

A Structural Average of Labeled Merge Trees for Uncertainty Visualization

Lin Yan, Yusu Wang, Elizabeth Munch, Ellen Gasparovic, and Bei Wang

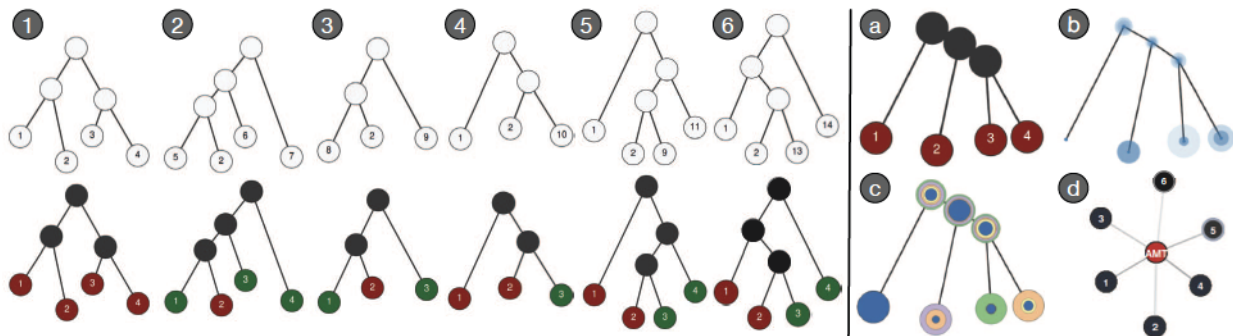


Fig. 1. Computing a 1-center tree as a structural average for an input ensemble of six leaf-labeled merge trees in partial agreement. Top row on the left: six input leaf-labeled merge trees, each visualized by a node-link diagram. Bottom row on the left: each input tree is visualized with original labels in red and updated labels in green. Right: various uncertainty visualizations for the 1-center tree: (a) 1-center tree with leaf labels; (b) variational vertex consistency plot; (c) statistical vertex consistency plot; (d): summary plot that shows the interleaving distance between each input tree and the 1-center tree (denoted as AMT).

Abstract—Physical phenomena in science and engineering are frequently modeled using scalar fields. In scalar field topology, graph-based topological descriptors such as merge trees, contour trees, and Reeb graphs are commonly used to characterize topological changes in the (sub)level sets of scalar fields. One of the biggest challenges and opportunities to advance topology-based visualization is to understand and incorporate uncertainty into such topological descriptors to effectively reason about their underlying data. In this paper, we study a structural average of a set of labeled merge trees and use it to encode uncertainty in data. Specifically, we compute a 1-center tree that minimizes its maximum distance to any other tree in the set under a well-defined metric called the interleaving distance. We provide heuristic strategies that compute structural averages of merge trees whose labels do not fully agree. We further provide an interactive visualization system that resembles a numerical calculator that takes as input a set of merge trees and outputs a tree as their structural average. We also highlight structural similarities between the input and the average and incorporate uncertainty information for visual exploration. We develop a novel measure of uncertainty, referred to as consistency, via a metric-space view of the input trees. Finally, we demonstrate an application of our framework through merge trees that arise from ensembles of scalar fields. Our work is the first to employ interleaving distances and consistency to study a global, mathematically rigorous, structural average of merge trees in the context of uncertainty visualization.

Index Terms—Topological data analysis, uncertainty visualization, merge trees

1 INTRODUCTION

In topological data analysis and visualization, topological descriptors have been used to understand and summarize the shape of complex data in science and engineering, ranging from gene expression of breast cancer tumors [44] to high-throughput screening of nanoporous materials [42]. For data modeled as scalar fields, the most popular descriptors include vector-based such as persistence diagrams [25] and barcodes [13, 32], graph-based such as merge trees [7], contour trees [15], and Reeb graphs [64], as well as complex-based such as Morse and Morse-Smale complexes [23, 24, 31].

These topological descriptors provide meaningful abstractions, reduce the amount of data to be processed, utilize sophisticated hierarchi-

cal representations that capture features at multiple scales, and enable progressive feature simplifications. However, as pointed out by Heine et al. [36], one of the biggest challenges and opportunities is to develop original approaches that *incorporate uncertainty* into topological descriptors to advance topology-based visualization. In this paper, we focus on *merge trees*, which are a special type of topological descriptor that tracks the evolution of connected components in the sublevel sets of scalar fields [47]. We compute *structural averages* of a set of labeled merge trees (referred to as an *ensemble*), and utilize such averages in uncertainty visualization. Our work is motivated from a statistical and a visualization perspective.

Statistics on topological descriptors. In statistics, the concept of an *average* refers to a measure of central tendency. Given a set of numbers, the notions of *mean*, *median*, and *mode* are considered as *numeric averages* in different contexts. In this paper, we study a certain *structural average* among a set of labeled merge trees. In particular, we would like to find a *metric 1-center* of a finite set of labeled merge trees; that is, an *average tree* that minimizes the maximum distance to any other tree in the set. Given a metric space (X, d) , a 1-center of a finite point set $P = \{p_1, \dots, p_m\} \subset X$ is a point $c \in X$ which satisfies $c = \arg \min_{x \in X} \max_{p \in P} d(x, p)$, i.e., $c \in X$ is the center of a minimum enclosing ball of P . Here, our metric space is the space of all labeled merge trees equipped with a metric referred to as the *interleaving distance for labeled merge trees* [48]. The 1-center is our notion of an average tree.

- Lin Yan is with University of Utah. E-mail: lin.yan@utah.edu.
- Yusu Wang is with Ohio State University. E-mail: yusu@cse.ohio-state.edu.
- Elizabeth Munch is with Michigan State University. E-mail: muncheli@egr.msu.edu.
- Ellen Gasparovic is with Union College. E-mail: gasparoe@union.edu.
- Bei Wang is with University of Utah. E-mail: beiwang@sci.utah.edu.

Manuscript received 31 Mar. 2019; accepted 1 Aug. 2019.

Date of publication 16 Aug. 2019; date of current version 20 Oct. 2019.

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.

Digital Object Identifier no. 10.1109/TVCG.2019.2934242

Our implementation in computing a 1-center is a first step toward performing statistical analysis of merge trees. By introducing a potentially extendable metric between labeled merge trees, we envision a complete suite of statistics describing the probability distribution of graph-based topological descriptors.

Visualization of uncertainty. Advances in technology such as increased bandwidth, storage, and computational power have led to a large amount of complex data. To effectively and accurately communicate such data to scientists via visualization, we should pay close attention to information about uncertainty, including accuracy, confidence, variability, and model bias [8].

Conveying uncertainty information through visualization, i.e., *uncertainty visualization*, is a very active area of research; e.g., see surveys [33, 62, 71]. Uncertainty information is often summarized via statistical quantities (such as mean, median, and standard deviation) and encoded with data via color, opacity, texture, glyphs, animation, etc. [62]. Uncertainty for 1D scalar fields is often expressed *point-wise* as error bars or boxplots [59]. In the case of 2D and 3D scalar fields, uncertainty can be encoded *feature-wise* within spatial domains (such as contours [57, 58, 75], 2D surfaces [60], and 3D volumes [22, 61]), as well as *structure-wise* as trees [41] or lattice graphs [19].

The challenges associated with the study of structure-wise uncertainty in graph-based descriptors can be attributed, in part, to the difficulty in characterizing and quantifying their structural uncertainty. In particular, ensemble datasets combining multiple realizations of a phenomenon are often used to mitigate the effects of uncertainty. State-of-the-art approaches typically aggregate over inherent data dimensions to reduce the structure and size of the data. Unfortunately, these aggregations can lose the richness in both global and local structures. For instance, a scalar field obtained by averaging the values from all the ensemble scalar field members at each data point (as a form of aggregation) was treated previously as the *ensemble mean*, and the contour tree of the ensemble mean was referred to as the *mean contour tree* [77].

In this regard, our work differs from previous approaches significantly. First, instead of computing a merge tree from an average scalar field, we calculate an average merge tree *directly* from a set of input trees that perhaps arise from an ensemble, and use such an average to encode uncertainty. Second, constructing an average labeled merge tree has a clean, mathematical foundation via a metric-space view, and does not rely on ad hoc operations. To the best of our knowledge, our work is the first to employ interleaving distances and consistency measures to study a mathematically rigorous structural average of merge trees in the context of uncertainty visualization.

Contributions. In this paper, we compute and visualize a structural average of labeled merge trees. Our work builds upon theoretical foundations regarding interleaving distances between labeled merge trees [30]. Such a distance is chosen because, unlike many of the other metrics available for merge trees, it is easily computable. It is also stable with respect to the input function data [48], and has a well-defined 1-center as well as geodesics. Our main contributions are the algorithms, implementations, and visualization design in moving from theory to practice, all of which are highly nontrivial.

- We provide an interactive visualization system to demonstrate the utilities of our proposed algorithms. It resembles a numerical calculator that takes as input an ensemble of leaf-labeled merge trees and outputs a tree as their structural average.
- We introduce heuristic strategies that complete the labelings for an ensemble of merge trees whose labels do not fully agree in order to provide structural averages.
- We highlight structural similarities between the input and the average tree and incorporate uncertainty information for visual exploration. To achieve this, we develop a novel measure of uncertainty for each vertex in the tree, via a metric-space view of the input trees. This measure is also flexible, allowing a local-global tradeoff in understanding structure variations.
- We demonstrate an application of our framework through merge trees that arise from ensembles of scalar fields.

Our framework is applicable to merge trees of the most general form; we give an additional example of applying our framework to merge trees derived from neuron morphology in Appx. F. Our visualization tool and algorithms are released open source under MIT license on Github: <https://github.com/tdavislab/amt>.

2 RELATED WORK

We review the most relevant literature on uncertainty visualization of scalar fields, with a focus on graph-based topological descriptors and topological features. Our work is primarily concerned with structure-wise data uncertainty of scalar fields; for vector and tensor fields, see an overview in [36].

Graph-based topological descriptors. Graph-based topological descriptors include merge trees [7] (also known as barrier trees [27] or join trees [15]), contour trees [15], Reeb graphs [64], mapper graphs [67], and joint contour nets [14]. These descriptors are graph-based representations to illustrate how the topology of level sets or sublevel sets of scalar fields changes with a scalar value parameter.

For a topological space X equipped with a function $f: X \rightarrow \mathbb{R}$, the merge tree encodes the connected components of the sublevel sets $f^{-1}(-\infty, a]$ for $a \in \mathbb{R}$. A closely related descriptor, the Reeb graph, encodes the connected components of the level sets $f^{-1}(a)$ instead. The contour tree [15] is a type of Reeb graph when X is simply connected. All these descriptors are related to Morse theory [46] and level-set topology through relations among critical points. They are widely applied in scientific visualization [49, 74, 76].

Since a contour tree of a function $f: X \rightarrow \mathbb{R}$ can be constructed by carefully combining the merge trees of f and $-f$ in linear time [15], merge tree visualization shares the same design space as that of a contour tree. Contour trees are often visualized with node-link diagrams in two or three dimensions [15, 51, 52, 70]. Such diagrams are simple and powerful tools for abstract data representations [74], contour extraction [73], and data explorations in various application domains [2, 52]. Many attempts have been made in contour tree visualization to overcome difficulties in visual interpretation, visual clutter, and missing topological features [78].

Uncertainty visualization of topological features. Critical points and contours (level sets, iso-surfaces) are important features for the study of scalar field topology. When data is affected by uncertainty, visualization of such topological features should adapt accordingly.

Mihai and Westermann [45] measure the likelihood of the occurrence of critical points with respect to both the positions and types of the critical points. Specifically, when the data uncertainty is described by a Gaussian distribution, confidence intervals are derived for the gradient and the determinant and trace of the Hessian matrix in scalar field ensembles to infer confidence regions for critical points [45]. Gunther et al. [35] characterize critical points and their spatial relation for 2D uncertain scalar fields, where each vertex in a regular grid is assigned a probability density function (PDF) describing its scalar value. They identify so-called mandatory critical points – spatial regions and function ranges where critical points have to occur in any realization of the input based on the PDF.

To visualize the effect of uncertainty on contours, envelopes within a gridded domain are extracted to indicate in which volume the contour will lie (with a certain confidence) [50, 80]. Uncertainty associated with a contour can also be rendered via animation [10] or as a collection of points where each point is displayed from “its original location along the surface normal by an amount proportional to the uncertainty at that point” [34]. Positional and geometrical variations of contours are captured by variability in gradients for uncertain scalar fields [54]. Positional uncertainty of contours could also be encoded by spatial correlation [55, 56] or numerical sensitivity [57]. Building on the notions of functional boxplots and data depth, contour boxplots [75] display statistical quantities analogous to the mean, median, and order statistics for ensembles of contours.

Finally, from an algorithmic perspective, probabilistic marching cubes [58] and positionally uncertain iso-contours [57] study the uncertainties inherent in computing the visual representations.

Uncertainty visualization of graph-based descriptors. Lee et al. have introduced CandidTree, which merges two trees into one and visualizes both location and subtree structural uncertainty [41]. Wu et al. have developed an interactive visualization tool that uses contour trees as abstract data representations to explore data-level uncertainty, contour and topology variability [77]. Kraus has employed grayscale morphology to visualize uncertain substructures in contour trees [39]. Zhang et al. have proposed sampling-based Monte Carlo methods to study contour trees of uncertain terrains, where uncertainty lies in the height function described by a probability distribution [81]. The work most relevant to ours is [77], where a mean contour tree is computed as the contour tree of the mean of an ensemble. However, our work differs significantly from [77] in that instead of computing a tree from an average of the ensemble members, we compute an average tree directly from a set of input trees (that potentially arise from ensemble members).

Distances between topological structures. Recently, many metrics have been proposed for merge trees, often by way of a restriction from a metric on the more general Reeb graph [3–7, 16, 20, 47, 69]. In this paper, we focus on the interleaving distance for labeled merge trees [48]. This distance is an example of an interleaving distance between persistence modules [17], which is brought to graph-based descriptors such as merge trees [47] and Reeb graphs [20] via category theory [11, 21].

The application of the interleaving distance to labeled merge trees can also be viewed as an interpretation of a metric for phylogenetic trees [12]. Our framework differs from the previous work as it relies on a clean and simple metric-space view of input trees that is equipped with geodesics, as well as easy-to-implement algorithms.

3 TECHNICAL BACKGROUND

We first review mathematical notions in graph theory and computational topology, including merge trees, labeled merge trees, and the interleaving distance on labeled merge trees. We then introduce leaf-labeled merge trees with full, partial, or no label agreements, which are structures we deal with in our algorithms. We end this section by giving an example of a *scalar field induced merge tree*, that is, a merge tree that arises from a scalar function on a topological space.

3.1 Theoretical Foundations for Labeled Merge Trees

Our implementation and visualization design is built upon theoretical foundations established in [30], which focus on labeled merge trees of the most general form, and which we review below.

Merge trees and labeled merge trees. A *tree* is an undirected graph in which any two vertices (nodes) are connected by a unique path. A *rooted tree* is a tree in which a special vertex is chosen to be the *root*. The *degree* of a vertex is the number of edges incident to the vertex. We assume the root is of degree 1.

Let $T = (V, E)$ denote a rooted tree with vertex set V and edge set E . The leaves $L \subset V$ of T are the nonroot vertices of degree 1; other nonroot vertices are *internal vertices*.

Definition 3.1. A *merge tree* is a pair (T, f) consisting of a rooted tree T and a function $f: V \rightarrow \mathbb{R} \cup \{\infty\}$ such that $f(u) \neq f(v)$ for all $uv \in E$, $f(v) = \infty$ if and only if v is the root, and every nonroot vertex has exactly one neighbor with a higher function value.

A scalar field induced merge tree is a special case of a merge tree in Definition 3.1, which we review in Sec. 3.3. We require $f(v) = \infty$ for the root v for technical reasons. We also require $f(u) \neq f(v)$ for all $uv \in T$; in practice, this can be achieved by the simulation of simplicity [26]. For any pair of vertices $u, v \in V$, we write $a(u, v) \in V$ for their *lowest common ancestor*; that is, the unique vertex of minimum function value such that the unique path from $a(u, v)$ to either u or v strictly decreases in the value of f . Then $f(a(u, v))$ denotes its function value. For ease of notation, let $[n] := \{1, \dots, n\}$ denote a label set.

Definition 3.2. A *labeled merge tree* \mathcal{T} is a triple (T, f, μ) that consists of a merge tree (T, f) along with a map $\mu: [n] \rightarrow V$ called a *labeling* that is surjective on the set of leaves $L \subset V$. When we need to emphasize the number of labels, we will call this an n -labeled merge tree.

Let $|T| = m$ be the size of the merge tree, or in other words, the number of nonroot vertices in the tree. Note that the definition permits trees with labeled internal vertices as well as vertices with multiple labels; see Fig. 2 (left) for an example. Surjectivity on L means that n , an arbitrary positive integer, satisfies $n \geq |L|$. For practical purposes, we will usually have $n \leq m$, but this is by no means required. Finally, a merge tree (T, f) without a labeling is referred to as an *unlabeled merge tree*.

Induced matrix and interleaving distance. We can build a matrix from a labeled merge tree. The *induced matrix* of an n -labeled merge tree $\mathcal{T} = (T, f, \mu)$ is the symmetric matrix $\mathcal{D} \in \mathbb{R}^{n \times n}$ where $\mathcal{D}_{ij} = f(a(\mu(i), \mu(j)))$. See Fig. 2 (middle) for an example of an induced matrix. The reason for turning the trees into matrices is that we can immediately use a natural choice of distance between the matrices to obtain a distance between the trees as follows:

Definition 3.3. Given two n -labeled merge trees $\mathcal{T}^1 = (T_1, f_1, \mu_1)$ and $\mathcal{T}^2 = (T_2, f_2, \mu_2)$, the (*labeled*) *interleaving distance* between them is defined to be the L_∞ distance between their corresponding induced matrices \mathcal{D}^1 and \mathcal{D}^2 , namely $d_l(\mathcal{T}^1, \mathcal{T}^2) = \|\mathcal{D}^1 - \mathcal{D}^2\|_\infty$.

Notice that Definition 3.3 requires that the two given trees have the same labeling, so that they are both n -labeled merge trees.

Labeled merge tree of a valid matrix. We can also turn certain matrices back into trees. A symmetric matrix $\mathcal{D} \in \mathbb{R}^{n \times n}$ is called *valid* if $\mathcal{D}_{ii} \leq \mathcal{D}_{ij}$ for all i, j . A valid matrix encodes a function f on the complete graph K of n vertices modeled after n labels, with function value \mathcal{D}_{ii} on the vertex i and \mathcal{D}_{ij} on the edge ij .

The *labeled merge tree of a valid matrix* $\mathcal{D} \in \mathbb{R}^{n \times n}$, denoted $\text{MergeTree}(\mathcal{D})$, is the labeled merge tree of the complete graph with the induced function f . For a fixed $a \in \mathbb{R}$, let $K_a = f^{-1}(-\infty, a]$. Such a merge tree is obtained by tracking the connected components of K_a as a increases. See Fig. 2 (right) for an example and Sec. 4 for implementation details.

A valid matrix \mathcal{D} is called *ultra* if $\mathcal{D}_{ij} \leq \max\{\mathcal{D}_{ik}, \mathcal{D}_{kj}\}$. The induced matrix of a labeled merge tree is an ultra matrix; in fact, this construction induces a bijection between the space of labeled merge trees and the space of ultra matrices [30].

It has been shown that the interleaving distance defined above enjoys nice properties including stability [48]; in particular, two theoretical results from [30] that are central to our paper are included here for completeness. Theorem 3.1, which is a proposition from [30], lays the foundation for computing a 1-center for a set of labeled merge trees.

Theorem 3.1 (Proposition LMT 1-Center [30]). *Given n -labeled merge trees $\mathcal{T}^1, \dots, \mathcal{T}^k$ and their corresponding induced ultra matrices $\mathcal{D}^1, \dots, \mathcal{D}^k$, let \mathcal{D} be the element-wise 1-center of $\mathcal{D}^1, \dots, \mathcal{D}^k$. Set $\mathcal{T} = \text{MergeTree}(\mathcal{D})$. Then \mathcal{T} is a 1-center of the labeled merge trees $\mathcal{T}^1, \dots, \mathcal{T}^k$.*

Note that when using the L_∞ distance, the 1-center for the matrices is not unique. However, it is easy enough to compute one of these, which we will make use of in our implementation described in Sec. 4. The next theorem implies the existence of an animated morphing between trees, which is also used in our implementation.

Theorem 3.2 (Corollary LMT Geodesics [30]). *Given any two n -labeled merge trees \mathcal{T}^1 and \mathcal{T}^2 and their corresponding induced ultra matrices \mathcal{D}^1 and \mathcal{D}^2 , the family of merge trees $\{\mathcal{T}^\lambda := \text{MergeTree}(\mathcal{D}^\lambda) \mid \lambda \in [0, 1]\}$ defines a geodesic between \mathcal{T}^1 and \mathcal{T}^2 in the metric d_l where $\mathcal{D}^\lambda = (1 - \lambda)\mathcal{D}^1 + \lambda\mathcal{D}^2$.*

3.2 Leaf-Labeled Merge Trees

We will focus on a restricted class of labeled merge trees defined as follows. Recall $[n] := \{1, \dots, n\}$.

Definition 3.4. A *leaf-labeled merge tree* \mathcal{T} is a triple (T, f, ω) that consists of a merge tree (T, f) together with a labeling $\omega: S \rightarrow L$ that is surjective on the set of leaves L , where $S \subseteq [n]$.

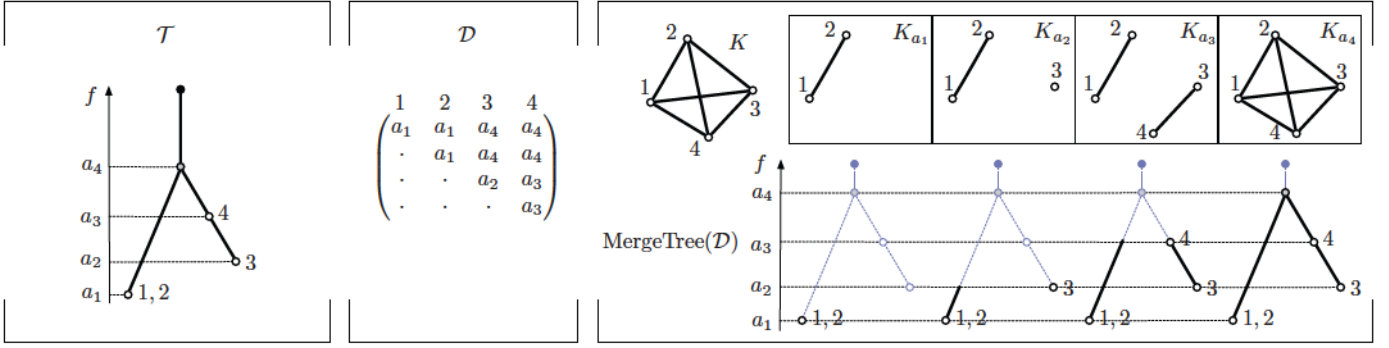


Fig. 2. An example of a labeled merge tree (left) that permits labels on internal vertices and multiple labels on a single leaf, together with its induced matrix (middle). The solid black vertex is the root, and white vertices are labeled. The tree has four nonroot vertices and two leaves. Right: construction of a labeled merge tree from the ultra matrix that recovers the original tree.

Unlike Definition 3.2, leaf-labeled merge trees allow labels only on the leaves. Also, unlike the traditional definition of a *leaf-labeled tree* where each of its leaves is labeled by precisely one element from a given label set [37], Definition 3.4 permits multiple labels on the same leaf. The set of leaf-labeled merge trees can be considered as a subset of labeled merge trees by considering ω as a restriction of μ to a subset of $[n]$. Up to reindexing the label set, the properties involving induced matrices (Theorems 3.1 and 3.2) and interleaving distance (Definition 3.3) therefore apply to leaf-labeled merge trees.

Full agreement, partial agreement and disagreement. In our implementation, we work with input data with various levels of labeling, so we discuss some terminology for various types of missing information next. First, an ensemble of leaf-labeled merge trees $\{(T_i, f_i, \omega_i)\}_{i=1}^k$ is in *full agreement* if the trees all share the same label set; that is, if the domain of ω_i is the same for all i . Second, an ensemble of leaf-labeled merge trees $\{(T_i, f_i, \omega_i)\}_{i=1}^k$ is in *partial agreement* if the trees are not in full agreement and $\bigcap_{i=1}^k S_i \neq \emptyset$. Third, an ensemble of leaf-labeled merge trees $\{(T_i, f_i, \omega_i)\}_{i=1}^k$ is in *disagreement* if $\bigcap_{i=1}^k S_i = \emptyset$. Ensembles of the second and the third type can also be treated as being *partially labeled* and *unlabeled*, respectively, although we rarely use these terminologies here.

3.3 Merge Trees from Scalar Field Ensembles

Of particular interest to visualization is merge trees that arise from an ensemble of scalar fields. Each tree captures the connectivity of the sublevel sets of a scalar field. In this paper, we refer to it as a *scalar field induced merge tree*.

Given a scalar function f defined on a topological space \mathbb{X} , $f: \mathbb{X} \rightarrow \mathbb{R}$, let $\mathbb{X}_a := f^{-1}(-\infty, a]$ denote the sublevel set of f for some $a \in \mathbb{R}$. Two points $x, y \in \mathbb{X}$ are *equivalent*, $x \sim y$, if they have the same function value, so $f(x) = f(y) = a$, and if they belong to the same component of the sublevel set \mathbb{X}_a . The quotient space \mathbb{X} with respect to the above equivalence relation, \mathbb{X}/\sim , is referred to as a scalar field induced merge tree to differentiate it from the more general Definition 3.1. We still refer to such a tree as a merge tree when it is clear from the context.

A scalar field induced merge tree is constructed by tracking the evolution of the components of \mathbb{X}_a as we vary the parameter a . Fig. 3 gives an example: leaves represent the creation of a component at a local minimum, internal vertices (saddles) represent the merging of components, and the root represents the entire space as a single component; see a straight line drawing in Fig. 3(c). We can further augment a scalar field induced merge tree with noncritical points, producing an *augmented merge tree* [15, Section 3.3]; otherwise it is *unaugmented*. In this paper, we always visualize an augmented merge tree by embedding its vertices (both critical and noncritical) inside the graph of f , as in Fig. 3(b).

3.4 Distance Between Vertices

Not only do we use distances between the trees themselves, but we also make use of a couple of different distances between vertices of a given merge tree for our implementation. Such distances impose a

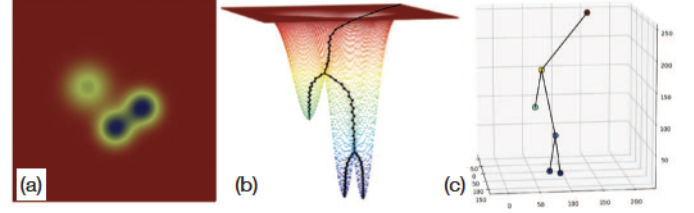


Fig. 3. (a) A 2D scalar field f is generated by a mixture of three Gaussian functions. It is visualized using a rainbow color map: red means high and blue means low values. (b) The graph of f , i.e., the set of all ordered pairs $(x, f(x))$, is visualized together with the corresponding augmented merge tree of f . (c) A straight line drawing of an (unaugmented) merge tree of f in 3D.

metric-space view of input trees and are useful both when handling merge trees with partial or no agreement (Sec. 4.2), and in helping to define consistency of vertices in Sec. 5.

Assume we are given a labeled merge tree $\mathcal{T} = (T, f, \omega)$. The first distance is an intrinsic distance on the merge tree T induced by the function f . Specifically, the *intrinsic tree distance between a pair of vertices* $x, y \in V$ is defined to be $d_T(x, y) = |f(x) - f(a(x, y))| + |f(a(x, y)) - f(y)|$. Note that this is exactly the path length between the vertices if we give every edge a weight equal to the difference in function values of its endpoints.

Now suppose, additionally, that \mathcal{T} has an embedding ι in \mathbb{R}^d , $\iota: |T| \rightarrow \mathbb{R}^d$ (in our experiments, $d = 2$ or 3). The second distance, the *Euclidean distance between a pair of vertices* $x, y \in V$, is the L_2 distance between their embeddings, $d_E(x, y) = \|\iota(x) - \iota(y)\|_2$.

4 COMPUTING 1-CENTERS OF LABELED MERGE TREES

Moving from theory to practice, we now discuss implementation details for computing a 1-center of leaf-labeled merge trees. We compute a structural average under three different scenarios: full agreement, partial agreement, and disagreement. For simplicity of explanation, we focus on leaf-labeled merge trees. Our algorithms and uncertainty encodings (Sec. 5) can be easily adapted to handle general merge trees with labeled internal vertices (Definition 3.2).

4.1 Full Agreement

We start with the simplest case: given an ensemble of leaf-labeled merge trees in full agreement, we compute its 1-center as a structural average. As specified by Theorem 3.1, the main idea is to compute an element-wise 1-center of their induced ultra matrices and convert the resulting valid matrix into a new merge tree.

We start with an ensemble of k leaf-labeled merge trees in full agreement, $\mathcal{T}^1, \dots, \mathcal{T}^k$, i.e., they all share the same label set $S \subseteq [n]$, with $|S| = s$. On a high level, our algorithm has three simple steps:

- F1. Represent each \mathcal{T}^i by its induced ultra matrix \mathcal{D}^i for $1 \leq i \leq k$.

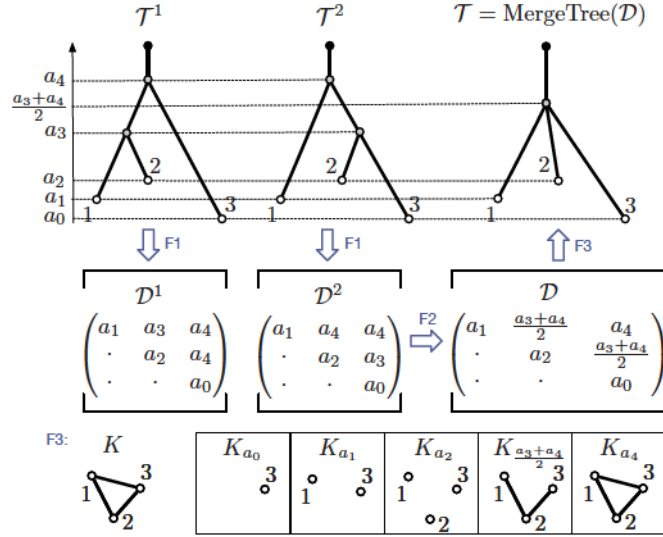


Fig. 4. Computing a 1-center for a pair of leaf-labeled merge trees in full agreement. The top row shows two input trees \mathcal{T}^1 and \mathcal{T}^2 (with three labeled leaves each) together with the output 1-center \mathcal{T} . The middle row displays the induced ultra matrices \mathcal{D}^1 and \mathcal{D}^2 together with their element-wise 1-center, matrix \mathcal{D} , which encodes a function f on the complete graph K . The bottom row shows K and its sublevel set filtration.

F2. Compute the element-wise 1-center \mathcal{D} of $\mathcal{D}^1, \dots, \mathcal{D}^k$.

F3. Turn \mathcal{D} back into a merge tree $\mathcal{T} = \text{MergeTree}(\mathcal{D})$.

An example for two leaf-labeled merge trees \mathcal{T}^1 and \mathcal{T}^2 in full agreement is given in Fig. 4.

Assume each tree $\mathcal{T}^i = (T_i, f_i, \omega_i)$ induces an ultra matrix $\mathcal{D}^i \in \mathbb{R}^{s \times s}$ indexed by the shared label set $S \subset [n]$. For step F2, we construct a choice of 1-center by computing the 1-center of each element,

$$\mathcal{D}_{ij} = \frac{1}{2} \left| \max\{\mathcal{D}_{ij}^\ell\}_{\ell=1}^k - \min\{\mathcal{D}_{ij}^\ell\}_{\ell=1}^k \right|. \quad (1)$$

Intuitively, \mathcal{D}_{ij} is the center of a minimum enclosing ball of points $\mathcal{D}_{ij}^1, \dots, \mathcal{D}_{ij}^k$ on a line, i.e., the midpoint between the smallest and the largest values along the line.

For Step F3, we turn \mathcal{D} back into a new merge tree $\mathcal{T} = \text{MergeTree}(\mathcal{D})$. Despite the fact that the matrices \mathcal{D}^ℓ are ultra matrices, the construction of Eq. (1) is not guaranteed to be ultra. However, it will be a valid matrix, and thus it encodes a function f on the complete graph K among n labels as vertices, $f: K \rightarrow \mathbb{R}$. Specifically, each vertex $v_i \in K$ has value $f(v_i) = \mathcal{D}_{ii}$; while each edge $e_{ij} = (v_i, v_j) \in K$ has function value $f(e_{ij}) = \mathcal{D}_{ij}$. That \mathcal{D} being a valid matrix guarantees that the function value of an edge is larger than or equal to that of each of its endpoint. The returned merge tree follows the standard merge tree construction described below.

Denote the sublevel set of K for $a \in \mathbb{R}$ by $K_a := f^{-1}((-\infty, a])$. The merge tree of f is defined on the vertex set of K and keeps track of the connected components in K_a . As a increases from $-\infty$ to ∞ , vertices in K creates new components in the tree, and edges in K either connect two vertices already from the same components, or merge existing components. We adapt Kruskal's algorithm that runs in time $O(|E|a(|V|))$ (a denotes the inverse of the Ackermann function), processes vertices/edges in a sorted order, and maintains connected components in a disjoint set data structure [28, 29] that supports fast component identification and merging [68].

By Theorem 3.1, \mathcal{T} is a 1-center of the leaf-labeled merge trees. Such a 1-center is not necessarily unique; nevertheless it serves as a justifiable structural average as it minimizes the maximum distance to any other tree in the set under d_f . Note that we allow multiple labels on the same leaf to ensure that this definition is well defined. See an example from Fig. 6 generated from our visualization system.

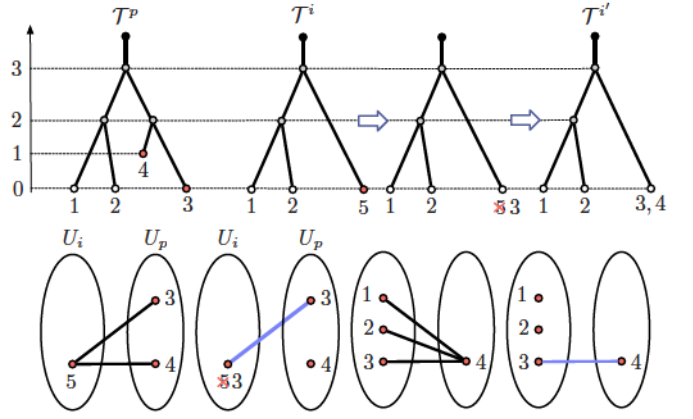


Fig. 5. Updating the labeling of \mathcal{T}^i against a pivot tree \mathcal{T}^p .

4.2 Partial Agreement

We now discuss strategies to turn incomplete information in terms of labeling into complete information so that the algorithm of Sec. 4.1 can be utilized. Given an ensemble of trees whose labels do not fully agree, finding the best ways to label the remaining vertices boils down to finding the best correspondences under the interleaving distance, which is unfortunately NP-hard [1]. We thus aim to develop a heuristic approach that is efficient and effective in practice for both the partial agreement and disagreement cases.

Assume we are given an ensemble of leaf-labeled merge trees $\mathcal{T}_1, \dots, \mathcal{T}_k$ in partial agreement. If S_i is the label set for tree \mathcal{T}_i , let $S = \bigcap_{i=1}^k S_i$ be the shared set of labels. For simplicity, we assume leaves do not have multiple labels for the input trees. On a high level, our algorithm is as follows:

- P1. Select a pivot tree $\mathcal{T}^p = (T_p, f_p, \omega_p)$ with the largest number of leaves among the input trees. Let S_p be its label set.
- P2. Convert each ensemble member $\mathcal{T}^i = (T_i, f_i, \omega_i)$ to a labeled tree $\mathcal{T}^{i'} = (T_i, f_i, \omega_i')$ by updating its labeling using the label set S_p combined with minimum weight matching.
- P3. Compute a 1-center of the trees $\mathcal{T}_1, \dots, \mathcal{T}_k$ using Sec. 4.1.

The key step is P2, for which we first give a pictorial toy example in Fig. 5 (top). The main idea is to study the structural similarities between unmatched leaves by comparing their distances (a linear combination of d_T and d_E) to the matched leaves and solving a minimum weight matching problem. We also give an example in Fig. 1 generated from our visualization system.

Algorithmic details. Recall that a *matching* in a bipartite graph is a set of the edges chosen in such a way that no two edges share an endpoint. A vertex is *matched* if it is an endpoint of one of the edges in the matching. Otherwise it is *unmatched*. A *maximum matching* is a matching of a maximum number of edges. An *assignment* (or a *minimum weight matching*) problem, in our setting, is the problem of finding, in a weighted bipartite graph, a maximum matching in which the sum of weights of the edges is as small as possible.

We describe our algorithm with the toy example given in Fig. 5. Given a pivot tree \mathcal{T}^p with a label set $S_p = \{1, 2, 3, 4\}$, we update the labeling for an ensemble member \mathcal{T}^i using S_p . The matched (shared) label set between the two trees is $S = \{1, 2\}$, and the unmatched label sets in \mathcal{T}^p and \mathcal{T}^i are $U_p = \{3, 4\}$ and $U_i = \{5\}$, respectively. Our goal is to assign new labels to U_i from the set U_p as follows.

First, build D_p to be a pairwise distance matrix with rows corresponding to unmatched labels U_p and columns corresponding to matched labels S . For $x \in U_p$ and $y \in S$, $D_p(x, y)$ is the distance between leaves $\omega_p(x)$ and $\omega_p(y)$. The matrix D_i is built similarly for tree \mathcal{T}_i . Note that both D_p and D_i have the same number of columns, but potentially a different number of rows. In our example, using the tree distance d_T , we have $D_p = \begin{pmatrix} 6 & 6 \\ 5 & 5 \end{pmatrix}$ and $D_i = \begin{pmatrix} 6 & 6 \end{pmatrix}$.

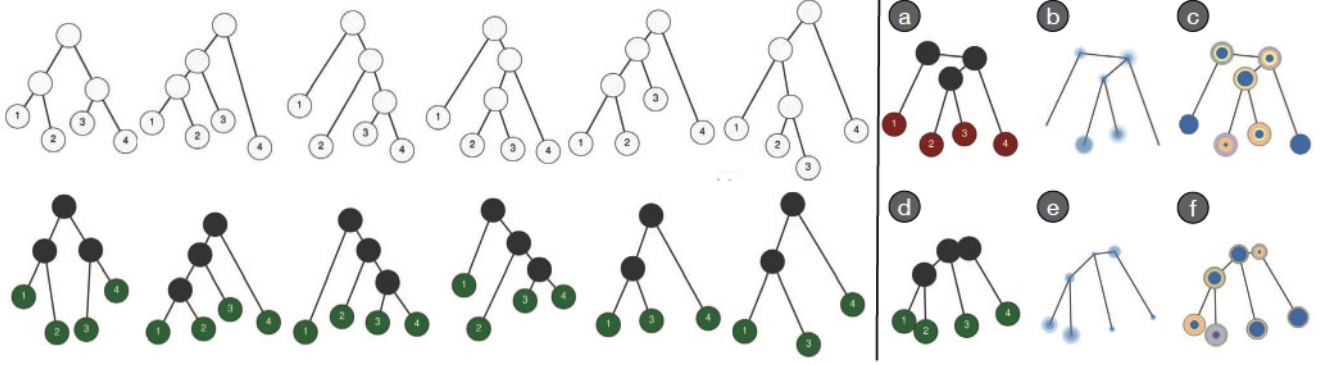


Fig. 6. Computing a 1-center for an ensemble of six leaf-labeled trees in full agreement (top row) and disagreement (bottom row). (a, d) 1-center tree; (b, e) variational vertex consistencies plots; (c, f) statistical vertex consistency plots. For the ensemble in disagreement, original labels are omitted; updated labels are green.

Second, construct a complete, weighted bipartite graph with vertex sets U_i and U_p where the weight c_{xy} between a label $x \in U_i$ and $y \in U_p$ is given by the L_2 distance between the rows from D_i and D_p , respectively. We find a minimum weight matching in this bipartite graph that gives an assignment of the rows of D_i (unmatched labels in an ensemble member) to the rows of D_p (matched labels in the pivot tree).

Formally, let b be a Boolean matrix where $b_{xy} = 1$ if and only if a leaf $x \in U_i$ is assigned to a leaf $y \in U_p$. Then the optimal assignment has a cost: $\min_{\mathbf{b}} \sum_{x,y} c_{xy} b_{xy}$. We then use the SciPy Python library to find a minimum cost matching, which employs the Hungarian algorithm [40]. Since \mathcal{T}^p was chosen to have the maximum number of leaves over all trees, $|U_i| \leq |U_p|$, and thus an optimal assignment will saturate all vertices in U_i . Thus, the output is a complete assignment $\eta : U_i \rightarrow U_p$ of minimal cost. We then define $\omega'_i(x) = \eta(x)$ for $x \in U_i$, and $\omega'_i(x) = \omega_i(x)$ otherwise.

At this point, we might still have unmatched labels U'_p in the pivot tree \mathcal{T}^p . We use a greedy algorithm to assign each such label to a leaf in the input tree based on its local structure similarity. In the example, we compute an updated matrix D'_p between the remaining unmatched label $U'_p = \{4\}$ and the matched labels $S' = \{1, 2, 3\}$, obtaining $D'_p = \begin{pmatrix} 5 & 5 & 3 \end{pmatrix}$. We then compute the pairwise distance matrix between all

matched labels for \mathcal{T}^i as $D'_i = \begin{pmatrix} 0 & 2 & 6 \\ 2 & 0 & 6 \\ 6 & 6 & 0 \end{pmatrix}$. The weight c'_{xy} between

a label $x \in S'_i$ and a label $y \in U'_p$ is again given by the L_2 distance between the rows from D'_i and D'_p , respectively. The leaf with the label x now obtains an additional label y when c'_{xy} achieves its minimum. In the toy example, the leaf in \mathcal{T}^i with a label 3 obtains an additional label 4. In other words, the labeling ω' is further updated by including the label 4 in its domain.

4.3 Disagreement

To compute a 1-center for an ensemble of leaf-labeled trees in disagreement, we again employ a labeling strategy that converts the trees to leaf-labeled trees in full agreement, and then apply the algorithm in Sec. 4.1. To find the labels, however, we now assume each tree $\mathcal{T}_i = (T_i, f_i, \omega_i)$ comes with a geometric embedding $t_i : |T_i| \rightarrow \mathbb{R}^d$. Our algorithm below is very similar to the partial agreement case; see Fig. 7 for an example.

- D1. Select a pivot tree \mathcal{T}^p with the largest number of leaves. Let S_p be its label set.
- D2. Update the initial labeling of each ensemble member \mathcal{T}^i by creating a new labeling using the label set S_p and a minimum weight matching. The updated tree is denoted as \mathcal{T}^i .
- D3. Compute a 1-center of the trees $\mathcal{T}^1, \dots, \mathcal{T}^k$ following Sec. 4.1.

In Step D2, for each ensemble member \mathcal{T}^i , we assume $S_i \cap S_p = \emptyset$ (otherwise, we follow the identical algorithm in Sec. 4.2). We create

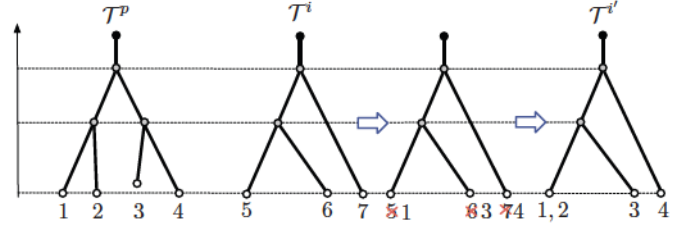


Fig. 7. Updating the labeling of \mathcal{T}^i against a pivot tree \mathcal{T}^p .

a weighted, complete bipartite graph between the label sets S_i and S_p , where the weight c_{xy} between a label $x \in S_i$ and a label $y \in S_p$ is their Euclidean distance in the embedded space, $c_{xy} = d_E(t_i(x), t_p(y))$. We again solve a minimum weight matching problem. The output is a complete assignment $\eta : S_i \rightarrow S_p$ of minimal cost. We then define $\omega'_i(x) = \eta(x)$ for $x \in S_i$. Since $|S_i| \leq |S_p|$, for any unmatched label $y \in S_p$, we follow a similar strategy as in Sec. 4.2, and ω'_i is updated accordingly.

5 ENCODING UNCERTAINTY

Given an ensemble of labeled merge trees $\mathcal{T}^1, \dots, \mathcal{T}^k$ and their 1-center \mathcal{T} , we work toward a visualization that highlights the structural consistency between each member \mathcal{T}^i and \mathcal{T} . To this end, we develop a novel measure of uncertainty via a metric space view of trees in the ensemble. This measure is also flexible, allowing a local-global tradeoff in understanding structure variations. In particular, we compute and visualize *vertex consistency* following a strategy based on Gaussian-weighted cosine similarity. Our consistency measures apply to general labeled merge trees as in Definition 3.2, not just leaf-labeled ones.

5.1 Computing Consistency for a Pair of Trees

Cosine similarity and weighted cosine similarity. Cosine similarity is a measure of similarity between two nonzero vectors that measures the cosine of the angle between them. Let $A = [A_1, \dots, A_m]$ and $B = [B_1, \dots, B_m]$ be two nonzero vectors of length m . Their cosine similarity is $\text{sim}(A, B) = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2} \sqrt{\sum_i B_i^2}}$. With respect to the parameter δ , we use a *Gaussian-weighted cosine similarity* given by

$$\text{sim}_\delta(A, B) = \frac{\sum_i \left(e^{-\frac{A_i^2 + B_i^2}{\delta^2}} A_i B_i \right)}{\sqrt{\left(\sum_i e^{-\frac{2A_i^2}{\delta^2}} A_i^2 \right)} \cdot \sqrt{\left(\sum_i e^{-\frac{2B_i^2}{\delta^2}} B_i^2 \right)}}. \quad (2)$$

Note that this similarity provides a “soft-threshold” for entries of A and B , so that entries smaller than $O(\delta)$ play a more important role in the similarity measure. In general, smaller values of the A_i ’s and B_i ’s are more important. If we remove all exponential factors in Eq. (2), then it becomes the standard cosine similarity.

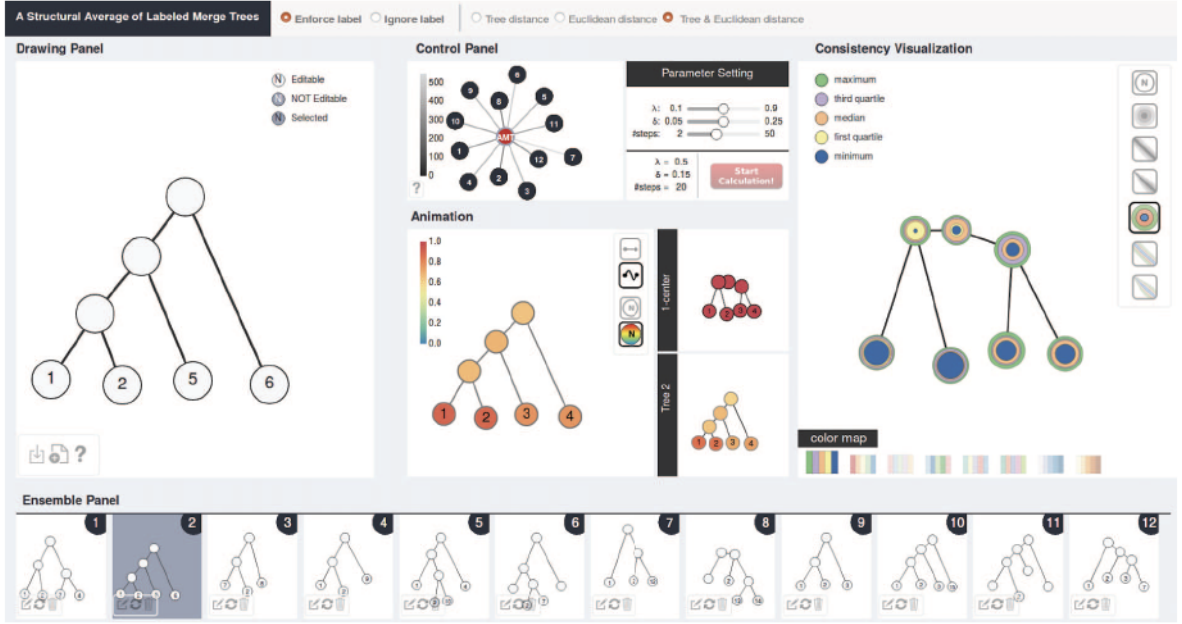


Fig. 8. User interface for the interactive visualization of labeled merge trees and their 1-center.

Vertex consistency. Given two labeled merge trees $\mathcal{T}^1 = (T_1, f_1, \mu_1)$ and $\mathcal{T}^2 = (T_2, f_2, \mu_2)$, we want to compute the consistency of vertices in \mathcal{T}^1 with respect to vertices in \mathcal{T}^2 . For simplicity, we assume that μ_1 and μ_2 are bijective on the set of vertices V_1 and V_2 , respectively (thus all vertices of \mathcal{T}^1 and \mathcal{T}^2 have unique labels). Let $|S_1| = |S_2| = s$ be the number of labels.

Let $V_1 = \{v_1, \dots, v_s\}$ and $V_2 = \{w_1, \dots, w_s\}$ be the labeled vertices for T_1 and T_2 , respectively, where v_i corresponds to w_i . Assume the chosen metrics between vertices in T_1 and T_2 are d_1 and d_2 , respectively, and fix a label l . We wish to measure the “consistency” of node v_l of T_1 with node w_l of T_2 . To do this, consider the two vectors $A = [d_1(v_l, v_1), \dots, d_1(v_l, v_s)]$ and $B = [d_2(w_l, w_1), \dots, d_2(w_l, w_s)]$. That is, A (resp. B) is the vector of distances from v_l (resp. w_l) to all other vertices (ordered by the labels). Intuitively, A (resp. B) summarizes how all other vertices relate to v_l (resp. to w_l) from a metric point of view. Hence, to measure the consistency between v_l and w_l , we use the Gaussian-weighted cosine similarity between these two vectors; that is, the *vertex consistency* between v_l and w_l is $\text{sim}_\delta(v_l, w_l) := \text{sim}_\delta(A, B)$. This consistency (similarity) value ranges from 0 to 1, and it is 1 if and only if $A = B$.

Intuitively, δ is a locality parameter: distances larger than, say, 3δ , will essentially be ignored, so this in effect is a soft thresholding where only nearby neighbors of v_l and w_l are being considered. By adjusting δ , we can change the neighborhood size. Setting $\delta = \infty$, we recover the traditional cosine similarity. The larger $\text{sim}(v_l, w_l)$ is, the more consistent vertices v_l and w_l are with each other.

We define the *vertex consistency* of \mathcal{T}^1 (with respect to \mathcal{T}^2) as a function defined on the vertices of \mathcal{T}^1 , $\alpha_1: V_1 \rightarrow \mathbb{R}$. For a vertex with a label l , $\alpha_1(v_l) = \text{sim}_\delta(v_l, w_l)$ for $v_l \in V_1$ and $w_l \in V_2$. $\alpha_2: V_2 \rightarrow \mathbb{R}$ is defined similarly.

5.2 Visual Encoding of Consistency for an Ensemble

We describe visual encodings for vertex consistencies for an ensemble member, variational consistencies, and statistical consistencies for the 1-center tree.

We encode vertex consistency for an ensemble member \mathcal{T}^i (with respect to the 1-center \mathcal{T}) using glyphs; see Fig. 9(left). Specifically, given a vertex consistency $\alpha_i: V_i \rightarrow \mathbb{R}$, the radius of each circular glyph at a vertex $v \in V_i$ scales proportional with $\alpha_i(v)$.

We encode variations in vertex consistencies for the 1-center tree \mathcal{T} using visual primitives inspired by [65, 66]; see Fig. 9(middle). For the 1-center tree with vertex set V , multiple vertex consistency functions are defined with respect to k ensemble members, $\alpha_1, \dots, \alpha_k: V \rightarrow \mathbb{R}$.

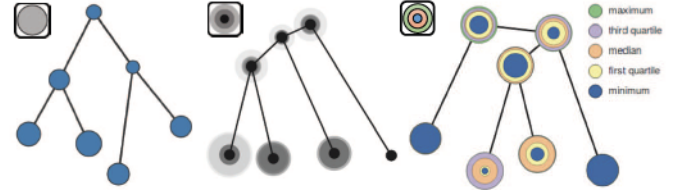


Fig. 9. Left: circular glyphs are used to encode vertex consistencies for an ensemble member. Graduated circular glyphs are used to encode variational (middle) and statistical (right) vertex consistencies for two 1-center trees.

Let $\bar{\alpha}$ be their mean value. Let $\alpha'_i := |\alpha_i - \bar{\alpha}|$, and we compute the sequence of variations $\{\alpha'_1, \dots, \alpha'_k\}$ as deviations from the mean $\bar{\alpha}$. The radius r_i of the i -th circular glyph is $r_i = \frac{g\alpha'_i}{2\bar{\alpha}}$ [66], where g is the desired spacing between glyphs, and α' is the maximum difference of any α_i to the mean of $\bar{\alpha}$ in the entire ensemble. The smaller the glyphs are, the more consistent the ensemble members are with respect to the 1-center. A small core indicates few outliers whereas a wider core indicates more deviation within members [66]. In our system, variational consistencies are rendered in a sequential colormap using a single hue.

Inspired by box plots, we visualize the distribution of vertex consistencies at the 1-center \mathcal{T} . For $\alpha_1, \dots, \alpha_k$ defined at a vertex $v \in V$ of \mathcal{T} , we compute their minimum, first quartile, median, third quartile, and maximum and apply graduated glyphs, as shown in Fig. 9(right). In our system, statistical consistencies are rendered using a miscellaneous colormap. Finally, the vertex consistency can be extended to edge consistency, see Appx. B for details.

6 INTERACTIVE VISUALIZATION AND USAGE SCENARIOS

We provide an interactive visualization system that takes as input an ensemble of leaf-labeled merge trees and outputs a 1-center tree as their structural average. The system incorporates 1-center computation, animation and uncertainty visualization; see the supplementary video for a demo and Appx. A for implementation and Appx. C for design details. It is implemented using D3.js, Ajax, Flask, and Python.

Its user interface is shown in Fig. 8. The **drawing panel** allows the user to draw individual merge trees using node-link diagrams and assign initial labels to the vertices. Each tree is then added to the **ensemble panel**, where ensemble members can be selected, deleted and edited. The **control panel** provides various options in computing

a 1-center tree, whereas an animated sequence between an input tree and the 1-center tree is provided within the **animation panel**. Various consistency measures are visualized in the **consistency visualization panel**; see Appx. C for design details.

We now describe how a user interacts with our visualization system under various usage scenarios. Thanks to *consistency*, a novel measure of uncertainty for vertices in a tree, our system helps the user perform tasks that were previously challenging. In particular, we can:

- Use the system as a structural calculator: it takes as input a set of leaf-labeled merge trees and outputs a 1-center tree as their structural average.
- Perform label diagnostics and correction to reduce data uncertainty and improve structural consistency.
- Understand structural similarities in a dynamic setting via animations from an ensemble member to the 1-center tree.

In addition, our system and its underlying algorithms allow us to study the tradeoff between local and global consistency measures in capturing structural similarities between ensemble members and their structural average, as well as to investigate heuristics in labeling strategies by exploring the tradeoff between intrinsic and extrinsic metrics (see Appx. D and Appx. E for examples).

6.1 Computing 1-Centers as Structural Averages

Suppose we have a numerical calculator. A typical usage scenario to obtain an average is to add a set of k input numbers and divide the sum by k . We want to perform similar operations, not with numbers, but with complex structures such as merge trees.

As illustrated in Fig. 1, when a user provides as input an ensemble of six leaf-labeled merge trees in partial agreement (top row on the left), the system first applies a (re)labeling strategy to update unmatched labels among the ensemble members while preserving the matched labels (bottom row on the left). For instance, *Tree 2* has three of its leaves relabeled based on their structural similarities to the pivot *Tree 1*. The labeling on the 1-center tree (a) shows leaf correspondences between the input and the output. The system incorporates uncertainty information on the 1-center tree: variations (b) and distributions (c) in vertex consistencies are encoded by graduated circular glyphs.

We explore the structural variations of the input ensemble via the summary plot (d), where all the ensemble members (in this particular input) have roughly the same interleaving distance to the 1-center (a). In the variational consistency plot (b), we see that there is a small variation in the vertex consistency measure around leaf 1 for the 1-center as indicated by a tiny blue circular glyph, which means that all six ensemble members are highly consistent in the local neighborhood of leaf 1. This is also captured by the statistical consistency plot (c) where the minimum, maximum, and medium consistency values coincide at leaf 1.

Using our visualization system, the user can perform computations and uncertainty visualization for trees in full agreement and disagreement in a similar fashion; see Fig. 6. For instance, the variational vertex consistency plot in Fig. 6(b) shows that all six input trees have almost no structural variation at leaves 1 and 4, whereas the distribution of vertex consistencies is much less concentrated at leaves 2 and 3 in the statistical consistency plot Fig. 6(c), which can be explained in the sense that leaf 1 and leaf 4 in each input tree all have a common lowest ancestor at the root with similar height values.

In summary, our system enables us to explore the structural variations between input trees and their 1-center by incorporating various means of consistency-based uncertainty visualization; see the supplementary video for a demo.

6.2 Label Diagnostics and Correction

Apart from the default functionality in computing a 1-center, perhaps more importantly, we could use our interactive tool to perform label diagnostics and correction to reduce data uncertainty among ensemble members and improve structural consistency.

We give an example in Fig. 10. The initial input provided by the user contains an ensemble of six leaf-labeled merge trees in partial

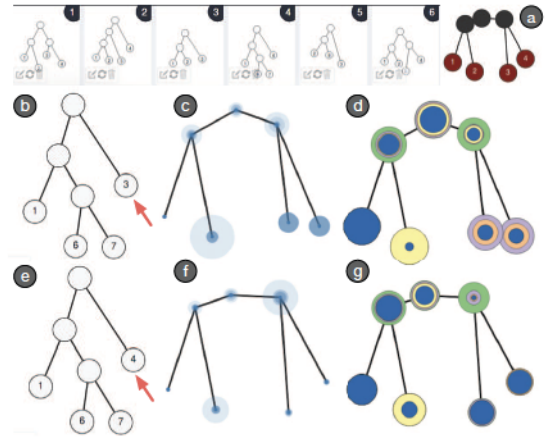


Fig. 10. Label diagnostics and correction. Given an input ensemble of six partially labeled merge trees, by changing a possibly inaccurate initial label from 3 (b) to 4 (e) in *Tree 4*, we reduce the structural variation for the 1-center, comparing (c) and (f).

agreement. Upon close inspection of the variational consistency plot, we notice that leaves 3 and 4 in the 1-center have high variations in their consistency (c). However, changing a (possibly inaccurate) label in *Tree 4* from 3 to 4 greatly reduces the variational consistency of the 1-center at leaves 2, 3, and 4, comparing (c) with (f). Meanwhile, such a correction also improves the statistical consistency, as shown in (g) where summary statistics (minimum, medium, etc.) coincide at these leaves; see the supplementary video for a demo. However, such a manual intervention is not practical for large merge trees. We believe the same strategy could be applied to a simplified merge tree on a coarser level. For larger trees, we leave it to future work to automatically identify outliers and suggest rules for modification.

6.3 Animation Along A Geodesic Path

Furthermore, we could understand structural similarities in a dynamic setting via animations. An animated sequence between an ensemble member and the 1-center shows how one deforms to the other via a geodesic (by Theorem 3.2), as well as the evolution of consistency measures at individual nodes through this process. As shown in Fig. 11, a 10-step animation between an ensemble member (a) and a 1-center (b) with different leaf sizes showcases the structural changes between them; e.g., the creation of new internal vertices along the geodesic. We can also observe the increase in vertex consistencies (red means high and green means low consistency), in particular, for vertex 3, along the same geodesic. See the supplementary video for a demo.

7 APPLICATION TO SCALAR FIELD ENSEMBLES

Our framework can be applied to study the structural variation of an ensemble of scalar fields. We give two illustrative examples. The first example, Fig. 12, contains a pair of scalar fields that gives rise to a pair of merge trees very similar to Fig. 4 in Section 4.1. Each scalar field is visualized as a 150×150 image in (a)-(b). We use the *Topology ToolKit* [72] to generate the augmented merge trees in (c)-(d).

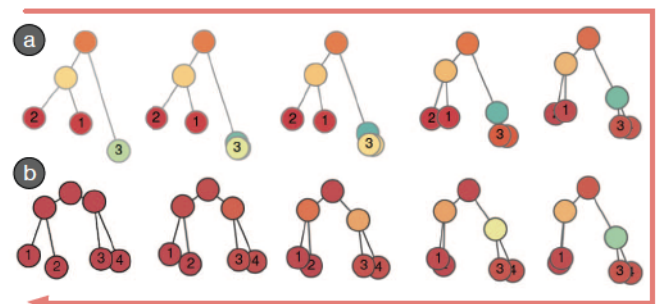


Fig. 11. An animated sequence between an ensemble member (a) and a 1-center tree (b) via a geodesic.

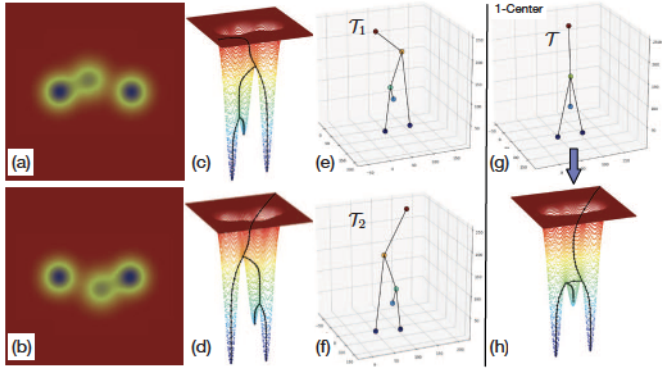


Fig. 12. Computing a 1-center for scalar field induced merge trees. On the left, each row contains from left to right: a scalar field visualized using a rainbow color map; an augmented merge tree visualized together with the graph of its corresponding scalar field; and a straight line drawing of a merge tree in \mathbb{R}^3 . On the right: (g) a straight line drawing of the 1-center tree in \mathbb{R}^3 ; (h) a reverse-engineered scalar field that gives rise to the 1-center tree.

Noncritical points are ignored for the 1-center computation as we need only leaves and their lowest common ancestors to compute the ultra matrices. We employ the labeling strategy described in Section 4.3 to find correspondences between the leaves, by taking into consideration their geometric embeddings.

The second example in Fig. 13 begins with an ensemble of three merge trees that arise from scalar fields generated by mixtures of Gaussians. We again apply our algorithm in Section 4.3 to compute the 1-center tree. In this example, merge trees \mathcal{T}_1 , \mathcal{T}_2 , and \mathcal{T}_3 contain 3, 4, and 4 leaves, respectively. According to Sec. 4.3, \mathcal{T}_2 (or \mathcal{T}_3) is selected as the pivot tree. We use the Euclidean distance between the geometric embeddings of leaves to assign labels. It is easy to verify that the computed 1-center tree \mathcal{T} is indeed of equal distance to the ensemble members \mathcal{T}_1 , \mathcal{T}_2 , and \mathcal{T}_3 , and therefore represents a structural average of the scalar field ensemble.

We now focus on uncertainty visualization using our encodings proposed in Sec. 5 and interactive visualization system in Sec. 6. As illustrated in Fig. 14, the vertex consistency plot (b) for \mathcal{T}_2 indicates that leaf 2 has low consistency against the 1-center tree; similarly in

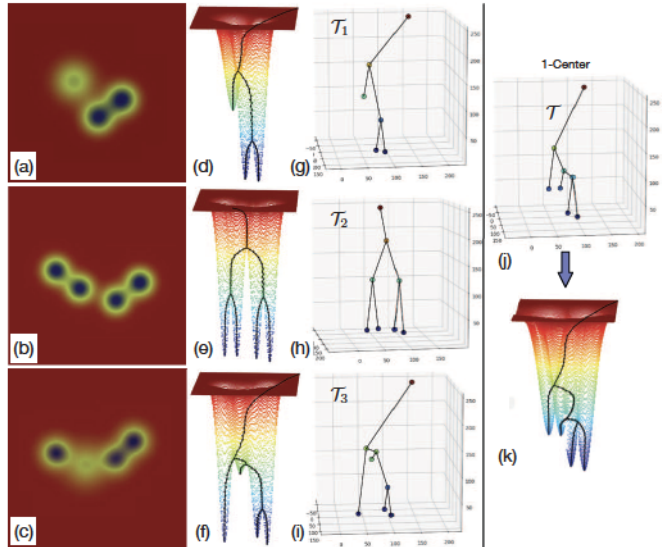


Fig. 13. Computing a 1-center for merge trees that arise from scalar fields generated by mixtures of Gaussians. Left: each row contains a scalar field visualized as an image, its augmented merge tree, and a straight line drawing of the corresponding (unaugmented) merge tree in \mathbb{R}^3 . Right: (j) 1-center tree; and (k) a reverse-engineered scalar field that gives rise to the 1-center tree.

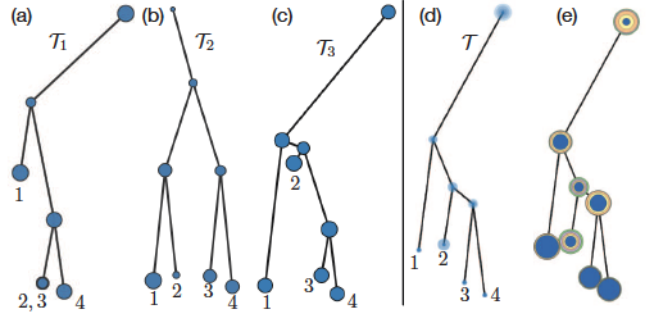


Fig. 14. (a)-(c): Vertex consistencies for each input tree. Leaf labels are inferred based on Euclidean distances among geometric embeddings of vertices. (d) Variational and (e) statistical vertex consistency plot for the 1-center tree.

the variational vertex consistency plot (d), leaf 2 in \mathcal{T} has the highest variation whereas leaves 1, 3, and 4 have low variations across ensemble members. This can be explained as all three input trees share similar local structures surrounding leaves 1, 3, and 4.

In addition to our proposed uncertainty visualization, we could further investigate the scalar field that gives rise to a 1-center tree. Such a scalar field is certainly not unique, and we have reverse-engineered a candidate by carefully positioning the critical vertices on the 2D domain with radius basis functions, as shown in Fig. 12(h) and Fig. 13(k), respectively. It is part of our on-going work to reverse-engineer *good* scalar fields that give rise to a given 1-center tree using optimization techniques with various constraints.

8 CONCLUSION AND DISCUSSION

We provide an interactive visualization system that computes and visualizes a structural average of an ensemble of leaf-labeled merge trees. We develop a novel measure of uncertainty, referred to as *consistency*, via a metric space view of the input trees. This measure is flexible in allowing a local-global tradeoff in understanding structure variations. Our results are the first steps toward statistical analysis of as well as uncertainty visualization for an ensemble of complex topological descriptors such as merge trees.

There are many future directions. From an algorithmic perspective, we are interested in computing a 1-center that is robust to outliers, or an effective algorithm to compute 1-mean/1-median. We also want to have a systematic investigation of various heuristic labeling algorithms. We are investigating ways to extend our framework to compute 1-center contour trees; as a contour tree of a function f can be constructed by carefully combining merge trees of f and $-f$. From a visualization perspective, we will evaluate the effectiveness of various visual encodings for consistency measures, as well as the visualization of 1-center trees at scale. We will also explore other applications of the animation between merge trees, e.g., in studying shape morphologies in computer graphs.

Finally, we hope that our visualization framework could help enhance topology- and geometry-based modern tools (e.g., [38, 43]) in neuron morphology analysis for biologists; see Appx. F for an example. We are also working on helping simulation scientists to take advantage of our framework in detecting abnormalities and outliers within simulation ensembles.

ACKNOWLEDGMENTS

YW was partially supported by NSF CCF-1740761, RI-1815697, DMS-1547357, and NIH R01EB022899. EM was partially supported by NSF CCF-1907591 and CMMI-1562012. LY and BW were partially supported by NSF DBI-1661375, IIS-1513616 and NIH 1R01EB022876. We thank participants of Dagstuhl Seminar 17292: Topology, Computation and Data Analysis in July 2017 for stimulating discussions on Reeb graphs. We are grateful to the Institute for Computational and Experimental Research in Mathematics (ICERM) for supporting us through the Collaborate@ICERM program in August 2018. Youjia Zhou created the reverse-engineered scalar fields in Figs. 12 and 13.

REFERENCES

- [1] P. K. Agarwal, K. Fox, A. Nath, A. Sidiropoulos, and Y. Wang. Computing the Gromov-Hausdorff distance for metric trees. *ACM Transactions on Algorithms*, 14(2):24:1–24:20, 2018.
- [2] C. L. Bajaj, V. Pascucci, and D. R. Schikore. The contour spectrum. *Proceedings of Visualization*, pp. 167–173, 1997.
- [3] U. Bauer, B. Di Fabio, and C. Landi. An edit distance for Reeb graphs. *Proceedings of the Eurographics Workshop on 3D Object Retrieval*, pp. 27–34, 2016.
- [4] U. Bauer, X. Ge, and Y. Wang. Measuring distance between reeb graphs. *Proceedings of the 30th Annual Symposium on Computational Geometry*, p. 464, 2014.
- [5] U. Bauer, C. Landi, and F. Memoli. The Reeb graph edit distance is universal. *Foundations of Computational Mathematics*, 2017.
- [6] U. Bauer, E. Munch, and Y. Wang. Strong equivalence of the interleaving and functional distortion metrics for Reeb graphs. In L. Arge and J. Pach, eds., *31st International Symposium on Computational Geometry*, vol. 34 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 461–475. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2015.
- [7] K. Beketayev, D. Yeliussizov, D. Morozov, G. Weber, and B. Hamann. Measuring the distance between merge trees. *Topological Methods in Data Analysis and Visualization III: Theory, Algorithms, and Applications, Mathematics and Visualization*, pp. 151–166, 2014.
- [8] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz. Overview and state-of-the-art of uncertainty visualization. In M. Chen, H. Hagen, C. Hansen, C. R. Johnson, and A. Kauffman, eds., *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*, pp. 3–27. Springer, 2014.
- [9] K. M. Brown, G. Barrionuevo, A. J. Canty, V. De Paola, J. A. Hirsch, G. S. Jefferis, J. Lu, M. Snippe, I. Sugihara, and G. A. Ascoli. The DIADEM data sets: representative light microscopy images of neuronal morphology to advance automation of digital reconstructions. *Neuroinformatics*, 9(2-3):143–157, 2011.
- [10] R. Brown. Animated visual vibrations as an uncertainty visualisation technique. *Proceedings of the 2nd international conference on Computer graphics and interactive techniques in Australasia and South East Asia*, pp. 84–89, 2004.
- [11] P. Bubenik, V. de Silva, and J. Scott. Metrics for generalized persistence modules. *Foundations of Computational Mathematics*, 15(6):1501–1531, 2014.
- [12] G. Cardona, A. Mir, F. Rosselló, L. Rotger, and D. Sánchez. Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf. *BMC Bioinformatics*, 14(1):3, 2013.
- [13] G. Carlsson, A. J. Zomorodian, A. Collins, and L. J. Guibas. Persistence barcodes for shapes. *Proceedings of the Eurographs/ACM SIGGRAPH Symposium on Geometry Processing*, pp. 124–135, 2004.
- [14] H. Carr and D. Duke. Joint contour nets: Computation and properties. *IEEE Pacific Visualization Symposium*, pp. 161–168, 2013.
- [15] H. Carr, J. Snoeyink, and U. Axen. Computing contour trees in all dimensions. *Computational Geometry*, 24(2):75–94, 2003.
- [16] M. Carrière and S. Oudot. Local equivalence and intrinsic metrics between Reeb graphs. In B. Aronov and M. J. Katz, eds., *33rd International Symposium on Computational Geometry*, vol. 77 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 25:1–25:15. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2017.
- [17] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, and S. Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the 25th annual symposium on Computational Geometry*, pp. 237–246, 2009.
- [18] H. Chen, H. Xiao, T. Liu, and H. Peng. SmartTracing: self-learning-based neuron reconstruction. *Brain informatics*, 2(3):135–144, 2015.
- [19] C. Collins, S. Carpendale, and G. Penn. Visualization of uncertainty in lattices to support decision-making. *Proceedings of Eurographics / IEEE VGTC Symposium on Visualization*, pp. 51–58, 2007.
- [20] V. de Silva, E. Munch, and A. Patel. Categorized Reeb graphs. *Discrete and Computational Geometry*, pp. 1–53, 2016.
- [21] V. de Silva, E. Munch, and A. Stefanou. Theory of interleavings on categories with a flow. *Theory and Applications of Categories*, 33(21):583–607, 2018.
- [22] S. Djuriclov, K. Kim, P. F. J. Lermusiaux, and A. Pang. Volume rendering data with uncertainty information. *Data Visualization*, 2001.
- [23] H. Edelsbrunner, J. Harer, V. Natarajan, and V. Pascucci. Morse-Smale complexes for piece-wise linear 3-manifolds. *Proceedings of the 19th Annual symposium on Computational Geometry*, pp. 361–370, 2003.
- [24] H. Edelsbrunner, J. Harer, and A. J. Zomorodian. Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete and Computational Geometry*, 30(87-107), 2003.
- [25] H. Edelsbrunner, D. Letscher, and A. J. Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28:511–533, 2002.
- [26] H. Edelsbrunner and E. P. Mücke. Simulation of simplicity: A technique to cope with degenerate cases in geometric algorithms. *ACM Transactions on Graphics*, 9:66–104, 1990.
- [27] C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger. Barrier trees of degenerate landscapes. *Zeitschrift für Physikalische Chemie*, 216(2), 2002.
- [28] Z. Galil and G. F. Italiano. Data structures and algorithms for disjoint set union problems. *ACM Computing Surveys (CSUR)*, 23(3):319–344, 1991.
- [29] B. A. Galler and M. J. Fischer. An improved equivalence algorithm. *Communications of the ACM*, 7(5):301–303, 1964.
- [30] E. Gasparovic, E. Munch, S. Oudot, K. Turner, B. Wang, and Y. Wang. Intrinsic interleaving distance for merge trees. *arXiv eprint ArXiv:1908.00063*, 2019.
- [31] S. Gerber and K. Potter. Data analysis with the Morse-Smale complex: The msr package for R. *Journal of Statistical Software*, 50(2), 2012.
- [32] R. Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45:61–75, 2008.
- [33] H. Griethe and H. Schumann. The visualization of uncertain data: Methods and problems. *Proceedings of SimVis*, pp. 143–156, 2006.
- [34] G. Grigoryan and P. Rheingans. Point-based probabilistic surfaces to show surface uncertainty. *IEEE Transaction on Visualization and Computer Graphics*, 10(5), 2004.
- [35] D. Günther, J. Salmon, and J. Tierny. Mandatory critical points of 2d uncertain scalar fields. *Computer Graphics Forum*, 33(3):31–40, 2014.
- [36] C. Heine, H. Leitte, M. Hlawitschka, F. Iurich, L. D. Floriani, G. Scheuermann, H. Hagen, and C. Garth. A survey of topology-based methods in visualization. *Computer Graphics Forum*, 35(3):643–667, 2016.
- [37] V. P. Johnson. *Enumeration Results on Leaf Labeled Trees*. PhD thesis, University of South Carolina, 2012.
- [38] L. Kanari, P. Dlotko, M. Sciamiero, R. Levi, J. Shillcock, K. Hess, and H. Markram. A topological representation of branching neuronal morphologies. *Neuroinformatics*, 16(1):3–13, 2018.
- [39] M. Kraus. Visualization of uncertain contour trees. *Proceedings of the International Conference on Information Visualization Theory and Applications*, pp. 132–139, 2010.
- [40] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.
- [41] B. Lee, G. G. Robertson, M. Czerwinski, and C. S. Parr. CandidTree: visualizing structural uncertainty in similar hierarchies. *Information Visualization*, 6(3):233–246, 2007.
- [42] Y. Lee, S. D. Barthel, P. Dlotko, S. M. Moosavi, K. Hess, and B. Smit. High-throughput screening approach for nanoporous materials genome using topological data analysis: Application to zeolites. *Journal of Chemical Theory and Computation*, 14(8):4427–4437, 2018.
- [43] Y. Li, D. Wang, G. A. Ascoli, P. Mitra, and Y. Wang. Metrics for comparing neuronal tree shapes based on persistent homology. *PLOS ONE*, 12(8):e0182184, 2017.
- [44] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Veldemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson. Extracting insights from the shape of complex data using topology. *Scientific Reports*, 3, 2013.
- [45] M. Mihai and R. Westermann. Visualizing the stability of critical points in uncertain scalar fields. *Computers & Graphics*, 41:13–25, 2014.
- [46] J. Milnor. *Morse Theory*. Princeton University Press, New Jersey, 1963.
- [47] D. Morozov, K. Beketayev, and G. Weber. Interleaving distance between merge trees. *Proceedings of Topology-Based Methods in Visualization (TopoInVis)*, 2013.
- [48] E. Munch and A. Stefanou. The ℓ^∞ -cophenetic metric for phylogenetic trees as an interleaving distance. In *Research in Data Science*, Association for Women in Mathematics Series, pp. 109–127. Springer International Publishing, 2019.
- [49] P. Oesterling, C. Heine, H. Jaenicke, 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(11):1547–1559, 2011.

- [50] A. T. Pang, C. M. Wittenbrink, and S. K. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997.
- [51] V. Pascucci and K. Cole-McLaughlin. Efficient computation of the topology of level sets. *Proceedings of the conference on Visualization*, pp. 187–194, 2002.
- [52] V. Pascucci, K. Cole-McLaughlin, and G. Scorzelli. Multi-resolution computation and presentation of contour trees. *Proceedings of the IASTED Conference on Visualization, Imaging, and Image Processing*, pp. 452–290, 2004.
- [53] H. Peng, Z. Ruan, F. Long, J. H. Simpson, and E. Myers. V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets. *Nature Biotechnology*, 28(4):348–353, 2010.
- [54] T. Pfaffelmoser, M. Mihai, and R. Westermann. Visualizing the variability of gradients in uncertain 2D scalar fields. *IEEE Transactions on Visualization and Computer Graphics*, 19(11):1948–1961, 2013.
- [55] T. Pfaffelmoser and R. Westermann. Visualization of global correlation structures in uncertain 2D scalar fields. *Computer Graphic Forum*, 31(3pt2):1025–1034, 2012.
- [56] T. Pfaffelmoser and R. Westermann. Correlation visualization for structural uncertainty analysis. *International Journal for Uncertainty Quantification*, pp. 171–186, 2013.
- [57] K. Potthow and H.-C. Hege. Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1393–1406, 2011.
- [58] K. Pöthkow, B. Weber, and H.-C. Hege. Probabilistic marching cubes. *Computer Graphics Forum*, 30(3):931–940, 2011.
- [59] K. Potter. Methods for presenting statistical information: The box plot. *Visualization of Large and Unstructured Data Sets, GI-Edition Lecture Notes in Informatics*, S4:97–106, 2006.
- [60] K. Potter, R. M. Kirby, D. Xiu, and C. R. Johnson. Interactive visualization of probability and cumulative density functions. *International Journal for Uncertainty Quantification*, 2(4):397–412, 2012.
- [61] K. Potter, J. Krüger, and C. J. Johnson. Towards the visualization of multi-dimensional stochastic distribution data. *Proceedings of the International Conference on Computer Graphics and Visualization*, 2008.
- [62] K. Potter, P. Rosen, and C. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. *Uncertainty Quantification in Scientific Computing*, 377:226–249, 2012.
- [63] T. Quan, H. Zhou, J. Li, S. Li, A. Li, Y. Li, X. Lv, Q. Luo, H. Gong, and S. Zeng. NeuroGPS-Tree: automatic reconstruction of large-scale neuronal populations with dense neurites. *Nature methods*, 13(1):51, 2016.
- [64] G. Reeb. Sur les points singuliers d’une forme de pfaff complètement intergrable ou d’une fonction numérique (on the singular points of a complete integral pfaff form or of a numerical function). *Comptes Rendus Acad. Science Paris*, 222:847–849, 1946.
- [65] J. Sanyal, S. Zhang, G. Bhattacharya, P. Amburn, and R. Moorhead. A user study to compare four uncertainty visualization methods for 1D and 2D datasets. *IEEE Transactions of Visualization and Computer Graphics*, 15(6):1209–1218, 2009.
- [66] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE Transactions of Visualization and Computer Graphics*, 16(6):1421–1430, 2010.
- [67] G. Singh, F. Mémoli, and G. Carlsson. Topological methods for the analysis of high dimensional data sets and 3D object recognition. In *Eurographics Symposium on Point-Based Graphics*, pp. 91–100, 2007.
- [68] D. Smirnov and D. Morozov. Triplet merge trees. In *Mathematics and Visualization*. Springer, 2017.
- [69] R. Sridharamurthy, T. B. Masood, A. Kamakshidasan, and V. Natarajan. Edit distance between merge trees. *IEEE Transactions on Visualization and Computer Graphics*, 2018.
- [70] S. Takahashi, Y. Takeshima, and I. Fujishiro. Topological volume skeletonization and its application to transfer function design. *Graphical Models*, 66(1):24–49, 2004.
- [71] J. Thomson, E. Hetzler, A. MacEachren, M. Gahegan, and M. Pavel. A typology for visualizing uncertainty. *Proceedings of SPIE, Visualization and Data Analysis*, 5669, 2005.
- [72] J. Tierny, G. Favelier, J. A. Levine, C. Gueunet, and M. Michaux. The topology toolkit. *IEEE Transactions on Visualization and Computer Graphics*, 24(1), 2018.
- [73] M. Van Kreveld, R. van Oostrum, C. Bajaj, V. Pascucci, and D. Schikore. Contour trees and small seed sets for isosurface traversal. *Proceedings of the 13th Annual Symposium on Computational Geometry*, pp. 212–220, 1997.
- [74] G. Weber, P.-T. Bremer, and V. Pascucci. Topological landscapes: A terrain metaphor for scientific data. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1416–1423, 2007.
- [75] R. Whitaker, M. Mirzargar, and R. Kirby. Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2713–2722, 2013.
- [76] W. Widanagamaachchi, A. Jacques, B. Wang, E. Crosman, P.-T. Bremer, V. Pascucci, and J. Horel. Exploring the evolution of pressure-perturbations to understand atmospheric phenomena. *IEEE Pacific Visualization Symposium*, 2017.
- [77] K. Wu and S. Zhang. A contour tree based visualization for exploring data with uncertainty. *International Journal for Uncertainty Quantification*, 3(3), 2013.
- [78] K. Wu and S. Zhang. Visualizing 2D scalar fields with hierarchical topology. *IEEE Pacific Visualization Symposium*, pp. 141–145, 2015.
- [79] H. Xiao and H. Peng. APP2: automatic tracing of 3D neuron morphology based on hierarchical pruning of a gray-weighted image distance-tree. *Bioinformatics*, 29(11):1448–1454, 2013.
- [80] B. Zehner, N. Watanabe, and O. Kolditz. Visualization of gridded scalar data with uncertainty in geosciences. *Computers & Geosciences*, 36(10):1268–1275, 2010.
- [81] W. Zhang, P. K. Agarwal, and S. Mukherjee. Contour trees of uncertain terrains. *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 43, 2015.