, proposed by Beketayev et al. [4], and use it as the second error measure. Details on the difference between the two selected measures are discussed in Sections 2.2, 3.3.

The chosen error measures require the real function to be known, thus they cannot be used if only a set of samples are available. However, they provide a powerful evaluation capability, given both the real function and its sampling, to compare common approximation methods. In our case, we evaluate the approximation quality of different types of meshes, as we vary several parameters, e.g., the density of the sampling. Such evaluation reveals common types of problems and leads to recommendations of preferred types of meshes under various conditions.

We emphasize that the key advantage of our work, when compared to F-measure of Correa and Lindstrom [1], is that our error measures implicitly distinguish the noise in the data, resulting in robust measurements resilient to perturbations of the input. This property is crucial when working with scientific data, where noise is a serious obstacle to any analysis.

The main contributions of this paper are:

- The use of two error measures, ε_B and ε_M, where the former is the bottleneck distance between persistence diagrams and the latter is the distance between merge trees, both with implicit ability to distinguish the noise, leading to better understanding of error levels in topological structures.
- The evaluation study of common mesh constructions using ε_B and ε_M.
- Recommendations for each type of mesh based on the evaluation results.
- Detailed analysis of proposed error measures, including their behavior for real-world data.

Section 2 presents related work in scalar field topology and persistent homology that provides a background for error measures. It also reviews commonly used types of meshes that are used later in evaluation. In Section 3, we present the evaluation study of common mesh constructions using the selected error measures. The study uses three sets of functions as ground truth: parametrically generated functions, objective functions, and real-world data sets. In Section 3.2, we form recommendations for various settings based on the evaluation study results. Finally, in Section 4 we summarize our work and provide venues to explore in the future.

2. Related Work

In this section we discuss the related work and theoretical background necessary for defining the new error measures [4]. We also discuss common types of mesh construction methods that are used for the approximation of scalar functions.

2.1. *Scalar Field Topology*

Scalar field topology characterizes data by topology changes of its level sets. Given a smooth, real-valued function without degenerate critical points, level set topology changes only at isolated critical points [5]. Several structures relate critical points to each other.

The *contour tree* [6,7,8] tracks the level sets of the function by recording their births (at minima), merges or splits (at saddles), and deaths (at maxima) in a tree. It is used in a variety of high-dimensional scalar field visualization techniques [9,10,11]. Contour trees correctly encode topology of the level sets of functions on simply connected domains. The contour tree is a special case of the *Reeb graph* [12], which can correctly represent connectivity of level sets for all functions by allowing loops in the resulting graphs. Another structure, Morse-Smale complex [13,14], used, for example, to visually explore high-dimensional scalar fields [15], segments the function into the regions of uniform gradient flow and encodes geometric information.

2.2. *Persistent Homology and Merge Trees*

The concept of homology in algebraic topology offers an approach to studying the topology of the sub-level sets. We refer to the textbook by Munkres [16] for the detailed introduction to homology. Informally, it describes the cycles in a topological space: the number of components, loops, voids, and so on. In this work, we are only interested in 0–dimensional cycles, i.e., the connected components.

Persistent homology tracks changes to the connected components in sub-level sets of a scalar function. We say that a component is born in the sub-level set $f^{-1}(-\infty, b]$ if its homology class does not exist in any sub-level set $f^{-1}(-\infty, b-\varepsilon]$. This class dies in the sub-level set $f^{-1}(\infty, d]$ if its homology class merges with another class that exists in a sub-level set $f^{-1}(-\infty, b']$ with $b' < b$. When a component is born at b and dies at d, we record a pair (b,d) in the (0–dimensional) persistence diagram of the function f, denoted $D(f)$. For technical reasons, we add to $D(f)$ infinitely many copies of every point (a,a) on the diagonal.

Persistence diagrams reflect the importance of topological features of the function: the larger the difference $d-b$ of any point, the more we would have to change the function to eliminate the underlying feature. Thus, persistence diagrams let us distinguish between real features in the data and the noise.

Cohen-Steiner et al. [3] proved the stability of persistence diagrams with respect to the bottleneck distance, $d_B(D(f), D(g))$. It is defined as the infimum over all bijections, $\gamma : D(f) \to D(g)$, of the largest distance between the corresponding points,

$$d_B(D(f), D(g)) = \inf_{\gamma} \sup_{u \in D(f)} ||u - \gamma(u)||_\infty.$$

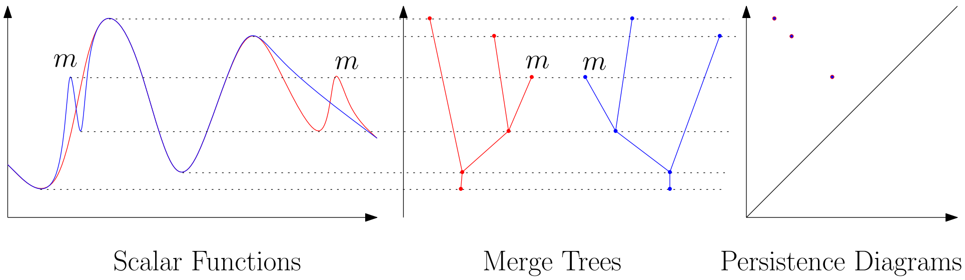


Fig. 3: The bottleneck distance between persistence diagrams fails to capture the difference in terms of the nesting of connected components of sub-level sets, while the distance between merge trees is able to quantify it.

We use the bottleneck distance d_B between the persistence diagrams of the real and the approximated functions as a basis for our ε_B error measure.

While persistence diagrams convey importance of topological features of the function, it records limited amount of information about connected components of sub-level sets. A structure called a *merge tree* is known to track the evolution of connected components of sub-/super-level sets, as they appear at minima (maxima) and merge with other connected components at merge saddles. We note that the merge tree here is related to a barrier tree [17,18] (also known as join/split tree) in scalar field topology through critical points [5], as the former tracks sub-/super-level sets, while the latter tracks level sets.

Since the goal is to quantify the topological difference between the real and the approximated functions, one can express such difference in terms of merge trees. We adopt a recently proposed distance definition between merge trees [4] as d_M, where the authors define the distance as a minimal cost of obtaining order-preserving isomorphism between merge trees with an additional consideration of the function value matching between critical points. The main idea behind such definition is that it allows to quantify the difference not only in terms of the persistence of connected components of sub-level sets, in which case persistence diagrams can be used, but also allows to take into account a difference in terms of the nesting of sub-level sets, resulting in a more precise difference quantification, see Figure 3.

We adopt the distance between merge trees as our second error measureε_M, assuming that the merge tree of the real function is the ground truth, and the merge tree of an approximated function has topological error, thus the distance between the two yields a measure of topological error.

Several other publications address the problem of quantifying topological difference. For example, Biasotti et al. [19] proposed a version of the Reeb graph that additionally encodes various geometric attributes, for which a similarity measure is defined using a graph matching algorithm. However, such measures do not explicitly reflect persistence, while both the bottleneck distance between persistence diagrams and the distance between merge trees do. Doraiswamy et al. [20] with a proposition of a saliency plot that summarizes relative importance of all topological features, and Bubenik [21] proposed the bottleneck distance for persistence landscapes, a topological descriptor that extends persistence diagrams.

Recently, the ε*-interleaving distance* was introduced by Chazal et al. [22] as yet another means to compare topological structures. Adaptions were proposed for specific structures, including merge trees [23], the Reeb graph [24], and extremum graphs and cluster trees [25,26]. However, no practically viable algorithms are known for approaches using this distance measure as its naive implementation leads to exponential complexity, which is the reason why we did not consider this distance measure in our effort.

2.3. *Sampling and Mesh Types*

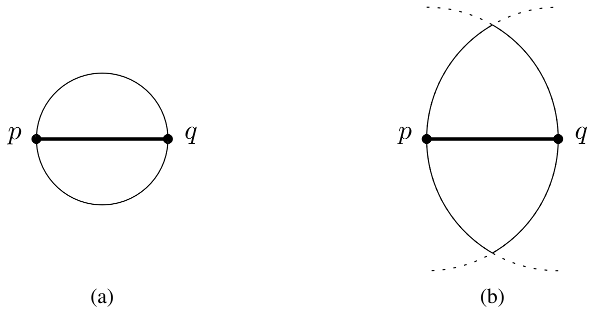
Two well-known sampling strategies, *regular* and *random*, are often used to generate real-world data sets. Most of the random sampling methods were developed in statistics [27,28] as a response to the need for rigorous representative selection of the subsets from the full set. In case of scalar functions, this translates into the selection of discrete points in the domain of the function. Regular sampling is a selection of points in the domain of the function with a fixed step in each dimension. Random sampling is a probabilistic selection of points within the domain, usually with requirements such as uniform density or variability. In this work, we use random sampling of scalar data without any constraints.

Since we deal with sampled scalar data that can lie in higher dimensions, we skip approximation schemes that try to interpolate (using polynomials or splines) inside the whole domain by decomposing it into cells. We rather focus on a simple approximation scheme that connects the samples via a mesh and linearly interpolates the function along the edges of the mesh. In some cases, we will use cell decomposition methods, but only for the mesh generation purposes.

There are a number of methods that work with non-regular sampling. One example is the Delaunay triangulation. It connects all the sample points into simplices of the same dimensionality as the underlying domain and, for each, guarantees that there are no sample points inside its circumsphere. It is frequently used because it produces average sized simplices. We use a mesh, generated by this method, in our evaluation.

Computing triangulated meshes can be expensive. So it is common to use sparser meshes. One example are meshes generated by empty region graphs [29]. As the name suggests, these meshes are constructed by connecting any two sample points whenever their “region” is empty, i.e., free of other sample points. Depending on the definition of “region”, they assume different names; for example, *Gabriel graph* or *relative neighborhood graph* meshes, see Figure 4. The former was used in the work by Oesterling et al. [11]. The latter was suggested as a primary choice for data with prohibitive dimensionality or size by Correa and Lindstrom [1]. We include these two types of meshes into our evaluation.

Finally, we consider meshes that are generated by parameterized neighborhood graphs. The first one is a *k–nearest neighbors* graph that constructs the mesh by connecting each point with its k nearest points. This type of mesh is used in number of studies in scalar field topology [8,10]. Since there is no fixed way of selecting the value of k, in the evaluation we use the mesh generated by the minimum value of k such that the mesh is connected, unless



(a)

(b)

Fig. 4: In an empty region graph, an edge between any two points p and q exists, if their "region" is empty, i.e., free of other points. (a) For the Gabriel graph, the "region" is a circle with an edge (p,q) as a diameter. (b) For the relative neighborhood graph, the "region" is an intersection of two half-circles, with centers at p and q, and a radius (p,q).

stated otherwise. Second, we consider a mesh, generated by Vietoris–Rips complex [30], as a comparison alternative for k–nearest neighbors mesh. This complex is constructed by placing a simplex for every set of points within pairwise distance r from each other. When the underlying data lies on some sub-manifold of the ambient space, a Vietoris–Rips complex serves as an approximation of this manifold. Similar to the k–nearest neighbors mesh, we use the minimum value of r that makes its mesh connected as a default.

2.4. *Evaluation of Mesh Types*

Although different combinations of sampling strategies and meshes are often used, the question of how far the approximated topology diverges from the original is rarely addressed. Usually, it is just assumed that the selected combination of sampling and mesh construction sufficiently approximates the original function.

An explicit attempt to address this question appears in the work of Correa and Lindstrom [1], who evaluated different mesh types in terms of the correct extrema discovery. They quantified the number of false positive and true negative cases of extrema classification and computed normalized harmonic mean, which they called the F-measure. They also proposed improved *relaxed* mesh types, given the results of evaluation based on the F-measure and additional observations. Such mesh types are obtained by relaxing the containment requirement for different empty region graphs [1]. Maljovec et al. [31] further evaluated mesh types and how they affect approximation of Morse complexes. We follow the general idea of evaluating mesh types, however we propose different measures to capture the error that are based on quantifying sub-level set topology.

Indeed, small perturbations of the data can generate an arbitrarily large number of false extrema. On the other hand, it is easy to recognize most of them as noise precisely because they result from a small perturbation, and, therefore, their persistence is low. Moreover, they can be explicitly eliminated with a small change of the function [32]. By focusing only on the number of extrema, the F-measure overlook this crucial distinction: not all false extrema are created equal. In contrast, our work considers the persistence of the extrema. By doing so, we automatically de-emphasize the noise in our evaluation, in effect, measuring how well an approximation preserves important topological features of the function.

It should be noted that somewhat similar importance measures for critical points [33], and extremal lines and surfaces [34], were previously proposed. However, unlike those, our measures specifically target the topological error introduced by approximation for evaluation purposes.

3. Evaluation

In this section, we evaluate different types of meshes, used to approximate a function from a set of samples, based on how small the error is between the approximated and ground truth functions.

We note that both d_B and d_M, introduced in Section 2, are computed in function value units and, thus, can vary greatly for different functions. To eliminate this dependency, we normalize the distances based on their value range.

For the distance between merge trees, the smallest possible value is zero. To find the largest possible value, we consider the difference between the global extremum m and the root saddle r of a merge tree T. The difference equals $d_{T_g} = |m_g - r_g|$ for the merge tree T_g of the ground truth function g, and $d_{T_a} = |m_a - r_a|$ for the merge tree T_a of the approximated function a. The inequality $d_{T_g} \geq d_{T_a}$ always holds, since the function a is only a sampling of the function g. Further, the distance between merge trees is computed based only on function value differences between vertices of merge trees, thus it cannot exceed d_{T_g}. We use this difference to normalize distances, resulting in error measures ε_B and ε_M that we use further in our evaluation.

3.1. *Sampling Density and Dimensionality*

As discussed in Section 2, we consider five types of meshes in our evaluation: Delaunay mesh (DEL); relative neighborhood mesh (RNG); Gabriel mesh (GAB); k-nearest-neighbor (KNN) mesh; and Vietoris-Rips mesh (VR).

We start with a simple function to explain our evaluation. We consider a two-dimensional *SinCos* function, defined for (x,y) defined on $\{[0,4*\pi]\times[0,4*\pi]\}$, slightly tilted in order to resolve ambiguities between extrema values. We take a gradually increasing set of random samples within the domain, and compute an approximate version of the *SinCos* function for each set size. For each approximate function, we compute error measures. While both error measures were computed, considering the error measure ε_B was sufficient to demonstrate our observations, as the error measure ε_M showed similar behavior. Hence, we use the ε_B throughout the evaluation, and present a detailed analysis of the observed differences between ε_B and ε_M separately in Section 3.3.

As expected, and shown in Figure 5 we can see that an increase of the number of samples leads to a decrease of the error. The error converges towards zero, but becomes zero only when randomly selected samples sufficiently cover the function and all the critical points of the original function are sampled. However, depending on the type of mesh used to

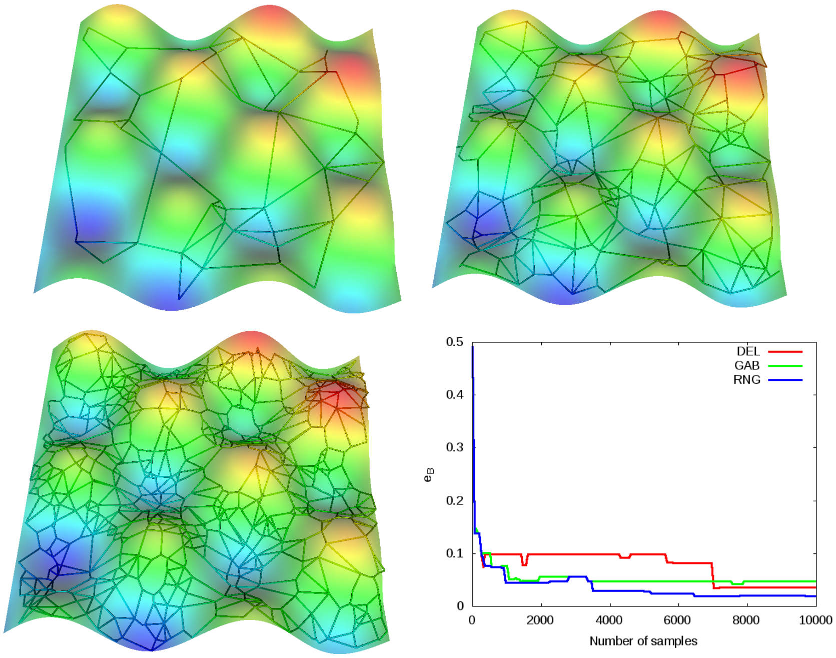


Fig. 5: As we increase the number of samples taken from *SinCos* function, the error decreases, with mesh type (DEL).

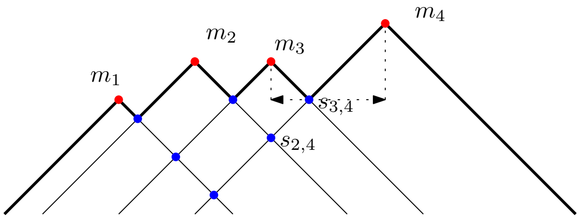


Fig. 6: The generated function is indicated by the bold black line, a supremum of mountain peaks created by maxima generators. Red nodes are peaks/maxima and blue nodes are possible saddles at intersections of each pair of mountains. Among those, saddle $s_{3,4}$ belongs to the given function (as it lies on the surface), while $s_{2,4}$ does not (as it lies under the surface). A function value and coordinates of any saddle can be explicitly computed, as we are given coordinates of each maxima (see dotted lines).

approximate the function from the set of samples, the behavior of the error can be different. We see that all three types of meshes (DEL, GAB, RNG) behave similarly in a sense of leading to smaller error and fast convergence towards zero.

To determine whether an observed behavior also is evident in other relatively simple

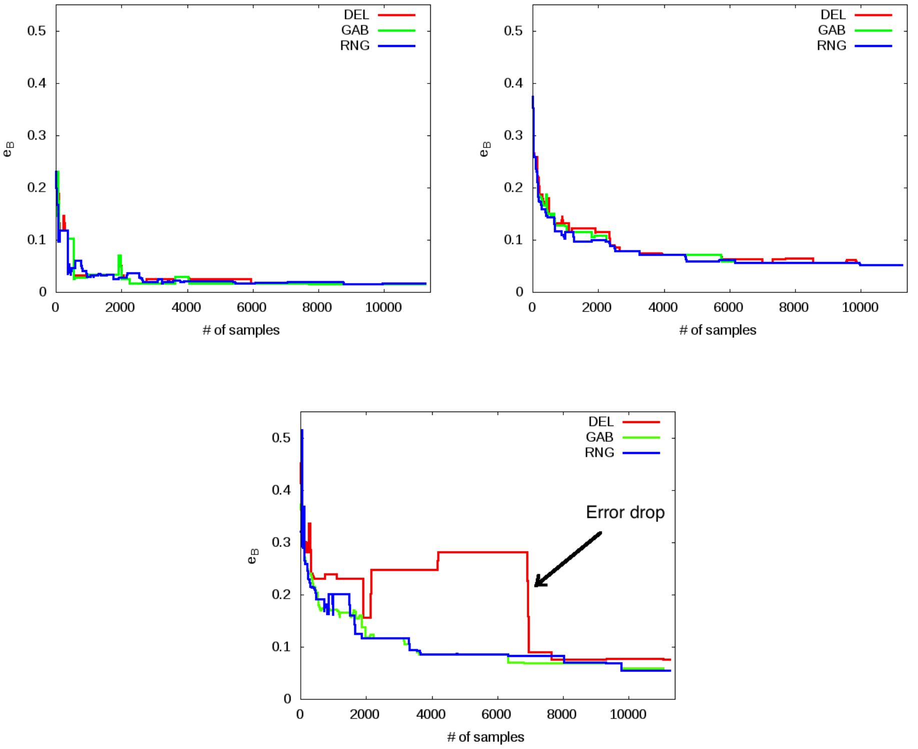


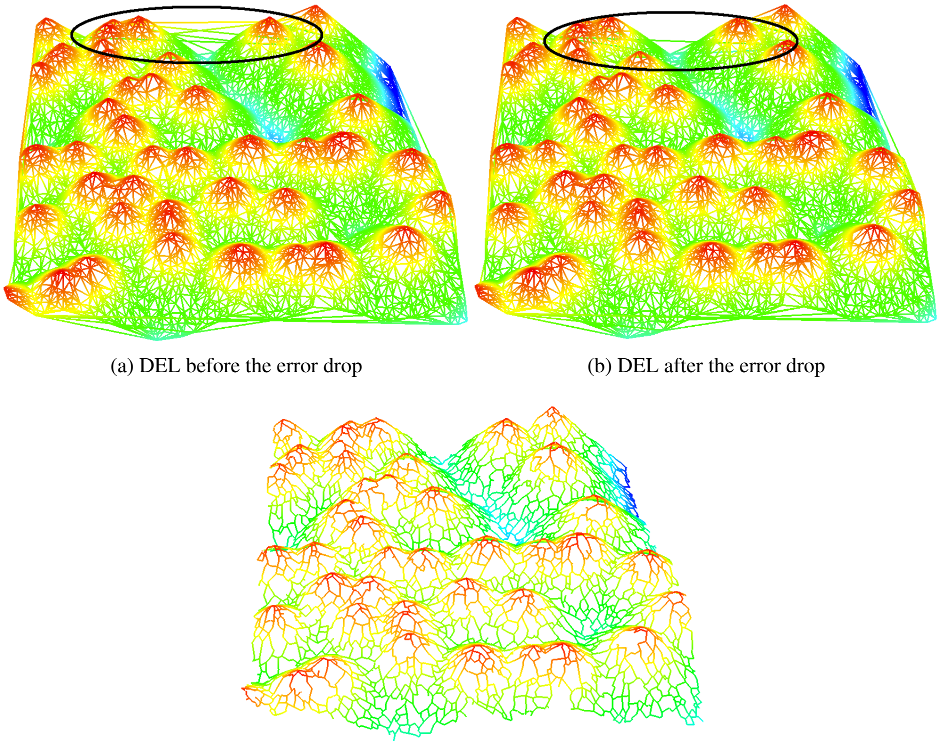
Fig. 7: The error measure ε_B for MoC functions with varying numbers of maxima (3, 20, 50). As expected, more maxima slow the error decrease rate, as it takes more samples to discover all maxima. For the $max3$ function, the error drops below 0.1 around 400 samples, for the $max20$ function around 1000 samples, for the $max50$ function around 3000 samples. We also note an irregular increase of the error for the DEL mesh, between 2100 and 6900 samples.

functions (moderate-dimensional functions with few critical points and no complex degeneracies), we generate a set of Mixture of Caps (MoC) functions as follows: we choose a parametric function that is an upper-envelope of cones, with a fixed slope coefficient k, with apexes at a set M of randomly generated maxima with given values $val(m)$:

$$f_{gt}(p) = \sup_{m \in M} (val(m) - k * d(m,p)),$$

where $d(m,p)$ is the Euclidean distance between two points, see Figure 6. The reason for choosing functions of this form is the fact that they make it possible to analytically compute (and control) exact coordinates and values of all the critical points of the ground truth function, allowing exact computation of its merge tree.

Our comparative experiments, where each uses a MoC function with various parameters, generally confirmed the behavior observed for the $SinCos$ function, see Figure 7. Moreover, the observed behavior persisted when we used additional randomly sampled



(a) DEL before the error drop

(b) DEL after the error drop

(c) RNG before the error drop

Fig. 8: An irregular increase of the error in Figure 7. We construct the DEL mesh before and after the error drop around 6900 samples. The DEL mesh before the error drop has many long edges along the boundaries of the domain, which increases the likelihood of over-connecting the maxima (which leads to reduction of the persistence of some maxima). The error decreases when the newly inserted samples reduce some of the long edges. We also show the RNG mesh before the error drop for comparison, and note that it is not affected by the problem of over-connecting boundary edges.

data sets, sets of the same cardinality for each function, thus accounting for the random nature of our approach especially when dealing with small sparse sets. We discovered a problem with the DEL mesh, which led to a sudden error increase for some functions, as shown in Figure 7(c). A more detailed analysis of this sudden error increase shows that in case of random sampling, the DEL mesh might have long edges along domain borders that can cause sudden spikes of error, see Figure 8.

Another parameter to evaluate is dimensionality. It is one of the main concerns in analysis of real-world data. Hence, we extend our evaluation to functions in higher dimensions to determine whether previously observed error behavior holds. The evaluation results for *MoC* functions with varying dimensionality are shown in Figure 9. While we see previ-

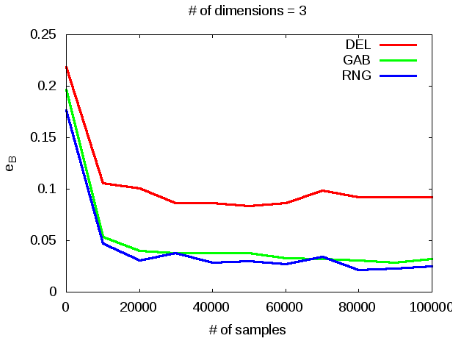


Fig. 9: With increase of dimension we still observe the expected decrease of error, although the rate of decrease becomes less steep, which is expected due to sparser density of samples. We note that the performance of the DEL mesh is relatively worse as it starts to over-connect the domain.

ously observed behavior, the decrease rate of error becomes slow due to sparser sample coverage of the domain in higher dimensions. A more detailed analysis of error decrease rates is provided in Section 3.3.

So far, we evaluated mesh errors based on sampling density and dimensionality. However, for simplicity we considered only DEL, GAB, and RNG meshes, because the KNN and VR meshes are parametric meshes, and they require additional evaluation that takes into account parameters used in their construction process (the number of nearest neighbors k for a KNN mesh, and the radius r for a VR mesh). We discuss experimental results for KNN and VR meshes, while using the RNG mesh for comparison purposes.

Since no optimal strategy is known for selecting the parameter k for a KNN mesh, we evaluate a minimally connected version of the KNN mesh. This version is based on using the lowest value of the parameter k, for which the mesh is connected. We note that k differs for every set of samples. Comparison of the minimally connected KNN mesh to the RNG mesh is shown in Figure 10. We see that the performance of the minimally connected KNN mesh is highly unstable, mainly due to the small number of considered neighbors k for each sample.

In comparison, the average number of neighbors for the minimally connected VR mesh is significantly higher. Hence, as expected, a similar evaluation of the minimally connected VR mesh shows more stable behavior and lower error, see Figure 10(left). Indeed, when fixing the number of neighbors for the KNN mesh to the average number of neighbors of the minimally connected VR mesh, we see a significant improvement, see Figure 10(right).

To evaluate the KNN mesh further, we performed an additional sweep of the parameter k. The resulting error function is given in Figure 11 as a surface, where we can see the valley that corresponds to the higher value(s) of the parameter k, for which the KNN mesh performs better. While these results do not provide definitive means to calculate the optimal value of the parameter k, we see that often choosing moderately high values of the

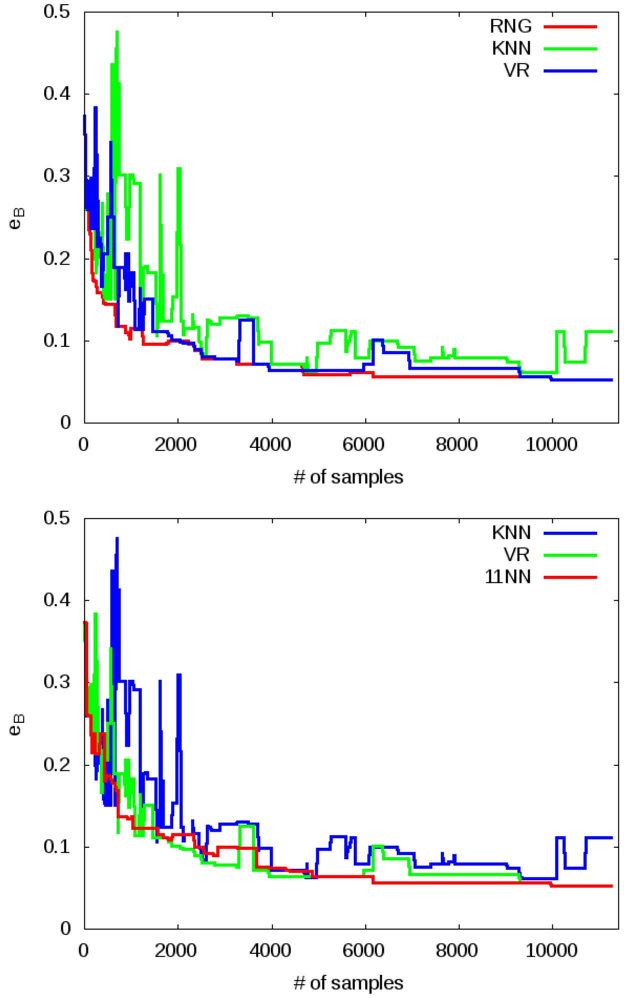


Fig. 10: (Left) The minimally connected KNN and VR meshes compared to the RNG mesh. The minimally connected VR mesh performs better as it approximates the function better. (Right) The average number of neighbors for the minimally connected VR mesh for the case on left equals 11. When setting the number of neighbors for the KNN mesh to 11 (hence called 11NN mesh), and comparing it to the minimally connected KNN and VR meshes, we see significantly more stable performance.

parameter *k* reduces the error. We note that the latter observation agrees with the similar results of the evaluation using F-measure, mentioned earlier.

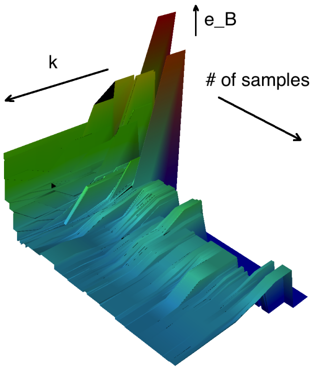


Fig. 11: The sweep of the parameter k versus the number of samples. The deep blue area on the right side corresponds to the values of k, for which the mesh is disconnected, thus no results are available. In nearly all cases we see the increase and then a decrease of the error, with growing k value.

3.2. *Mesh Evaluation Summary*

We summarize the results of the mesh evaluation: (1) the GAB and RNG meshes show stable good performance, with the RNG mesh being slightly better in cases with lower number of samples; (2) the relaxed RNG and GAB meshes perform similarly when compared to the original versions; (3) the performance of the minimally connected KNN mesh is highly unstable, while the minimally connected VR mesh performs significantly better; (4) for a small number of samples, the minimally connected KNN and VR meshes perform similarly when compared with the GAB and RNG meshes, thus their use can be justified for such cases.

We note that these results are applicable to functions that are similar to the types of functions we have used as ground truth, i.e., a function must have relatively slowly changing slopes (no spikes) and no complex topology (e.g., loops).

3.3. *Analysis of Error Measures*

We now present an analysis of the error measures. In particular, we first conduct a comparative analysis of ε_B and ε_M, and then of their relationship to the F-measure. We also provide experimental results of applying the measures to real-world data.

In Figure 12 we see how the error measures ε_B and ε_M compare. The evaluation results suggest that ε_M leads to a more precise error approximation, since it is consistently higher than or equal to ε_B, while both are lower than the infinity norm between the functions. The difference between the two measures reflects that merge trees are a finer representation of topology of the sub-level sets. Thus, ε_M not only captures the persistence, but also the structural error in the mesh. Figure 12 shows that as we increase the size of the sampling, the structural error for DEL, GAB, RNG meshes dissipates, while it stays high for KNN and VR meshes. Furthermore, the GAB mesh leads to the smallest error due to the structure

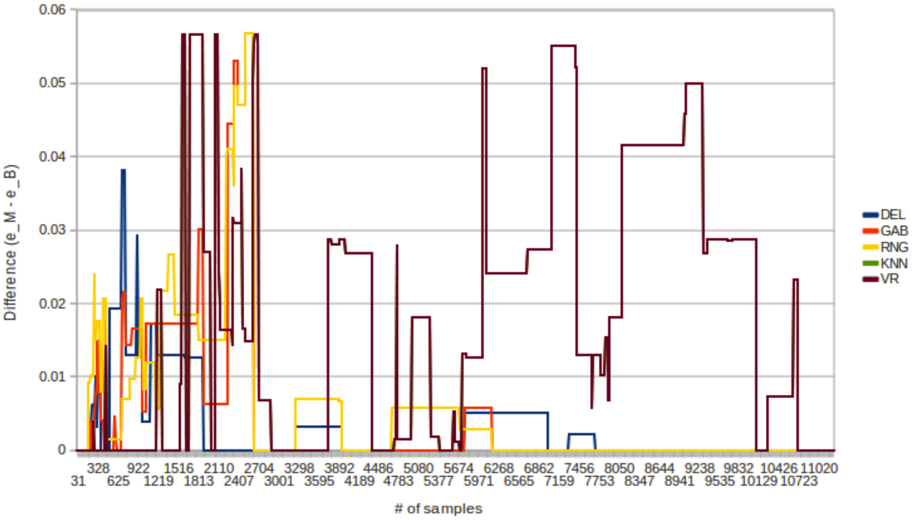


Fig. 12: Difference between the error measures show additional structural error that meshes impose.

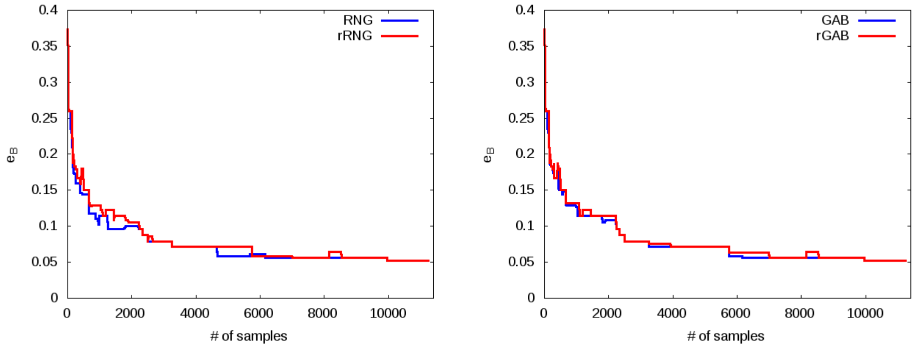


Fig. 13: Relaxed versions of RNG and GAB meshes perform very close to default versions in terms of proposed error measures, while according to the original study [1], relaxed versions significantly decrease the F-measure error.

of the mesh.

This observation is consistent with the earlier findings [4] that the bottleneck distance (the basis of the ε_B measure) fails to account for the structural difference captured by the distance between merge trees (the basis of the ε_M measure), see Figure 3.

Another question is how the proposed error measures compare to the F-measure [1]. We note that the F-measure is based on finding "false positive" (recall) and missed (precision) extrema, which requires the sampling set to include all the extrema of the ground truth func-

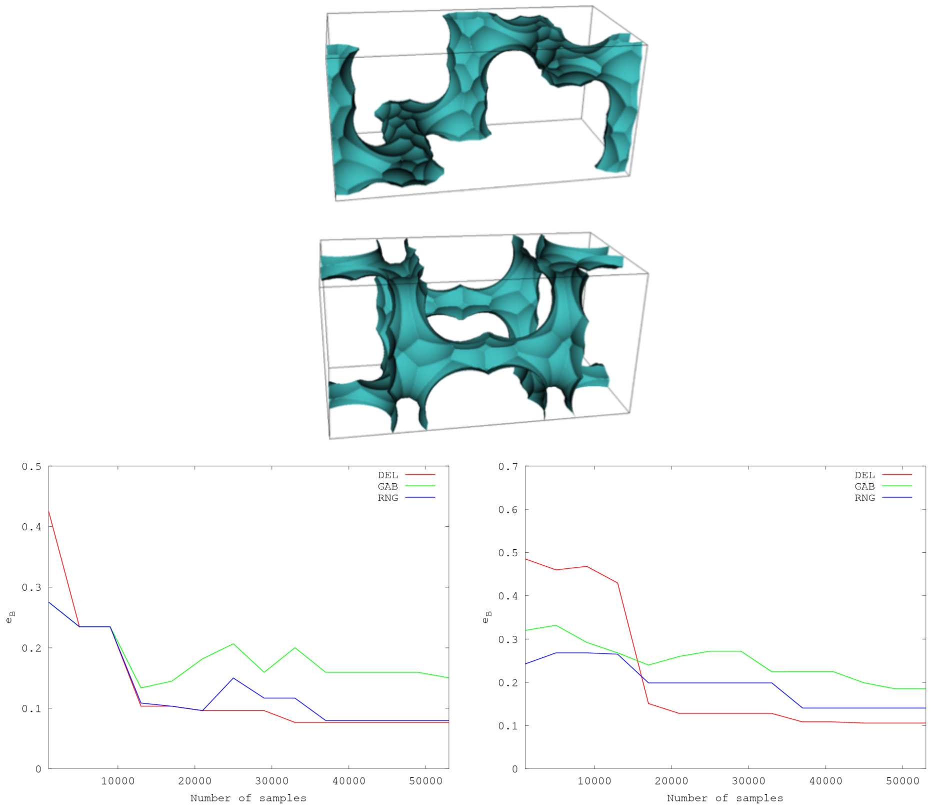


Fig. 14: Two types of porous zeolite materials (CHA, LTL) were selected. While for the first material (top) the GAB mesh performs slightly better, for the second material (bottom), the RNG mesh performs better. The DEL mesh performs consistently worse than the other two meshes.

tion. However, both the $\varepsilon_B, \varepsilon_M$ measures are based on persistence-based distances, making them independent of the information about the extrema, and, as a consequence, independent of the F-measure. Experimental results of comparing default and "relaxed" versions of RNG and GAB meshes in terms of the $\varepsilon_B, \varepsilon_M$ measures and the F-measure support this observation. Indeed, the original study [1] demonstrated that the "relaxed" versions of meshes significantly decrease the F-measure, i.e, decrease the error with regard to extrema. However, the "relaxed" versions of RNG and GAB meshes performed very close to the default versions in all our experiments with the ε_B and ε_M measures, see Figure 13.

3.3.1. *Error Analysis for Real-world Data*

Finally, we consider real-world data. We focus on simulated data set of porous zeolite materials (see Keffer et al. [35] for simulation details), in particular we select two types of zeolite materials (CHA, LTL), each presented as a 3D scalar function, sampled regularly within a domain. The sample sizes of each data set are $69 \times 69 \times 74$ and $91 \times 91 \times 38$ correspondingly. Since no ground truth is known, we assume that the triangulated highest resolution sampling represents the ground truth, and evaluate DEL, GAB, and RNG meshes by taking the random subsets of given samples, and computing error measures for them. Figure 14 shows that most of the observations we made previously hold, namely the gradual decrease of the error as a number of samples increase, and the problem with long edges in the DEL mesh when the number of samples is low. This result is important, as it demonstrates that proposed measures can be applied to functions, even if they do not satisfy Lipschitz continuity condition, which is the case for majority of experimental and simulated data.

One additional observation can be made in the Figure 14. As the number of samples approach the original, we can see that DEL mesh becomes more precise. This is consistent with an expected behavior, as the ground truth is computed using a simplicial mesh.

4. Conclusions

We have presented a new approach to studying the approximation error in scalar field topology. In particular, we consider the loss of topological information, related to the sub-level sets of a function. For that purpose, we used the distance between merge trees and the bottleneck distance for persistence diagrams to define error measures for topological information loss. We offered an evaluation of the different types of meshes using the proposed error measures, and based on the results, discussed performances of selected types of meshes, leading to several recommendations.

We note that the produced recommendations are applicable to functions similar to the selected class of functions. Two features of the selected class of functions that distinguishes it are: functions are relatively slow changing; functions have no complex topology. We leave an evaluation of meshes for various other classes of functions to future work.

To address possible concerns about the proposed measures, we presented a detailed analysis of proposed measures. Namely, we investigated how they compare to each other and to the F-measure, as well as how proposed measures perform, when applied to real-world data. An important application of the presented work would be to use its results towards development of a topology-aware mesh, i.e., type of mesh that reduces the error in approximating the topological information.

Finally, an interesting venue to explore is the evaluation of the adaptive topological sampling methods, proposed by Maljovec et al. [36]

Acknowledgements

The authors would like to thank Aidos Abzhanov and Aidos Adilkhanov for their helpful comments. This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231, through the grant "Towards Exascale: High Performance Visualization and Analytics," program manager Dr. Lucy Nowell. This work was also funded by the Ministry of Education and Science of the Republic of Kazakhstan under the target program “NU-Berkeley Strategic Initiative in Warm-dense Matter, Advanced Materials and Energy Sources for 2014-2018” (Program No. 0115PK03029). The authors further thank Kanat Baigarin and Olzhas Makhambetov, both with Nazarbayev University, for their organizational support and their administrative assistance with grant funds for the Kazakhstani portion of the research.

References

1. C. Correa and P. Lindstrom, Towards robust topology of sparsely sampled data, *IEEE Transactions on Visualization and Computer Graphics* **17**.
2. S. Takahashi, G. M. Nielson, Y. Takeshima and I. Fujishiro, Topological volume skeletonization using adaptive tetrahedralization, in *Proceedings of the Geometric Modeling and Processing 2004* (IEEE Computer Society, Washington, DC, USA, 2004), GMP ’04, pp. 227–236.
3. D. Cohen-Steiner, H. Edelsbrunner and J. Harer, Stability of persistence diagrams, in *Proceedings of 21st Annual Symposium on Computational Geometry* (ACM, 2005), pp. 263–271.
4. K. Beketayev, D. Yeliussizov, D. Morozov, G. H. Weber and B. Hamann, Measuring the distance between merge trees, in *Topological Methods in Data Analysis and Visualization III*, eds. P.-T. Bremer, I. Hotz, V. Pascucci and R. Peikert (Springer International Publishing, 2014), Mathematics and Visualization, pp. 151–165.
5. J. W. Milnor, *Morse Theory* (Princeton University Press, Princeton, New Jersey, 1963).
6. R. L. Boyell and H. Ruston, Hybrid techniques for real-time radar simulation, in *Proceedings of the 1963 Fall Joint Computer Conference* (IEEE, 1963), pp. 445–458.
7. H. Carr, J. Snoeyink and U. Axen, Computing contour trees in all dimensions, *Computational Geometry – Theory and Applications* **24** (2003) 75.
8. S. Takahashi, L. Fujishiro and M. Okada, Applying manifold learning to plotting approximate contour trees, *IEEE Transactions on Visualization and Computer Graphics* **15** (2009) 1185.
9. G. H. Weber, P.-T. Bremer and V. Pascucci, Topological landscapes: A terrain metaphor for scientific data, *IEEE Transactions on Visualization and Computer Graphics* **13** (2007) 1416.
10. W. Harvey and Y. Wang, Topological landscape ensembles for visualization of scalar-valued functions, *Computer Graphics Forum* **29** (2010) 993.
11. P. Oesterling, C. Heine, H. Janicke, G. Scheuermann and G. Heyer, Visualization of high-dimensional point clouds using their density distribution’s topology, *IEEE Transactions on Visualization and Computer Graphics* **17** (2011) 1547.
12. G. Reeb, Sur les points singuliers d’une forme de pfaff complement intergrable ou d’une fonction numerique, *Comptes Rendus Acad. Science Paris* **222** (1946) 847.
13. H. Edelsbrunner, J. Harer and A. Zomorodian, Hierarchical Morse-Smale complexes for piecewise linear 2-manifold, *Discrete & Computational Geometry* **30** (2003) 87.
14. H. Edelsbrunner, J. Harer, V. Natarajan and V. Pascucci, Morse-Smale complexes for piecewise linear 3-manifolds, in *Proceedings of the 19th ACM Symposium on Computational Geometry* (2003), pp. 361–370.
15. S. Gerber, P.-T. Bremer, V. Pascucci and R. Whitaker, Visual exploration of high dimensional scalar functions, *IEEE Transactions on Visualization and Computer Graphics* **16**.
16. J. R. Munkres, *Elements of Algebraic Topology* (Addison-Wesley, Redwood City, California, 1984).
17. C. Flamm, I. L. Hofacker, P. Stadler and M. Wolfinger, Barrier trees of degenerate landscapes, *Physical Chemistry* **216** (2002) 155.
18. C. Heine, G. Scheuermann, C. Flamm, I. L. Hofacker and P. F. Stadler, Visualization of barrier tree sequences, *IEEE Transactions on Visualization and Computer Graphics* **12** (2006) 781.
19. S. Biasotti, L. De Floriani, B. Falcidieno, P. Frosini, D. Giorgi, C. Landi, L. Papaleo and

M. Spagnuolo, Describing shapes by geometrical-topological properties of real functions, *ACM Computing Surveys* **40** (2008) 12:1.

20. H. Doraiswamy, N. Shivashankar, V. Natarajan and Y. Wang, Topological saliency, *Computers & Graphics* **37** (2013) 787.
21. P. Bubenik, Statistical topological data analysis using persistence landscapes, *ArXiv e-prints* .
22. F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas and S. Oudot, Proximity of persistence modules and their diagrams, in *Proceedings of the twenty-fifth annual symposium on Computational geometry* (ACM, 2008), pp. 237–246.
23. D. Morozov, K. Beketayev and G. Weber, Interleaving distance between merge trees, in *Topology-Based Methods in Visualization* (2013).
24. U. Bauer, X. Ge and Y. Wang, Measuring distance between reeb graphs, in *Proceedings of the thirtieth annual symposium on Computational geometry* (ACM, 2014), p. 464.
25. V. Narayanan, D. M. Thomas and V. Natarajan, Distance between extremum graphs, in *Visualization Symposium (PacificVis), 2015 IEEE Pacific* (IEEE, 2015), pp. 263–270.
26. K. Jisu, Y.-C. Chen, S. Balakrishnan, A. Rinaldo and L. Wasserman, Statistical inference for cluster trees, in *Advances in Neural Information Processing Systems* (2016), pp. 1831–1839.
27. S. Lohr, *Sampling: Design and Analysis* (Brooks/Cole, Cengage Learning, 2010).
28. D. Bursztyn and D. M. Steinberg, Comparison of design for computer experiments, *Journal of Statistical Planning and Inference* **136** (2006) 1103.
29. J. Cardinal, S. Collette and S. Langerman, Empty region graphs, *Comp. Geom. Theory Appl.* **42** (2009) 183.
30. L. Vietoris, Über den höheren zusammenhang kompakter räume und eine klasse von zusammenhangstreuen abbildungen, *Mathematische Annalen* **97** (1927) 454.
31. D. Maljovec, A. Saha, P. Lindstrom, P.-T. Bremer, B. Wang, C. Correa and V. Pascucci, A comparative study of morse complex approximation using different neighborhood graphs, in *Topological Methods in Data Analysis and Visualization III* (Springer, 2013), Mathematics and Visualization.
32. H. Edelsbrunner, D. Morozov and V. Pascucci, Persistence–sensitive simplification functions on 2-manifolds, in *Proceedings of 22nd Annual Symposium on Computational Geometry* (ACM, 2006), pp. 127–134.
33. J. Reininghaus, N. Kotava, D. Guenther, J. Kasten, H. Hagen and I. Hotz, A scale space based persistence measure for critical points in 2d scalar fields, *IEEE Transactions on Visualization and Computer Graphics* **17** (2011) 2045.
34. D. Günther, H.-P. Seidel and T. Weinkauf, Extraction of dominant extremal structures in volumetric data using separatrix persistence, *Comp. Graph. Forum* **31** (2012) 2554.
35. D. Keffer, V. Gupta, D. Kim, E. Lenz, H. T. Davis and A. V. McCormick, A compendium of potential energy maps of zeolites and molecular sieves, *Journal of Molecular Graphics and Modelling* **14** (1996) 108.
36. D. Maljovec, B. Wang, A. Kupresanin, G. Johannesson, V. Pascucci and P.-T. Bremer, Adaptive sampling with topological scores, Working with Uncertainty Workshop, October 2011.