





Fast Comparative Analysis of Merge Trees Using Locality Sensitive Hashing

Weiran Lyu , Raghavendra Sridharamurthy , Jeff M. Phillips , and Bei Wang 

Abstract—Scalar field comparison is a fundamental task in scientific visualization. In topological data analysis, we compare topological descriptors of scalar fields—such as persistence diagrams and merge trees—because they provide succinct and robust abstract representations. Several similarity measures for topological descriptors seem to be both asymptotically and practically efficient with polynomial time algorithms, but they do not scale well when handling large-scale, time-varying scientific data and ensembles. In this paper, we propose a new framework to facilitate the comparative analysis of merge trees, inspired by tools from locality sensitive hashing (LSH). LSH hashes similar objects into the same hash buckets with high probability. We propose two new similarity measures for merge trees that can be computed via LSH, using new extensions to Recursive MinHash and subpath signature, respectively. Our similarity measures are extremely efficient to compute and closely resemble the results of existing measures such as merge tree edit distance or geometric interleaving distance. Our experiments demonstrate the utility of our LSH framework in applications such as shape matching, clustering, key event detection, and ensemble summarization.

Index Terms—Merge trees, locality sensitive hashing, comparative analysis, topological data analysis, scientific visualization

1 INTRODUCTION

Measuring the similarity between objects is fundamental in data analysis. Particularly, it is important to quantify the proximity of objects to one another when they are notably similar, whereas measuring the distance between significantly dissimilar objects is often of lesser concern. A measure of similarity is the key to identifying repeated patterns, retrieving similar objects, building data clustering, and performing nearest neighbor search. It can also be used as a generalized inner product for kernel methods for a variety of embedding, classification, and regression tasks. Choosing a meaningful similarity measure that can be computed efficiently is critical for performing advanced data analysis on any dataset, especially for visual analytics that require real-time and interactive feedback.

The need for meaningful and efficient similarity measures is especially true for scientific data analysis and visualization. In scientific computing, a large number of numerical simulations yield data in the forms of scalar fields, for example, temperature and surface atmospheric pressure from the Weather Research and Forecasting model. Moreover, the predominate way to facilitate efficient storage, analysis, and visualization of scalar fields is through various topological descriptors—from merge trees [13] to Morse–Smale complexes [23]—that represent the salient features of the underlying scientific phenomena.

To that end, a rich set of comparative measures is available for a number of topological descriptors, with applications in structural biology, climate science, combustion studies, neuroscience, computational physics and chemistry, and ecology. A key takeaway from a survey by Yan et al. [81] is that choosing the right similarity measure between topological descriptors is a recurring challenge because it tends to be both data-driven and dependent on specific applications. Rarely does a single measure fulfill all desired criteria, including acting as a metric or pseudometric, exhibiting stability and discriminative power, and being easy and efficient to compute.

The issue of computational efficiency is especially important in studying time-varying scientific data and ensembles at scale. For time-varying scalar fields, similarity measures between successive time steps are employed to identify periodic patterns, significant events, and

anomalies, as well as to facilitate feature tracking (e.g. [47, 56, 62, 63]). Computational efficiency is particularly notable, for instance, for tracking the evolution of extreme weather events (e.g., thunderstorms and hurricanes) using time-varying reanalyzed data such as temperature, wind, and moisture (e.g. [72]). For ensembles, similarity measures aid in the identification of clusters, outliers, and distinctive ensemble members (e.g., [34, 61, 82]). Efficient similarity measure holds particular significance in the examination of climate simulation ensembles, where thousands of climate model simulations with slight variations in parameter settings are utilized for climate projections. However, many similarity measures for topological descriptors suffer from challenges in both efficiency and scalability. While the complexity to compute the distance between a pair of merge trees of size n is at best $O(n^2)$ [81], in practice, these methods involve solving matching problems, resulting in high runtime. Furthermore, in the case of ensembles or time-varying data, any useful analysis often requires a large number of comparisons, potentially involving all-pairs comparison in the worst case.

The broader data analysis community has turned to locality sensitive hashing, or LSH [18, 38], to address these questions of efficiency and scalability. LSH is more flexible than embedding methods which require intermediate vector representations, and it is less specific than clustering methods which typically force fixed groupings. LSH uses random hash functions, not to index distinct objects, but to randomly group together similar objects. Each object is given a set of representative signatures through a random process. These random signatures match between objects proportionally to how similar the objects are. This property induces a mechanism, whereby objects can be allocated to (multiple) hash buckets according to their signatures, enabling highly efficient probabilistic retrieval of similar objects by exploiting the tendency for them to reside within the same hash bucket. Importantly, LSH avoids comparing all objects.

Although LSH has become a fundamental tool in most large-scale data analysis, it remains unexplored in topological data analysis and visualization. As datasets continue to grow, it is paramount to investigate how to extend LSH to comparing topological descriptors at scale. This paper focuses on LSH for merge trees. The merge tree captures topological relations between sublevel sets of a scalar field, and is shown to be quite useful in symmetry detection [54], shape matching and retrieval [62], feature tracking [56], summarization [62, 82], interactive exploration [49], and uncertainty visualization [82]. As most of the LSH algorithms require labeled structures, we use labeled merge trees [80, 82]. A labeled merge tree is applicable when there is a natural labeling for the nodes, or when a labeling may be inferred from the data [40]. A climate simulation ensemble produces a set of slightly varying scale fields (e.g., pressure) that give rise to slightly different merge trees with a shared domain. We may use the indices of mesh

• Weiran Lyu, Raghavendra Sridharamurthy, Jeff Phillips, and Bei Wang are with the University of Utah. E-mail: wlyu@sci.utah.edu, g.s.raghavendra@gmail.com, jeffp@cs.utah.edu, beiwang@sci.utah.edu.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxx/TVCG.201x.xxxxxx

nodes or the correspondences between underlying features (e.g., hurricane eyes) as the labeling, which is useful for feature tracking using merge trees (e.g., [79]). A labeled merge tree where labels encode geometric features (e.g., Euclidean coordinates or other node attributes) further enables geometry-aware comparisons of merge trees [80].

Contributions. In this paper, we provide a LSH framework to facilitate the comparative analysis of labeled merge trees. Our framework is an adaptation of existing methods like Recursive MinHash [20] and subpath signature [78] with critical extensions to our setting. The key contributions are as follows:

- We propose two new similarity measures for labeled merge trees that can be computed via LSH, using new extensions to Recursive MinHash and subpath signature, respectively.
- Our similarity measures are extremely efficient to compute and closely resemble the results of existing measures such as merge tree edit distance or geometric interleaving distance.
- We develop efficient and scalable algorithms for our LSH framework in comparing labeled merge trees.
- Our experiments demonstrate the utility of our LSH framework in applications such as shape matching, temporal scalar field and ensemble summarization, and in identifying transitions between data structures of time-varying datasets.
- We compare with existing methods in terms of accuracy and scalability. Our methods achieve $10 - 30\times$ speed-up on moderate datasets and high speed-up on a large ensemble.

Overall, our framework is the first of its kind in integrating the notion of LSH within topological data analysis and visualization. It demonstrates good efficiency in comparative analysis of merge trees at scale. Our LSH framework may be extended to handle other topological descriptors, in particular, extremum graphs [47] and contour trees [13].

2 RELATED WORK

Merge trees. We mainly focus on merge trees and labeled merge trees; see the surveys [37, 81] for other topological descriptors.

Merge trees capture the topology of sublevel sets of a real-valued function. They appear as an intermediate step in constructing contour trees [13]. In the past few decades, contour trees and merge trees have been used in various applications such as excess topology removal from isosurfaces [73], image analysis [45], topology controlled volume rendering [67], flexible isosurface generation [14], seed selection for segmentation [39], high-dimensional data analysis [48], uncertainty data exploration [74], cavity identification in biomolecules [5], symmetry detection [65], segmentation of volumetric data [7], and analysis of astronomical data [53]. Multiple methods exist to compute contour trees/merge trees in both serial and parallel; see [2, 13, 15, 16, 31–33, 68].

Labeled merge trees were first defined by Gasparovic et al. [28], followed by Yan et al. [82] and subsequently used by Yan et al. [80] in comparing time-varying scalar fields. The nodes of a merge tree are labeled based on their function values or geometrical properties.

Comparative analysis of topological descriptors. Comparison measures for topological descriptors such as merge trees and contour trees have to incorporate the structure along with the information about the scalar fields. Morozov et al. [46] introduced interleaving distance between merge trees, a stable and discriminative distance but without an efficient algorithm to compute it. Later Beketayev et al. [6] introduced branch decomposition distance, which considers all possible branch decompositions. Sridharamurthy et al. [62] introduced global merge tree edit distance for ordered and unordered trees extending tree edit distances, and provided an efficient algorithm to compute it. Sridharamurthy and Natarajan further extended it to local merge tree edit distance [63] that enables a hierarchical comparison of merge trees.

Gasparovic et al. [28] and Yan et al. [82] defined intrinsic interleaving distance for labeled merge trees, provided an algorithm to compute it with numerous applications. Pont et al. [50] introduced Wasserstein distance between merge trees with the facility to compute barycenters and provided applications to ensemble data. Yan et al. [80] introduced geometry-aware interleaving distance extending the intrinsic interleaving distance by incorporating geometric information to enhance its ap-

plicability to scientific data. Wetzel et al. [71] followed by Wetzel and Garth [70] introduced branch-decomposition independent edit distance and a deformation-based edit distance, both for comparing branch-decomposition of merge trees. Bollen et al. [8] introduced a stable edit distance addressing instability issues in previous edit distance-based measures. Wetzel et al. [69] eliminated the horizontal instability of edit distance-based measures by providing linear programming (LP) formulation of unconstrained edit distance. Qin et al. [52] hashed persistence diagrams into binary codes using a generative adversarial network to speed up comparisons. While hashing is also used in our framework, we do not need any learning-based approach to generate representations of the merge trees. There is another set of comparison measures (such as those based on histograms [55, 56] and the extended branch decomposition [54]). They are not metrics by definition but are simple, intuitive, and easy to compute.

Locality sensitive hashing. Hashing is the process of transforming data to values, oftentimes of fixed size, via a hash function. It has been effective in compressing data for fast access and comparison; see [21] for a survey. Locality sensitive hashing (LSH) utilizes a family of hash functions that map similar objects to the same hash buckets with high probability, making it useful for clustering and nearest neighbor search. Given a set of objects, LSH creates a sequence of discrete representatives for each object; the more these representatives match, the more similar the objects are deemed to be. Since the algorithm generates more representatives, it refines the notion of similarity, at the expense of higher computation cost. The basics of LSH were first introduced by Indyk and Motwani [38], followed by Gionis et al. [29]. Charikar [18] and later Chierichetti et al. [22] provided a more theoretical foundation along with necessary conditions for the existence of a LSH method for any similarity measure.

Wu et al. [76] provided a survey of LSH methods. These methods focus on evaluating a base measure—most commonly Jaccard similarity (MinHash), cosine similarity (SimHash), or string edit distance (SED). Ertl [25] provided an LSH framework for probabilistic Jaccard Similarity called ProbMinHash. Many LSH frameworks for SED have been defined, including by Zhang and Zhang [83–85], Marçais et al. [43], Chen and Shao [19], approximating similarity search under SED by Mccauley [44].

Coming to structured data such as sequences, trees, and graphs, Wu and Li [75] provided a survey of existing LSH methods. We discuss two relevant ones: hierarchical and kernel-based methods. Gollapudi and Panigrahy [30] proposed the idea of combining two min-hashes to enhance LSH for hierarchical structures like trees. Chi et al. [20] provided Recursive MinHash (RMH) for hierarchical structures by repeatedly using MinHash at each level of the hierarchy, building the hash representations from the bottom up.

Kernel-based methods generate vectorized representations of graphs while still being able to capture their structure. The most prominent kernel method for representing graphs and trees is the Weisfeiler-Lehman (WL) kernel by Shervashidze et al. [59]. Li et al. [41] defined Nested Subtree Hashing (NSH) which is an improvement over WL kernels, but still takes $O(n^2)$ time. Wu et al. [77] defined K -ary Tree Hashing (KATH). All three methods, WL, NSH, and KATH kernels, although applicable to trees, are more suited for general graph structures with cycles. Aioli et al. [3] defined the Subset Tree (SST) kernel, which also takes $O(n^2)$ time. Tatikonda and Parthasarathy [64] defined a kernel for trees based on pivots and least common ancestors, but the computation of the kernel takes $O(n^2)$ time. Shin and Ishikawa [60] and Xu et al. [78] defined subpath signature for trees to capture hierarchical relationships, providing an $O(n)$ signature construction. Signatures which asymptotically take $O(n^2)$ time are still useful for graphs, since graph comparison is a difficult problem to solve; however trees require signatures that are subquadratic to be of any use in practice. Hashing techniques have been used for other applications such as graph alignment [36] and relational graph matching [42].

In terms of LSH methods, the upper bound and lower bound proofs for tree edit distance (TED) [26, 27] along with the LSHability results [22] showed that we cannot have an LSH method for TED. Thus, we design an LSH method that emulates Jaccard similarity with some

hierarchical information encoded so that the tree structure can still be incorporated. We concentrate on extending hierarchical methods by Chi et al. [20] and some of the fast kernel methods such as subpath signature by Xu et al. [78] for their simplicity in implementation and potential for speed-ups.

3 BACKGROUND

3.1 Merge Trees and Labeled Merge Trees

Merge trees. Given a scalar field $f : \mathbb{X} \rightarrow \mathbb{R}$ defined on a topological space \mathbb{X} , the merge tree of the data (\mathbb{X}, f) captures the connectivity of its sublevel sets $f^{-1}(-\infty, t]$ (for some $t \in \mathbb{R}$). Mathematically, we identify two points $x, y \in \mathbb{X}$ to be equivalent (i.e., $x \sim y$) if $f(x) = f(y) = t$ and they belong to the same connected components of $f^{-1}(-\infty, t]$. The *merge tree* is the quotient space \mathbb{X}/\sim . The root of a merge tree is the global maximum, the leaves are the local minima, and the internal nodes are the merging saddles. We denote a merge tree by T , with a set of nodes $V(T)$ and a set of edges $E(T)$.

Labeled merge tree. A labeled merge tree consists of a merge tree T with a map $\pi : [n] \rightarrow V(T)$ that is surjective on the set of leaves [28, Def. 2.2], that is, all leaves and some internal nodes are labeled. Here, $[n] := \{1, 2, \dots, n\}$ denotes a set of labels. Fig. 1 shows merge trees and labeled merge trees for a 2D scalar field.

A node can have multiple labels. Labels can be added from one labeled merge tree to another unlabeled merge tree using a variety of methods [80]: *tree mapping* considers topological information captured by how the function values differ along the path containing the least common ancestor; *Euclidean mapping* considers geometric information captured by the Euclidean distance between the critical points; and *hybrid mapping* combines the two.

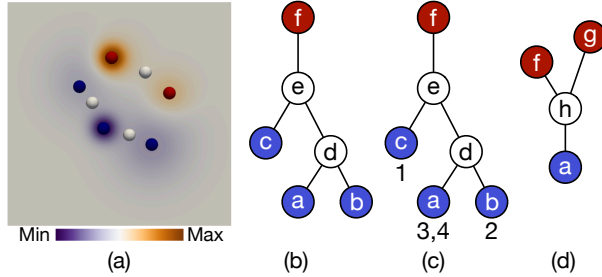


Fig. 1: Illustration of merge trees. (a) A scalar field f overlaid with critical points. (b) Merge tree. (c) Labeled merge tree with node labels $[3, 4]$, $[2]$, and $[1]$ for nodes a , b , and c , respectively. (d) Merge tree of $-f$.

Merge tree edit distance (d_E). Given two merge trees T_1 and T_2 , let $\mathcal{Q} := \mathcal{Q}(T_1, T_2)$ denote a set that contains sequences of edit operations (insert, delete, relabel) that transform T_1 to T_2 . $\gamma(Q)$ denotes the cost incurred over a sequence $Q \in \mathcal{Q}$; it is the sum of the cost of individual edit operations. Now, d_E is given by $d_E(T_1, T_2) = \min_{Q \in \mathcal{Q}}(\gamma(Q))$.

Geometry-aware interleaving distance (d_I). For T with a labeling π that incorporates both geometric and topological information, and a scalar function f , its *induced matrix* M consists of entries defined as $M_{ij} = f(\text{lca}(\pi(i), \pi(j)))$ where *lca* stands for the *lowest common ancestor*. Given two labeled merge trees T_1 and T_2 , d_I is given by the *cophonetic metric* (i.e., p -th norm) [12] between the induced matrices M_1 and M_2 , i.e., $d_I(T_1, T_2) = \|M_1 - M_2\|_p$. We use $p = \infty$.

3.2 Locality Sensitive Hashing

We define LSH and its variants that are relevant to this work. For a more detailed introduction, see the survey by Wu et al. [76].

Locality sensitive hashing. An LSH algorithm considers a space of objects \mathcal{Z} and a similarity function $s : \mathcal{Z} \times \mathcal{Z} \rightarrow [0, 1]$. It then uses a family of hash function \mathcal{H} , so that in expectation

$$\mathbb{E}_{h \sim \mathcal{H}}[h(p) = h(q)] = s(p, q).$$

In other words, given the randomness in the choice of harsh functions $h \sim \mathcal{H}$, the probability that $h(p) = h(q)$ equals their similarity $s(p, q)$ [38]. Therefore, the greater the similarity between objects p and q , the higher the probability that they will hash together. This is in

contrast to traditional hashing used for indexing a set, where we aim to avoid collisions between objects.

Then we bundle a set of hashes to amplify their effect, so objects more similar than a threshold collide with high probability. We harness this property to compute similarity or find similar objects without explicit comparison or search.

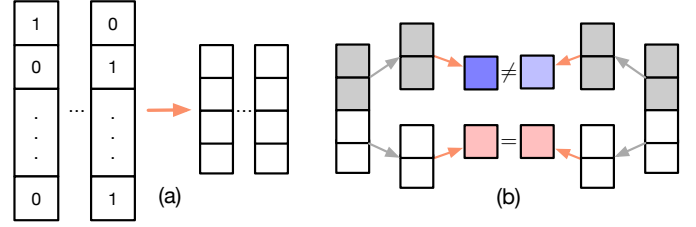


Fig. 2: (a) Generating MinHash: an $N \times M$ binary matrix is generated and then reduced to a dense $q \times M$ signature matrix using q -MinHash. Here, $q = 4$. (b) LSH: a signature is divided into bands of length r . Here, $r = 2$. LSH is applied to individual bands and candidate pairs are determined based on signature collision in any band (in red). The orange arrows are applications of hash functions.

MinHash. A MinHash is a common mechanism to create an LSH when each object $Z \in \mathcal{Z}$ is represented as a subset $Z \Rightarrow W \subset U$; where $U = \{u_1, \dots, u_N\}$ is some universal set of possible objects. The base element of a MinHash is a (random) permutation over $\sigma : U \rightarrow U$. However, the actual hash h returns only the minimum $h(W) = \arg \min_{w \in W} \sigma(w)$. As a result, we typically do not implement this as a full permutation (which can be expensive to store and compute), but σ can map $U \rightarrow \mathbb{R}$, so the minimum object can still be retrieved.

A classic insight [10] is that for a MinHash h on two sets $V, W \subset U$,

$$\mathbb{E}_{h \sim \mathcal{H}}[h(V) = h(W)] = \text{Jac}(V, W),$$

where $\text{Jac}(V, W) = |V \cap W|/|V \cup W|$ is the Jaccard similarity.

Therefore, to use a MinHash for any family of objects \mathcal{Z} , the key is to represent each $Z \in \mathcal{Z}$ into sets W of potential “views” or structured “subset-elements” of those base objects Z . Then, we can directly invoke MinHash to define a similarity of those base objects via the Jaccard similarity of their sub-objects.

We combine q MinHash functions h_1, \dots, h_q into a single hash function called an q -MinHash. For a set W , it concatenates these q signatures into a longer ordered set $[h_1(W), h_2(W), \dots, h_q(W)]$; see Fig. 2 for an illustration.

LSHability [18, 22] provides criteria for the existence of an LSH framework for a particular similarity measure. For a similarity s to admit an LSH framework, there are two necessary conditions:

1. $1 - s$ must be a metric distance function;
2. $1 - s$ must be isometrically embeddable in l_1 .

For an LSH algorithm to exist, both the distortion lower bound and upper bound should be 1 according to [22]. Garofalakis and Kumar proved that this is not the case for tree edit distance [26, 27]. Thus, in this work, we operate with the Jaccard similarity variants.

Amplifying similarity for similarity search. A single hash that has collisions proportional to object similarity by itself would make for a very noisy similarity search. For objects with a very high similarity (e.g., 0.9), they do not collide 10% of the time. To address this issue, we use a set of k randomly chosen hash functions $h_1, \dots, h_k \sim \mathcal{H}$ and employ *banding*.

We divide the k hash functions into b bands each with $r = k/b$ hash functions. A band is considered to have a collision for two objects only if *all* r hash functions in the band show a collision. In other words, these r -banded hash functions are combined into a single r -MinHash function. Such a process can be implemented efficiently by concatenating the hash values from each h_i in the band, and using a regular hash table to increase the specificity, thereby making it harder for objects to collide.

However, we counterbalance this increased specificity by decreasing it through b bands, where two objects are deemed to collide if they

collide in *any* band. Together, these adjustments sharpen the threshold for identifying a collision. Consequently, pairs with $s(p, q) = 0.9$ are much more likely to be identified as being similar, and pairs with $s(p, q) = 0.1$ are much less likely to be considered similar.

For a fixed k , we can adjust the desired specificity and the similarity threshold by changing r . In our case, we find that the hash function design is already very specific, so we use a very small r . Fig. 2 provides an illustration of conceptual steps of how LSH is used to identify similar objects efficiently. Objects W_i and W_j are considered similar when their hashed signatures collide in any band.

3.3 Subpath Signature for LSH

For graph-structured data, a common way to encode labeled graphs into sets W from some universal set U is to consider all subpaths. Given a parameter t , we set W as all sequences of t labels which can be formed by a path in the graph. The representation of an object as a set W means it can immediately be used in the MinHash framework. For rooted trees, Xu et al. [78] showed that it is effective to consider only directed paths from the root toward the leaves, without including the root itself.

A rooted-tree subpath signature is of interest because it not only is efficient but also provides theoretical bounds w.r.t. to the tree edit distance d_E with unit costs (see [78, Theorem 2]). Specifically considering d_E with operations consisting of insertion, deletion, relabelling, and subtree moves, if the subpath signatures of two trees for parameter t are the same, i.e., $S_t(T_1) = S_t(T_2)$, then

$$d_E(T_1, T_2) \leq n - \min(t - 1, \text{height}(T_1) + 1),$$

where $n := |T_1|$ ($n > 2$) and $t > 1$.

3.4 LSH for Hierarchical Data

In this section, we describe how LSH is used to specifically handle hierarchical data such as documents and fixed-height trees.

Recursive MinHash (RMH) [20] uses q -MinHash repeatedly by following a bottom up approach to compare hierarchical data.

- The method **recursively** performs the following steps starting from the lowest level, until the highest level is reached:
 1. Apply q -**MinHash** to data at the current level.
 2. **Reorganize** these q -MinHash vectors into q sets: insert the i th term from each q -MinHash vector into the i th set.
- Finally, we apply q -MinHash again to the reorganized vectors and **concatenate** the results into one single fingerprint.

Fig. 3 shows how RMH works for a toy example consisting of two levels, with $q = 4$. The sets $\{a, b, c\}, \{b, e\}, \{d, e, a\}$ comprise the bottom level. First, q -MinHash is applied on these sets, and the hashes are reorganized. Next, q -MinHash is generated for the reorganised sets to account for the top level. The final signature is the ordered set of these q -MinHash signatures for these reorganized sets. We describe the hash functions in detail in the supplement.

A single fingerprint is represented as an ordered set of k values. These values can then be split into b bands of $r = k/b$ elements, each within an LSH framework. The LSH parameters r and b do not need to correspond with the signature parameter q .

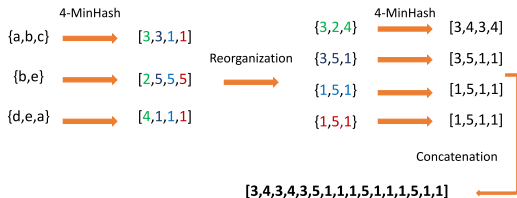


Fig. 3: An illustration of RMH.

4 LSH FOR COMPARING MERGE TREES

Our LSH framework introduces two similarity measures: one based on Recursive MinHash s_R , and another based on subpath signature s_S . To compare against distance matrices in our experiments, we convert each

similarity measure to a distance, denoted as d_R and d_S , respectively, where $d_R = 1 - s_R$ and $d_S = 1 - s_S$. We present an overview of our framework followed by a detailed description of the algorithms.

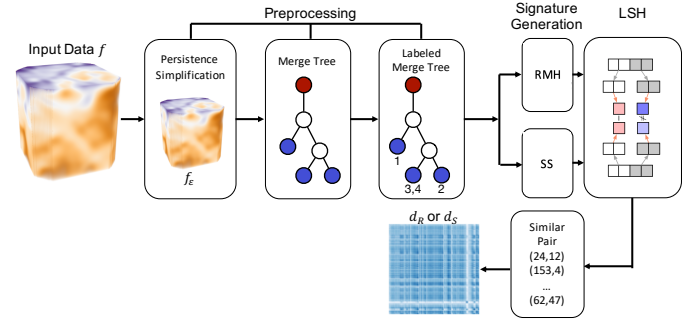


Fig. 4: A pipeline of LSH for comparing merge trees.

Overview. An overview of our pipeline is shown in Fig. 4. Given a set of scalar fields as input, we generate our d_R and d_S matrices as follows:

- **Preprocessing.** We first simplify each scalar field with a small persistence threshold to remove noise in the data [24]. We then compute its corresponding merge tree, followed by label assignment.
- **Signature generation.** We take the labeled merge trees as the input and generate their signatures using either RMH or subpath signature (SS) algorithms.
 - SS signature: For each labeled merge tree, we can identify the elements W of a universal set U of subpaths. Then, we directly use MinHash to obtain signatures.
 - RMH signature: We combine RMH with a hierarchy of trees, and apply it to labeled merge trees to generate signatures.
- **LSH.** We divide signatures into b bands each with r rows. If two objects collide in any band, then we mark these two objects as a similar pair. For empirical comparison, we generate our distance matrices d_R and d_S by collecting all similar pairs from the LSH.

4.1 Subpath Signature for Labeled Merge Trees

We extend the rooted-tree subpath signature method to labeled merge trees. Given labeled merge trees and a parameter t that denotes subpath length, we generate all subpaths of length t and collect them in a multiset. We include a path containing $t - 1$ dummy nodes leading to the root node. This ensures that all the subpaths would be of length t , even in the extreme case where the tree consists of only one node.

Fig. 5 shows an example on generating subpath signatures for a labeled merge tree with $t = 3$ and $q = 2$. After adding two dummy nodes, we start collecting all the subpaths of length three from the new root and traverse down the tree until every node has been visited. We then apply 2-MinHash to each merge tree to get a collection of vectors. Finally, we take these vectors as input for LSH.

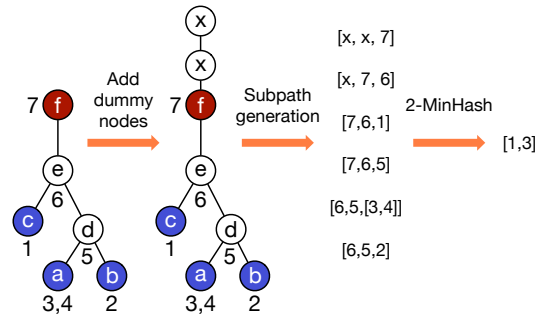


Fig. 5: Illustration of subpath signature for a labeled MT.

We design a simple modification of depth-first search (DFS) to generate subpaths. Intuitively, a traditional DFS returns a subpath of length 1. To modify it to return a subpath of length t , we do the following steps: whenever a node is flagged as visited in a traditional DFS, it is popped and returned. Instead of a single node, we implement

a $(t, t - 1)$ pop-push operation that pops t elements to provide us with the subpath and immediately pushes back $t - 1$ elements that have not been flagged as being visited.

In practice, if we implement the stack using an array, then we need not explicitly perform $(t, t - 1)$ pop-push operations since we can access the elements that are not at the top of the stack. Thus, we generate the multiset containing all the subpaths of length t . The supplement provides the pseudocode. [Algorithm 4](#) generates the subpath multiset. [Algorithm 3](#) is the modified DFS that is used internally by [Algorithm 4](#). It modifies traditional DFS with the extra step POP - PUSH($t, t - 1$), which takes $O(t)$ time. Thus, the running time is given by $O(n \times t)$ where n is the size of the tree. Since $t \ll n$, and $t < 10$ is a constant in practice for most cases, the running time is linear in the size of the trees.

4.2 RMH for Labeled Merge Trees

Our starting point is RMH, as described in [Sec. 3.4](#). RMH provides a hash signature for hierarchical structures. It was originally designed for documents where the hierarchy is fixed, consisting of words, sentences, paragraphs, and so on. However, for merge trees, the hierarchy (the tree structure) can be different for each object as it depends on the topology of the sublevel sets.

We first make the new observation that the RMH framework can be applied to any hierarchical object. Every reorganization step results in q sets. Then, in the next recursive step, q sets always lead to q different q -MinHash signatures. This is true no matter how many iterations of the hierarchy have been processed. Hence, the final concatenated signature is always of length $k = q^2$. Since the above process is independent of the height or structure of the tree, the signature can be compared across trees of different topology.

Second, unlike the traditional RMH where all nodes at a particular level are of the same type, in the case of merge trees, the nodes can be of different combinations. At a particular level, the nodes can be all extrema, a mixture of extrema and saddles, all saddles, or the root node. To address the combination of different nodes, we apply a recursion based on the type of the nodes rather than the level. This ensures that the MinHash of subtree rooted at a and the MinHash of subtree rooted at b will be subjected to reorganization only at the level corresponding to their lowest common ancestor, $lca(a, b)$. We modify the RMH to reflect this small but crucial difference so that it can be applied to merge trees; see [Algorithm 2](#) in the supplement for the pseudocode.

In [Fig. 6](#), we provide a toy example illustrating how RMH with 2-MinHash works for labeled merge trees. First, we apply 2-MinHash to all the labels of leaf nodes in our merge tree. Second, we reorganize the hash vectors of leaf nodes that share the lowest common ancestor, leaf nodes a and b , and their parent node d . Third, we truncate the tree to move one level up. We perform the three steps above recursively until we only have one node left, which is the root. Then, we apply 2-MinHash again to obtain the hash vectors and concatenate them into a single LSH signature as the input for LSH.

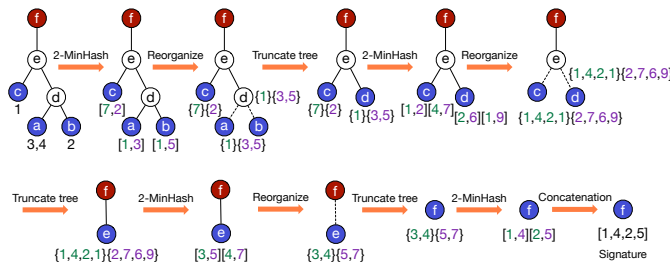


Fig. 6: An illustration of RMH on a labeled MT.

Using our observation that each recursive round always generates at most q sets, we provide a simplified analysis of the runtime compared to [\[20\]](#). We say a node is an *exposed leaf node* if it has not yet been truncated in the recursive process, and it has no child nodes in the truncated tree. In each round, each exposed leaf node needs to convert from at most q sets, each of size at most q^2 , to q MinHash signatures. This process takes $O(q^4)$ time. For binary trees of height L with n

nodes, there are at most $\min(2^L, n)$ exposed leaf nodes. Since the height decreases by one each round, the total runtime is

$$\sum_{z=1}^L O(\min(2^{L-z+1}, n)q^4) = O(Lnq^4).$$

Therefore, if q is a small constant (we typically use $q = 2$ or 4), then the runtime is near-linear in the size of the tree unless it is very unbalanced and the height L is super-logarithmic in the tree size.

4.3 Design Choices and Implementation Details

Labeled merge trees may contain nodes with multiple labels. To ensure that we do not miss any similar pairs, we incorporate multiple labels into our subpath, i.e., if node a has l labels then we generate a subpath which contains a , b , and c as $((a_1, \dots, a_l), b, c)$.

We follow the traditional LSH method that the signatures are divided into b bands of r rows each. A “match” between merge trees requires that for at least one band, all its rows have to match. Therefore, if r is larger, we require a closer match, and if b is larger, it is more forgiving as there are more chances to find an instance of all-rows correspondence. Since the relationship $k = r \times b$ is fixed, then for a specified k , we can tune the similarity threshold to define a “close pair” to be more specific by increasing r ; so adjusting r is like adjusting a threshold for similarity with any other distance. As we have observed, small r values (often $r = 1$) provide useful matches already.

We can also adjust the value of k : the larger the parameter k , the less variance in the matches found in this randomized process, but the more expensive the computation. We discover via experiments that a fairly moderate value of k (e.g., $k = 20$) works quite well.

If we have a very large number of trees, an additional hash function can be applied to report collisions within a single band. However, given the scale of our experiments (not involving millions of trees), it is feasible to calculate and report all collisions directly. This approach enhances the precision of our reporting for this paper.

Implementation details. We use the implementation in [\[80\]](#) to generate merge tree labels. We implement both d_R and d_S in Python. We also implement the algorithm to generate subpaths of length t required for d_S . We implement the MinHash first introduced by Broder [\[9\]](#).

5 EXPERIMENTS AND RESULTS

We experimentally validate the effectiveness of our LSH framework, in terms of **utility** and **efficiency**:

- We demonstrate that our new similarity measures are effective on a wide variety of examples, in recovering results from existing merge tree distances, and in some cases uncovering new scientific structures.
- We illustrate that our framework is significantly more efficient than standard distance measures. It achieves 10-30 \times speed-up on moderately-sized datasets, and (estimated) 800 \times on large ones.

Comparison with edit and interleaving distances. Our framework introduces two similarity measures: one based on Recursive MinHash s_R and another based on subpath signature s_S . We convert each similarity measure to a distance, denoted as d_R and d_S , respectively, where $d_R = 1 - s_R$ and $d_S = 1 - s_S$.

We compare our results with the merge tree edit distance [\[62\]](#) (referred to as edit distance for short, denoted as d_E) and the geometry-based interleaving distance [\[80\]](#) (denoted as d_I). We choose these two distances as they are both applicable to labeled merge trees but represent different types of similarity measures. d_E provides the best theoretical runtime among comparative measures that internally solve matching problems, whereas d_I decouples the computation externally into a labeling step and a comparison step.

In our comparison, we report d_E and d_I , and use a color scale to visually observe the thresholds when useful structures become apparent. We then demonstrate that our LSH framework could approximately match these structures as follows: for each fixed k , we adjust r so that the matched merge trees approximately correspond with those from the other distances. We visualize this optimal choice of r as a

Table 1: Detailed descriptions of all datasets and parameters used in all experiments. ε stands for persistence threshold.

Dataset	Dimensions	# Instances/Time steps	ε	Size of Merge Tree	k	r
Vortex Street	$192 \times 64 \times 48$	102	0.01	10 - 104 nodes	20, 40, 60, 80	1, 2
Shape Matching	varies	132	0.01	10 - 52 nodes	20, 40, 60, 80	1, 2, 4
Corner Flow	450×150	1500	0.01	12 - 78 nodes	20, 40, 60, 80	1, 2
Heated Flow	150×450	2000	0.01	2 - 148 nodes	20, 40, 60, 80	1, 2
Viscous Finger	$100 \times 100 \times 100$	5746	5	64 - 232 nodes	20, 40, 60, 80	1, 2

symmetric binary matrix. In some cases, the other distances do not find an interesting structure, but our LSH-based measures do.

We use the Topology Toolkit (TTK) [66] to visualize scalar fields and generate merge trees. We compute d_E and d_I using implementations provided in [62] and [80], respectively. We use a hybrid mapping strategy from [80] to generate labels for all datasets, since it encodes both geometric and topological information from the input. We perform our experiments on a standard laptop with i7 processor with 20 threads running at 3.5 GHz, with 32 GB memory. We use Moving Gaussian, a toy dataset from [80], TOSCA shape dataset from [58], Corner flow [4] and Heated flow [35] datasets from [51] (available at [17]), 3D vortex street dataset from [11] (available at [1]) and Viscous Fingers dataset [57].

5.1 An Illustrative Example

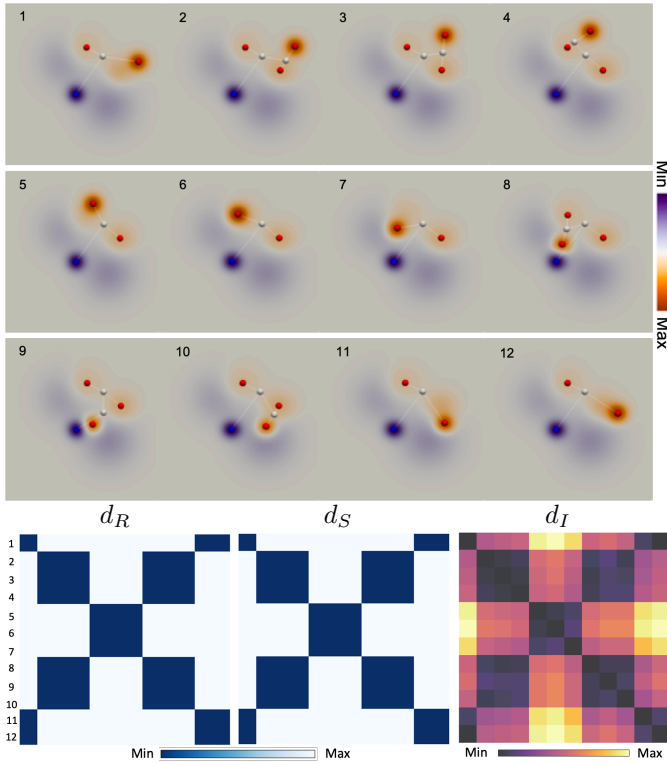


Fig. 7: Moving Gaussian dataset. Top: 12 time steps are visualized with embedded merge trees. Bottom: comparing merge trees with binary matrices for d_R and d_S , together with an interleaving distance d_I matrix.

We first consider a toy dataset, called Moving Gaussian, generated by placing a mixture of three Gaussian functions on a plane. The dataset consists of 12 time steps, where one of the three Gaussian functions moves counterclockwise around two fixed Gaussian functions [80]. We compute the merge trees of the inverse, which capture the relationships between local extrema and saddles. Fig. 7 (top) illustrate the set of scalar fields along with the merge trees.

This dataset contains natural clusters by design. Yan et al. [80] reported three clusters formed by time steps $\{2, 3, 4, 8, 9, 10\}$, $\{5, 6, 7\}$, and $\{1, 11, 12\}$, respectively; see Fig. 7 (bottom right).

We compare our LSH framework against the results generated with the interleaving distance in [80]. We apply a persistence threshold $\varepsilon = 0.02$ to separate features from noise in the scalar fields. We then

generate merge trees and labels for all 12 time steps and apply our LSH-based similarity measures.

To visualize our results, we mark an entry in a 12×12 binary matrix to be 1 if there exists a candidate pair by an LSH collision, otherwise 0. The binary matrices for both d_R and d_S are shown in Fig. 7 (bottom), along with a 12×12 interleaving distance d_I matrix from [80]. We observe similar clustering results from both d_R and d_S , confirming that our framework can be a good alternative to the interleaving distance.

We experiment with different values of k , b , and r . Our results match exactly the results in [80], when $k = 4$, $r = 4$, and $b = 1$ for d_R and $k = 8$, $r = 4$, and $b = 2$ for d_S . For subsequent experiments, Tab. 1 shows all datasets and parameters used in the experiments.

5.2 Shape Matching

Shape matching involves detecting similarities between shapes. Even though geometric methods have been very powerful in shape matching, we could still use topological descriptors to get reasonable results.

We use a TOSCA shape dataset, which contains various nonrigid shapes in different poses. Fig. 8 shows 10 classes of shapes with ground truth labels, where each class contains multiple poses of the same shape. We aim to classify similar shapes correctly using merge trees, irrespective of their poses.

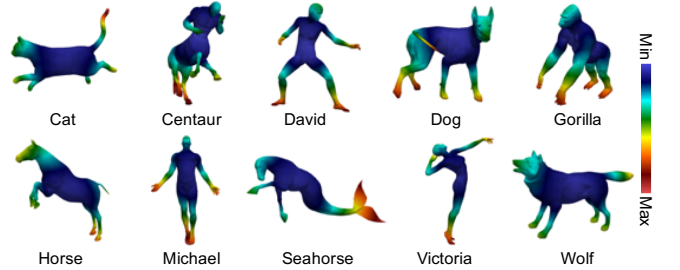


Fig. 8: TOSCA shape dataset contains 10 classes of shapes. One pose is selected to represent each class. Each mesh is colored by the average geodesic distance from a set of anchor points.

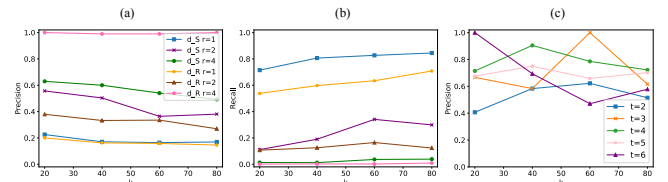


Fig. 9: TOSCA shape dataset: precision and recall in shape matching. (a)-(b): Precision and recall plot for d_R and d_S , by varying r w.r.t k . (c): Exploring parameter configurations for d_S .

We define the scalar field f to be the average geodesic distance from a set of anchor points on the mesh. We then generate merge trees of $-f$ and their node labels. We use this dataset to demonstrate the utility of our LSH framework in finding exact matching pairs. We also use it to illustrate parameter choices.

We first investigate how the choices of parameters k and r (so $b = k/r$) can affect the outcome of d_R and d_S in the context of shape matching where we have labeled ground truth. The same strategy extends to other datasets. Note that the primary computational cost of a LSH algorithm is computing the hash function for each data object, where k indeed can be used as a representation of runtime. In Fig. 16, we explicitly investigate the relationship between k and runtime.

Fig. 9 plots the average precision and recall as we vary r and k . For each shape x , a precision is computed as a ratio of the matched instances

with the same label over all matched instances. A pointwise *recall* is computed as a ratio between the matched instances with the same label versus all instances with that label. As r increases, precision increases, but recall decreases (as with any distance and similarity threshold). We see this for d_R and d_S in this plot considering $k \in [20, 40, 60, 80]$ and $r \in [1, 2, 4]$. We observe k does not affect precision and recall much and so we often opt for a smaller $k = 20$. We also favor a smaller $r = 1$, since this gives much better recall. We can always filter mismatches (because of not as good precision) with a direct computation of a slower distance over a smaller number of matched objects.

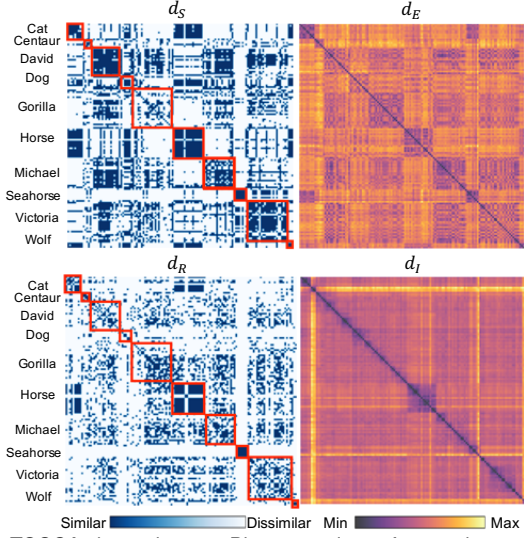


Fig. 10: TOSCA shape dataset. Binary matrices of d_R and d_S are shown with red squares which contain instances with the same ground truth class label. d_E and d_I distance matrices are provided for context.

Fig. 9 also varies subpath length t . We choose $t = 4$ to maintain a good precision without the subpath being too long, depending on the height of the merge trees.

We perform experiments on different parameter settings provided in Tab. 1. Fig. 10 shows the binary matrices for both d_R and d_S , with $k = 20$ and $r = 1$. Each binary matrix shows clear block structures along its diagonal, for example, in Cat, David, Horse, Michael, Seahorse, and Victoria classes. The results show that we obtain reasonable classification of some classes.

On the other hand, d_R and d_S (to some extent) perform imperfect classification between Gorilla, Horse, and Michael classes. For instance, we observe off diagonal blocks showing similarities between Horse and Cat. The reason is two-fold. First, merge trees have similar structures among these classes of shapes, and the merge tree itself is not always a good descriptor for capturing all the geometric details of a shape. Second, the labeling strategy, which currently labels only the leaf nodes, might also contribute toward the imperfect results. The LSH framework intuitively captures a variation of Jaccard similarity (based on hash buckets). Since the Jaccard similarity generally ignores hierarchy, d_R applies Jaccard similarity at multiple levels to alleviate this issue.

In Fig. 10, d_E and d_I matrices are also provided for context. In comparison, d_I captures less information as fewer classes are classified correctly by d_I . We observe that d_E gives good classification results for only some of the classes. For instance, d_E also exhibit off diagonal blocks showing similarities between Gorilla and Michael, David and Michael, and so on. The small discrepancies between d_E and our framework probably arise from modeling shapes with merge trees, not the similarity measures we employ.

5.3 Time-Varying and Ensemble Data Summarization

We demonstrate via experiments the utility of our LSH framework in summarizing time-varying scalar fields and ensembles.

5.3.1 Vortex Street

We first demonstrate temporal summarization using a 3D Bénard von Kármán vortex street dataset. We consider the magnitude of velocity

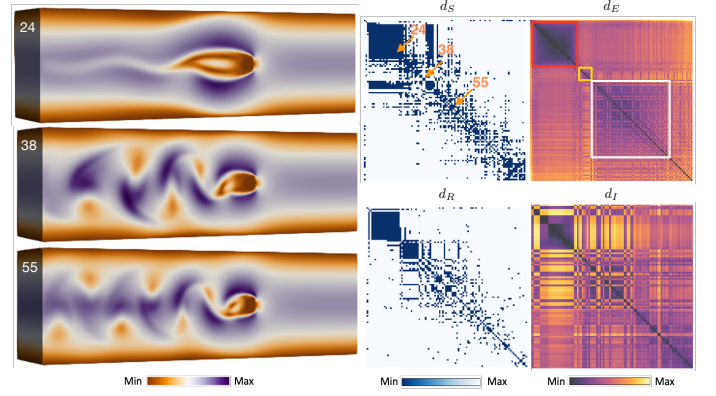


Fig. 11: Vortex Street dataset. Binary matrices for d_R and d_S are presented with selected time steps from different clusters. Distance matrices for d_I and d_E are provided for comparison.

as the scalar field. We generate the binary matrices according to the parameters provide in Tab. 1.

Fig. 11 (left) visualizes three instances with different structures along the time-varying dataset. This dataset contains a set of clusters capturing different phases of the flow behavior [62]. During the 1st phase, the flow evolves slowly, and the first 30 time steps are highly similar to one another, and time step 24 is selected as a representative. During the 2nd phase, the flow is transitioning toward a more periodic behavior, and time step 38 marks a transition to the next phase. During the 3rd phase, the vortex shedding phenomena (i.e., an oscillating flow behind a bluff body at certain velocities) become clearly visible, and time step 55 is selected as a representative.

Fig. 11 (right) includes binary matrices d_R ($k = 60, r = 1$) and d_S ($k = 20, r = 1, t = 4$). We observe that d_E clearly identifies the three phases as three clusters (blocks). In particular, it shows periodicity in the 3rd block that corresponds to vortex shedding. In the d_S matrix, we identify the 1st major cluster clearly, but the 2nd minor cluster and the 3rd cluster are not as obvious as those from d_I . In the d_R matrix, the vortex shedding is vaguely visible as periodic block patterns, similar to what is shown in the d_I matrix.

5.3.2 Viscous Finger

We use a Viscous Finger dataset to demonstrate the scalability of our LSH method in summarizing a large-scale ensemble dataset and uncover similarity between particular runs.

The ensemble is composed of 3D transient fluid flow obtained by a simulation with stochastic effects, formulating a special behavior named as viscous finger. We choose the ensemble with 0.44 as the resolution level. This ensemble contains 48 runs, each contains ≈ 120 time steps. To compare equally among different runs, we generate binary matrices across all runs with parameters in Tab. 1.

Fig. 12 (top) shows binary matrices d_S (with $k = 60$ and $r = 1$) and d_R ($k = 20$ and $r = 1$) across 48 runs. For this large dataset, both d_E and d_I become intractable; therefore, we compute only a submatrix of d_I for comparison. The binary matrix d_R has a lighter color compared to d_S as d_R shows less similarities than d_S .

We observe that most runs are similar due to a similar simulation process, except the end instance and the beginning instance connecting two runs where straight line patterns separate the entire matrix into blocks. From the zoomed-in d_I matrix, we also observe such block-wise pattern.

Our binary matrices identify two runs that perform differently, runs 44 and 48. We also observe similar patterns in the d_I matrix, where run 44 forms a clear diagonal block, meaning that instances inside the run are more similar internally than externally.

We select time step 120 from each of the runs 43, 44, 45, and 48, and render their isosurfaces respectively, shown in Fig. 13. We observe that runs 44 and 48 behave similarly with each other. These visualizations match the observations from the binary matrices in Fig. 12.

Given a large ensemble, our LSH framework manages to capture the global behavior due to its efficiency and scalability, whereas d_E

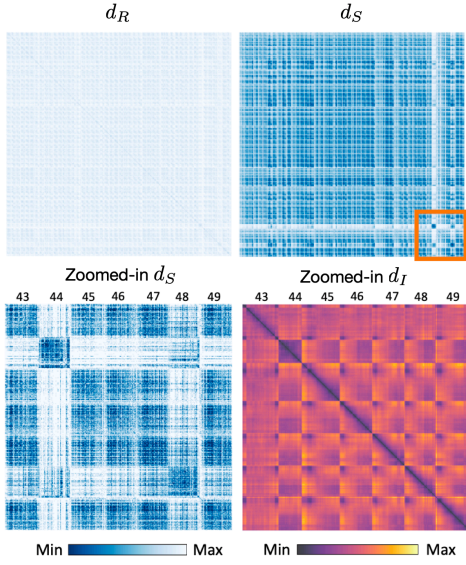


Fig. 12: Viscous Finger dataset. Top: binary matrices for d_R and d_S . Bottom: the zoom-in matrices are from run 43 to run 49.

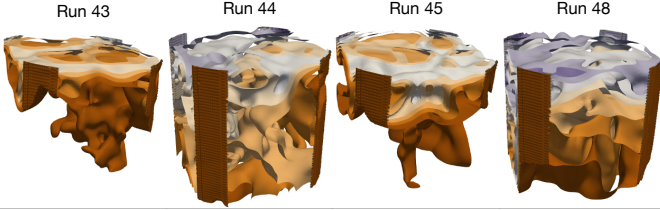


Fig. 13: Viscous Finger dataset. Isosurface rendering of time step 120 from runs 43, 44, 45, and 48, respectively.

and d_I fail to do so. In particular, both d_S and d_R allow us to discover similarities among the runs.

5.4 Identification of Clusters and Data Transitions

Here we show the utility of our framework in identifying clusters and structural transitions in large flow datasets.

5.4.1 Heated Flow Dataset

The Heated Flow dataset is a time-varying 2D dataset generated by flow around a heated cylinder using Boussinesq approximation where the flow contains many small vortices. We use the magnitude of the flow as the scalar field f . We compute merge trees of f and node labels. The binary matrices are generated based on the parameters in Tab. 1.

Fig. 14 shows binary matrices of d_S ($k = 60$, $r = 1$) and d_R ($k = 20$ and $r = 1$). Our framework helps to identify six clusters, highlighted as colored blocks: cluster 1 (grey), time steps 70 - 300; cluster 2 (grey), time steps 485 - 814; cluster 3 (orange), time steps 815 - 1180; cluster 4 (green), time steps 1181 - 1450; cluster 5 (grey), time steps 1451 - 1690; and cluster 6 (magenta), time steps 1691 - 2000.

We select two time steps, 867 and 897, 1200 and 1230, and 1700 and 1730, from each of the three clusters (orange, green, and magenta) to show their similarity within the cluster and dissimilarity outside the cluster. The colors on the labels correspond to different clusters.

d_E matrix visibly captures only the 1st cluster, but fails to show noticeable clusters for the rest of the dataset. d_I matrix, on the other hand, groups most time steps into one big cluster. Therefore, our LSH framework provides clearer clustering structure.

5.4.2 Corner Flow Dataset

The Corner Flow is a 2D dataset generated by flow around two cylinders. The flow is bound by walls with corners around which vortices form due to the presence of the cylinders. We use the velocity magnitude as the scalar field f . We generate the binary matrices using parameters in Tab. 1. We demonstrate the ability of our framework in capturing structural transitions of time-varying scalar fields.

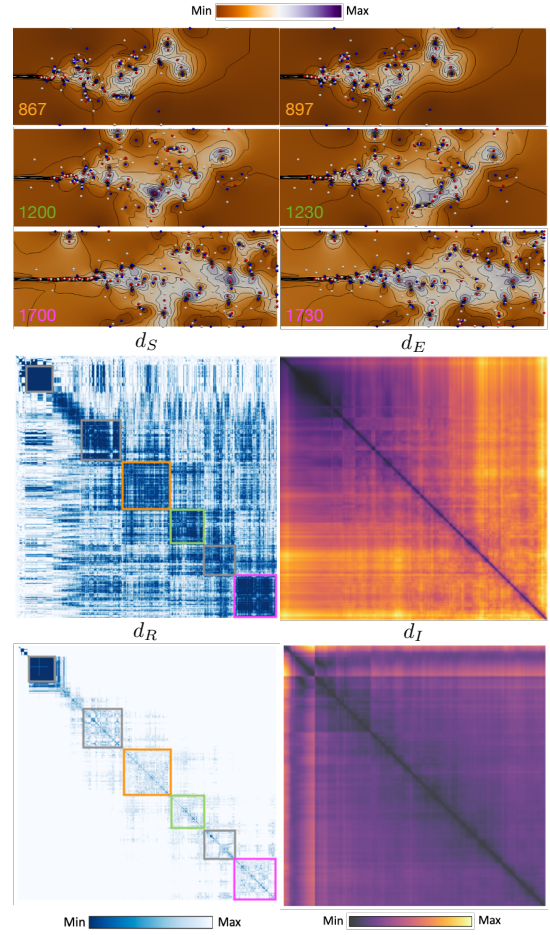


Fig. 14: Heated Flow dataset. Top: time steps from different clusters. Bottom: d_R and d_S . Different colored boxes represent different clusters.

Fig. 15 shows the binary matrices of d_R ($k = 20$, $r = 1$) and d_S ($t = 4$). We observe three pairs of time steps showing structural transitions, highlighted as arrows in d_S and d_R . From time step 773 to 774 (orange filled arrow), there is a structural transition. As fluid flows, the contour line expands and two max-saddle pairs disappear in 774. The critical points are highlighted in the white dashed box. From time step 1095 to 1096 (orange filled double arrow), there is another structural transition where a new min-saddle pair appears. From time step 1359 to 1360 (orange double arrow), a min-saddle pair disappears.

d_R matrix shows similar patterns as d_E matrix, where a small cluster connects to a large cluster for the 2nd half of the time-varying dataset. d_R matrix also presents similar patterns (i.e., two clusters) as the d_I matrix at the beginning of the time steps. The binary matrices of both d_R and d_S can capture transitions, whereas d_S provides a more precise structural detection. In conclusion, our LSH framework serves as a good alternative to existing distance measures.

Table 2: Runtime analysis of d_S and d_R . Each entry represents the LSH runtime with a corresponding k . All times are in seconds.

Dataset	Method	$k = 20$	$k = 40$	$k = 60$	$k = 80$
Vortex Street	d_S	0.18	0.21	0.23	0.27
	d_R	1.26	4.19	10.55	21.82
TOSCA	d_S	0.27	0.30	0.34	0.38
	d_R	1.28	4.03	10.07	20.77
Corner Flow	d_S	3.85	4.51	5.09	5.71
	d_R	37.49	121.17	307.22	637.55
Heated Flow	d_S	1.5	1.72	2.03	2.31
	d_R	76.9	254	663.2	1357.2
Viscous Fingers	d_S	45.1	55.18	65.45	72.18
	d_R	374	1156	2843	5636

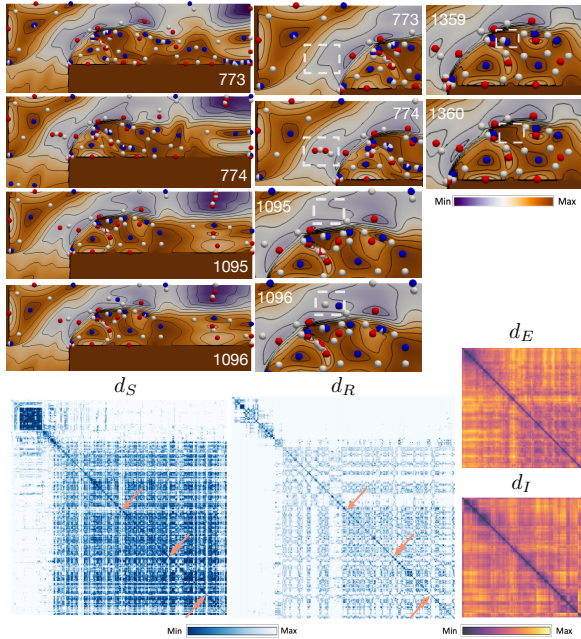


Fig. 15: Corner Flow dataset. d_S and d_R : binary matrices with visualized scalar fields to show structural transitions. d_E and d_I are also provided. Table 3: Runtime analysis across datasets: columns d_S and d_R represent the total runtime (the sum of LSH time and labeling time) for $k = 20$. The improved runtime are highlighted in bold. All times are in seconds. Numbers with asterisk are estimated.

Dataset	d_S	d_R	d_I	d_E
Vortex Street	38.08	39.26	65.5	48.0
TOSCA	11.72	12.72	24.9	109.7
Corner Flow	363.92	397.56	3456.0	6203.1
Heated Flow	2660.7	2736.1	23686	211688
Viscous Fingers	13825	14154	$\approx 455388^*$	$\approx 15965520^*$

5.5 Quantification of Runtime Improvements

Tab. 2 and Tab. 3 provide a runtime analysis of our LSH framework, which depends on two main factors. The first is the creation of k signatures, and the use of hash buckets to identify matchings. The second is the labeling of the merge trees following a hybrid strategy in [80]. Tab. 2 reports LSH runtime, for both d_S and d_R along with the labeling time in bold. For both d_S and d_R , the LSH runtime increases as k increases. We also observe that d_S is faster than d_R for all datasets.

From Tab. 2, we can see that as the size of the data grows, the LSH process has a very small runtime, and the runtime bottleneck is indeed the labeling process. We argue that this is acceptable in certain scenarios. For instance, assuming we are performing interactive analysis in real time over a fixed dataset, the labeling process can be precomputed (once), and only the LSH process needs to run multiple times on-the-fly.

For comparison, we report runtime for both d_E and d_I along with the total runtime for d_S and d_R for $k = 20$ in Tab. 3. The total runtime is the sum of LSH time and the labeling time. For the Vortex Street dataset, we have a similar runtime for most cases compared to d_E and d_I , where d_R gives slightly higher runtime when $k = 60$ and $k = 80$. For the TOSCA dataset, d_R gets slightly higher runtime compared to d_I , when $k = 80$; otherwise our LSH measures are faster to compute. We obtain approximately $3\times$ speed-up for d_R ($k = 80$) compared to d_E . Our LSH framework has approximately $1.5\times$ speed-up compared to d_I , and approximately $9\times$ speed-up compared to d_E for d_S . Our LSH framework takes less time for TOSCA than Vortex Street. Despite TOSCA containing more instances, the merge trees are generally smaller.

We observe a larger speedup for Corner Flow and Heated Flow. For both datasets, we have (approximately) $9\times$ speedup for d_S and $6\times$ speed-up for d_R compared to d_I . For Corner Flow, we have $17\times$ and $6\times$ speed-up compared to d_E for d_S and d_R , respectively. For Heated

Flow, we have $80\times$ for d_S and $5\times$ speed-up for d_R compared to d_E .

For Viscous Finger, the runtime is significantly improved. We compute d_I for only a subset of the data, since computing the entire d_I matrix is estimated to take 5 days; this runtime is estimated based on the recorded runtime for the first 40,000 comparisons.

We obtain $33\times$ (d_S) and $23\times$ (d_R) speed-up compared to d_I , even for $k = 80$. We are not able to compute d_E for the full dataset comprising of 5746 instances, due to its lack of scalability in computation. Therefore, we run d_E on one single run that contains 120 instances, and report a runtime of ≈ 6929 seconds, close to 2 hours. We estimate the runtime of d_E for the entire dataset, estimating across all 48 runs, which takes roughly $48^2 \times 6929 \approx 185$ days! Thus, our LSH approach is estimated to be about $822\times$ faster than d_E computation.

As data grow larger, we observe larger improvements, compared to d_E and d_I . d_E suffers from poor scalability due to two reasons. First, all pairs comparisons need to be explicitly performed. Second, each comparison involves solving a set of matching problems, which leads to high runtime in practice. Even for the Viscous Finger dataset, d_E and d_I distance matrices become impractical to compute, and our LSH framework becomes a necessity.

In addition, we perform an ablation runtime analysis as we vary k in Fig. 16. Our analysis shows the LSH runtime excluding the labeling process, with different k . d_S presents nearly linear behavior and d_R shows nearly quadratic behavior as a function of k . There are other costs associated with retrieving the query, but they are lower-order terms and do not depend on the size of the data object. Fig. 16 demonstrates that k is a good representation of runtime.

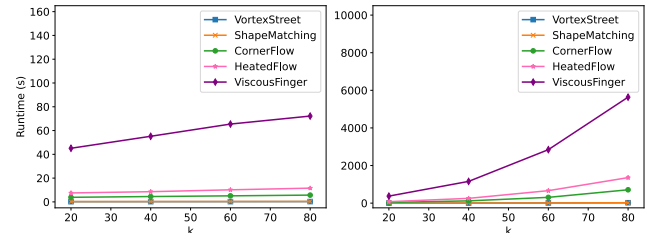


Fig. 16: Runtime analysis of d_S and d_R , excluding the labeling process. Left: k vs. runtime for d_S , $t = 4$. Right: k vs. runtime for d_R .

6 DISCUSSION AND LIMITATIONS

Our LSH framework using d_S and d_R , as expected, is efficient and scalable, in comparison with existing distance methods such as d_E and d_I . In fact, our framework is efficient on large datasets when d_E and d_I become intractable to compute. Furthermore, the d_S and d_R we propose offer enormous advantages in scalability, often replicate results of d_E and d_I , and sometimes capture new cluster features that d_E and d_I do not. Therefore, they could be used as efficient, alternative measures for comparing scalar fields at scale. As a preprocessing step, LSH can eliminate unnecessary candidates for comparative analysis, therefore reducing the size of input data for other distance computations.

A limitation of our LSH framework is that it does not recover the nearest-neighbor exactly as does d_E . With any LSH approach, we need to adjust parameters k and r , which control the trade-off between precision and recall. We find that even small values $k = 20$, $r = 1$ provide decent results, and recommend these parameters for similar problems to the ones we consider. Another factor affecting LSH is the labeling strategy. While the requirement for labeled merge tree seems like a limitation, the labels can be used to incorporate geometric information which is crucial in many applications [14, 61, 80]. We conjecture that our LSH results would improve with more stable labeling strategies for internal nodes. Making the labeling process more efficient will also benefit the overall runtime, which is left for future work.

ACKNOWLEDGEMENT

This work was supported by grants from DOE DE-SC0021015 and DE-SC0023157, as well as NSF DMS-2134223, IIS-1910733, IIS-1816149, CCF-2115677, and IIS-2311954.

REFERENCES

- [1] 3D flow around a confined square cylinder. <http://tinowewinkauf.net/notes/squarecylinder.html>, 2024. 6
- [2] A. Acharya and V. Natarajan. A parallel and memory efficient algorithm for constructing the contour tree. In *IEEE Pacific Visualization Symposium*, pp. 271–278. IEEE, 2015. doi: 10.1109/PACIFICVIS.2015.7156387 2
- [3] F. Aioli, G. Da San Martino, A. Sperduti, and A. Moschitti. Efficient kernel-based learning for trees. In *IEEE Symposium on Computational Intelligence and Data Mining*, pp. 308–315. IEEE, 2007. doi: 10.1109/CIDM.2007.368889 2
- [4] I. Baeza Rojo and T. Günther. Vector field topology of time-dependent flows in a steady reference frame. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):280–290, 2020. doi: 10.1109/TVCG.2019.2934375 6
- [5] C. Bajaj, A. Gillette, and S. Goswami. Topology based selection and curation of level sets. In H.-C. Hege, K. Polthier, and G. Scheuermann, eds., *Topology-Based Methods in Visualization II*, pp. 45–58. Springer, 2009. doi: 10.1007/978-3-540-88606-8_4 2
- [6] K. Beketayev, D. Yeliussizov, D. Morozov, G. H. Weber, and B. Hamann. Measuring the distance between merge trees. In P.-T. Bremer, I. Hotz, V. Pascucci, and R. Peikert, eds., *Topological Methods in Data Analysis and Visualization III*, pp. 151–165. Springer, 2014. doi: 10.1007/978-3-319-04099-8_10 2
- [7] T. Biedert and C. Garth. Contour tree depth images for large data visualization. In *Proc. 15th Eurographics Symposium on Parallel Graphics and Visualization*, pp. 77–86, 2015. doi: 10.2312/pgv.20151158 2
- [8] B. Bollen, P. Tennakoon, and J. A. Levine. Computing a stable distance on merge trees. *IEEE Transactions on Visualization and Computer Graphics*, 29(01):1168–1177, 2023. doi: 10.1109/TVCG.2022.3209395 2
- [9] A. Z. Broder. Identifying and filtering near-duplicate documents. In R. Giancarlo and D. Sankoff, eds., *Combinatorial Pattern Matching*, pp. 1–10. Springer, 2000. doi: 10.1007/3-540-45123-4_1 5
- [10] A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher. Min-wise independent permutations. *Journal of Computer and System Sciences*, 60(3):630–659, 2000. doi: 10.1006/jcss.1999.1690 3
- [11] S. Camarri, M. Buffoni, A. Iollo, and M. V. Salvetti. Simulation of the three-dimensional flow around a square cylinder between parallel walls at moderate Reynolds numbers. In *XVII Congresso di Meccanica Teorica ed Applicata*, pp. 11–15, 2005. doi: 10.1063/1.869879 6
- [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):3-13:13, 2013. doi: 10.1186/1471-2105-14-3 3
- [13] H. Carr, J. Snoeyink, and U. Axen. Computing contour trees in all dimensions. *Computational Geometry: Theory and Applications*, 24(2):75–94, 2003. doi: 10.1016/S0925-7721(02)00093-7 1, 2
- [14] H. Carr, J. Snoeyink, and M. Van De Panne. Flexible isosurfaces: Simplifying and displaying scalar topology using the contour tree. *Computational Geometry*, 43(1):42–58, 2010. doi: 10.1016/j.comgeo.2006.05.009 2, 9
- [15] H. A. Carr, C. M. Sewell, L.-T. Lo, and J. P. Ahrens. Hybrid data-parallel contour tree computation. In *Proc. Conference on Computer Graphics & Visual Computing*, pp. 73–80, 2016. doi: 10.2312/cgvc.20161299 2
- [16] H. A. Carr, G. H. Weber, C. M. Sewell, O. Rübél, P. Fasel, and J. P. Ahrens. Scalable contour tree computation by data parallel peak pruning. *IEEE Transactions on Visualization and Computer Graphics*, 27(4):2437–2454, 2019. doi: 10.1109/TVCG.2019.2948616 2
- [17] Computer Graphics Laboratory. <https://cgl.ethz.ch/research/visualization/data.php>, 2024. 6
- [18] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. 34th ACM Symposium on Theory of Computing*, pp. 380–388, 2002. doi: 10.1145/509907.509965 1, 2, 3
- [19] K. Chen and M. Shao. Locality-sensitive bucketing functions for the edit distance. *Algorithms for Molecular Biology*, 18(1):7, 2023. doi: 10.1186/s13015-023-00234-2 2
- [20] L. Chi, B. Li, and X. Zhu. Context-preserving hashing for fast text classification. In *Proc. SIAM International Conference on Data Mining*, pp. 100–108, 2014. doi: 10.1137/1.9781611973440.12 2, 3, 4, 5
- [21] L. Chi and X. Zhu. Hashing techniques: A survey and taxonomy. *ACM Computing Surveys*, 50(1):1–36, 2017. doi: 10.1145/3047307 2
- [22] F. Chierichetti, R. Kumar, A. Panconesi, and E. Terolli. On the distortion of locality sensitive hashing. *SIAM Journal on Computing*, 48(2):350–372, 2019. doi: 10.1137/17M1127752 2, 3
- [23] H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete & Computational Geometry*, 30:87–107, 2003. doi: 10.1007/s00454-003-2926-5 1
- [24] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28:511–533, 2002. doi: 10.1007/s00454-002-2885-2 4
- [25] O. Ertl. ProbMinHash – a class of locality-sensitive hash algorithms for the (probability) Jaccard similarity. *IEEE Transactions on Knowledge and Data Engineering*, 34(7):3491–3506, 2020. doi: 10.1109/TKDE.2020.3021176 2
- [26] M. Garofalakis and A. Kumar. Correlating XML data streams using tree-edit distance embeddings. In *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 143–154, 2003. doi: 10.1145/773153.773168 2, 3
- [27] M. Garofalakis and A. Kumar. XML stream processing using tree-edit distance embeddings. *ACM Transactions on Database Systems*, 30(1):279–332, 2005. doi: 10.1145/1061318.1061326 2, 3
- [28] 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. doi: 10.48550/arXiv.1908.00063 2, 3
- [29] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *International Conference on Very Large Data Bases (VLDB)*, vol. 99, pp. 518–529, 1999. 2
- [30] S. Gollapudi and R. Panigrahy. The power of two min-hashes for similarity search among hierarchical data objects. In *Proc. 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 211–220, 2008. doi: 10.1145/1376916.1376946 2
- [31] C. Gueunet, P. Fortin, J. Jomier, and J. Tierny. Contour forests: Fast multi-threaded augmented contour trees. In *IEEE Symposium on Large Data Analysis and Visualization*, pp. 85–92. IEEE, 2016. doi: 10.1109/LDAV.2016.7874333 2
- [32] C. Gueunet, P. Fortin, J. Jomier, and J. Tierny. Task-based augmented merge trees with Fibonacci heaps. In *IEEE Symposium on Large Data Analysis and Visualization*, pp. 6–15. IEEE, 2017. doi: 10.1109/LDAV.2017.8231846 2
- [33] C. Gueunet, P. Fortin, J. Jomier, and J. Tierny. Task-based augmented contour trees with Fibonacci heaps. *IEEE Transactions on Parallel and Distributed Systems*, 30(8):1889–1905, 2019. doi: 10.1109/TPDS.2019.2898436 2
- [34] D. Günther, J. Salmon, and J. Tierny. Mandatory critical points of 2D uncertain scalar fields. *Computer Graphics Forum*, 33(3):31–40, 2014. doi: 10.1111/cgf.12359 1
- [35] T. Günther, M. Gross, and H. Theisel. Generic objective vortices for flow visualization. *ACM Transactions on Graphics*, 36(4):141:1–141:11, 2017. doi: 10.1145/3072959.307368 6
- [36] M. Heimann, W. Lee, S. Pan, K.-Y. Chen, and D. Koutra. HashAlign: hash-based alignment of multiple graphs. In *Advances in Knowledge Discovery and Data Mining*, pp. 726–739. Springer, 2018. doi: 10.1007/978-3-319-93040-4_57 2
- [37] C. Heine, H. Leitte, M. Hlawitschka, F. Iuricich, 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. doi: 10.1111/cgf.12933 2
- [38] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. 30th Annual ACM Symposium on Theory of Computing*, pp. 604–613, 1998. doi: 10.1145/276698.276876 1, 2, 3
- [39] G. Johansson, K. Museth, and H. Carr. Flexible and topologically localized segmentation. In *Eurographics/IEEE-VGTC Symposium on Visualization*, pp. 179–186, 2007. doi: 10.2312/VisSym/EuroVis07/179-186 2
- [40] F. Lan, S. Parsa, and B. Wang. Labeled interleaving distance for Reeb graphs. *arXiv preprint arXiv:2306.01186*, 2023. doi: 10.48550/arXiv.2306.01186 1
- [41] B. Li, X. Zhu, L. Chi, and C. Zhang. Nested subtree hash kernels for large-scale graph classification over streams. In *IEEE International Conference on Data Mining*, pp. 399–408. IEEE, 2012. doi: 10.1109/ICDM.2012.101 2
- [42] H. Li, W. Wang, Z. Liu, Y. Niu, H. Wang, S. Zhao, Y. Liao, W. Yang, and X. Liu. A novel locality-sensitive hashing relational graph matching network for semantic textual similarity measurement. *Expert Systems with Applications*, 207:117832, 2022. doi: 10.1016/j.eswa.2022.117832 2
- [43] G. Marçais, D. DeBlasio, P. Pandey, and C. Kingsford. Locality-sensitive hashing for the edit distance. *Bioinformatics*, 35(14):i127–i135, 2019. doi: 10.1093/bioinformatics/btz354 2

- [44] S. McCauley. Approximate similarity search under edit distance using locality-sensitive hashing. In *24th International Conference on Database Theory*, pp. 21:1–21:22, 2021. doi: 10.4230/LIPIcs.ICDT.2021.21 2
- [45] S. Mizuta and T. Matsuda. Description of digital images by region-based contour trees. In *International Conference Image Analysis and Recognition*, pp. 549–558. Springer, 2005. doi: 10.1007/11559573_68 2
- [46] D. Morozov, K. Beketayev, and G. Weber. Interleaving distance between merge trees. *Topology-Based Methods in Visualization*, 2013. 2
- [47] V. Narayanan, D. M. Thomas, and V. Natarajan. Distance between extremum graphs. In *IEEE Pacific Visualization Symposium*, pp. 263–270, 2015. doi: 10.1109/PACIFICVIS.2015.7156386 1, 2
- [48] 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(11):1547–1559, 2011. doi: 10.1109/TVCG.2011.27 2
- [49] J. Poco, H. Doraiswamy, M. Talbert, J. Morisette, and C. T. Silva. Using maximum topology matching to explore differences in species distribution models. *IEEE Scientific Visualization Conference*, pp. 9–16, 2015. doi: 10.1109/SciVis.2015.7429486 1
- [50] M. Pont, J. Vidal, J. Delon, and J. Tierny. Wasserstein distances, geodesics and barycenters of merge trees. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):291–301, 2022. doi: 10.1109/TVCG.2021.3114839 2
- [51] S. Popinet. Free computational fluid dynamics. *ClusterWorld*, 2(6), 2004. 6
- [52] Y. Qin, B. T. Fasy, C. Wenk, and B. Summa. A domain-oblivious approach for learning concise representations of filtered topological spaces for clustering. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):302–312, 2021. doi: 10.1109/TVCG.2021.3114872 2
- [53] P. Rosen, A. Seth, E. Mills, A. Ginsburg, J. Kamenetzky, J. Kern, C. R. Johnson, and B. Wang. Using contour trees in the analysis and visualization of radio astronomy data cubes. In *Topological Methods in Data Analysis and Visualization VI*, pp. 87–108, 2021. doi: 10.1007/978-3-030-83500-2_6 2
- [54] H. Saikia, H.-P. Seidel, and T. Weinkauff. Extended branch decomposition graphs: Structural comparison of scalar data. *Computer Graphics Forum*, 33(3):41–50, 2014. doi: 10.1111/cgf.12360 1, 2
- [55] H. Saikia, H.-P. Seidel, and T. Weinkauff. Fast similarity search in scalar fields using merging histograms. In *Topological Methods in Data Analysis and Visualization IV*, pp. 121–134, 2017. doi: 10.1007/978-3-319-44684-4_7 2
- [56] H. Saikia and T. Weinkauff. Global feature tracking and similarity estimation in time-dependent scalar fields. *Computer Graphics Forum*, 36(3):1–11, 2017. doi: 10.1111/cgf.13163 1, 2
- [57] Scientific visualization contest. <http://www.uni-kl.de/scviscontest/>, 2016. 6
- [58] TOSCA. <http://tosca.cs.technion.ac.il/>, 2024. 6
- [59] N. Shervashidze, P. Schweitzer, E. J. Van Leeuwen, K. Mehlhorn, and K. M. Borgwardt. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011. 2
- [60] K. Shin and T. Ishikawa. Linear-time algorithms for the subpath kernel. In *Proc. 29th Annual Symposium on Combinatorial Pattern Matching*, pp. 22:1–22:13. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018. doi: 10.4230/LIPIcs.CPM.2018.22 2
- [61] M. Soler, M. Plainchault, B. Conche, and J. Tierny. Lifted Wasserstein matcher for fast and robust topology tracking. In *IEEE Symposium on Large Data Analysis and Visualization*, pp. 23–33, 2018. doi: 10.1109/LDAV.2018.8739196 1, 9
- [62] R. Sridharamurthy, T. B. Masood, A. Kamakshidasan, and V. Natarajan. Edit distance between merge trees. *IEEE Transactions on Visualization and Computer Graphics*, 26(3):1518–1531, 2020. doi: 10.1109/TVCG.2018.2873612 1, 2, 5, 6, 7
- [63] R. Sridharamurthy and V. Natarajan. Comparative analysis of merge trees using local tree edit distance. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1518–1530, 2023. doi: 10.1109/TVCG.2021.3122176 1, 2
- [64] S. Tatikonda and S. Parthasarathy. Hashing tree-structured data: Methods and applications. In *IEEE International Conference on Data Engineering*, pp. 429–440. IEEE, 2010. doi: 10.1109/ICDE.2010.5447882 2
- [65] D. M. Thomas and V. Natarajan. Multiscale symmetry detection in scalar fields by clustering contours. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2427–2436, 2014. doi: 10.1109/TVCG.2014.2346332 2
- [66] J. Tierny, G. Favelier, J. A. Levine, C. Gueunet, and M. Michaux. The Topology ToolKit. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):832–842, 2018. doi: 10.1109/TVCG.2017.2743938 6
- [67] G. H. Weber, S. E. Dillard, H. Carr, V. Pascucci, and B. Hamann. Topology-controlled volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 13(2):330–341, 2007. doi: 10.1109/TVCG.2007.47 2
- [68] K. Werner and C. Garth. Unordered task-parallel augmented merge tree construction. *IEEE Transactions on Visualization and Computer Graphics*, 27(8):3585–3596, 2021. doi: 10.1109/TVCG.2021.3076875 2
- [69] F. Wetzels, M. Anders, and C. Garth. Taming horizontal instability in merge trees: On the computation of a comprehensive deformation-based edit distance. In *Topological Data Analysis and Visualization*, pp. 82–92, 2023. doi: 10.1109/TopoInVis60193.2023.00015 2
- [70] F. Wetzels and C. Garth. A deformation-based edit distance for merge trees. In *Topological Data Analysis and Visualization*, pp. 29–38, 2022. doi: 10.1109/TopoInVis57755.2022.00010 2
- [71] F. Wetzels, H. Leitte, and C. Garth. Branch decomposition-independent edit distances for merge trees. *Computer Graphics Forum*, 41(3):367–378, 2022. doi: 10.1111/cgf.14547 2
- [72] 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. In *IEEE Pacific Visualization Symposium*, pp. 101–110, 2017. doi: 10.1109/PACIFICVIS.2017.8031584 1
- [73] Z. Wood, H. Hoppe, M. Desbrun, and P. Schröder. Removing excess topology from isosurfaces. *ACM Transactions on Graphics*, 23(2):190–208, 2004. doi: 10.1145/990002.990007 2
- [74] K. Wu and S. Zhang. A contour tree based visualization for exploring data with uncertainty. *International Journal for Uncertainty Quantification*, 3(3), 2013. doi: 10.1615/Int.J.UncertaintyQuantification.2012003956 2
- [75] W. Wu and B. Li. Locality sensitive hashing for structured data: A survey. *arXiv preprint arXiv:2204.11209*, 2022. doi: 10.48550/arXiv.2204.11209 2
- [76] W. Wu, B. Li, L. Chen, J. Gao, and C. Zhang. A review for weighted MinHash algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2553–2573, 2020. doi: 10.1109/TKDE.2020.3021067 2, 3
- [77] W. Wu, B. Li, L. Chen, X. Zhu, and C. Zhang. K -ary tree hashing for fast graph classification. *IEEE Transactions on Knowledge and Data Engineering*, 30(5):936–949, 2017. doi: 10.1109/TKDE.2017.2782278 2
- [78] Z. Xu, L. Niu, J. Ji, and Q. Li. Structure-preserving hashing for tree-structured data. *Signal, Image and Video Processing*, 16(8):2045–2053, 2022. doi: 10.1007/s11760-022-02166-7 2, 3, 4
- [79] L. Yan, H. Guo, T. Peterka, B. Wang, and J. Wang. TROPHY: A topologically robust physics-informed tracking framework for tropical cyclones. *IEEE Transactions on Visualization and Computer Graphics*, 30:1302–1312, 2024. doi: 10.1109/TVCG.2023.3326905 2
- [80] L. Yan, T. B. Masood, F. Rasheed, I. Hotz, and B. Wang. Geometry-aware merge tree comparisons for time-varying data with interleaving distances. *IEEE Transactions on Visualization and Computer Graphics*, 29(8):3489–3506, 2023. doi: 10.1109/TVCG.2022.3163349 1, 2, 3, 5, 6, 9
- [81] L. Yan, T. B. Masood, R. Sridharamurthy, F. Rasheed, V. Natarajan, I. Hotz, and B. Wang. Scalar field comparison with topological descriptors: Properties and applications for scientific visualization. *Computer Graphics Forum*, 40(3):599–633, 2021. doi: 10.1111/cgf.14331 1, 2
- [82] L. Yan, Y. Wang, E. Munch, E. Gasparovic, and B. Wang. A structural average of labeled merge trees for uncertainty visualization. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):832–842, 2020. doi: 10.1109/TVCG.2019.2934242 1, 2
- [83] H. Zhang and Q. Zhang. Embedjoin: efficient edit similarity joins via embeddings. In *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 585–594, 2017. doi: 10.1145/3097983.3098003 2
- [84] H. Zhang and Q. Zhang. MinJoin: efficient edit similarity joins via local hash minima. In *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1093–1103, 2019. doi: 10.1145/3292500.3330853 2
- [85] H. Zhang and Q. Zhang. MinSearch: an efficient algorithm for similarity search under edit distance. In *Proc. 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 566–576, 2020. doi: 10.1145/3394486.3403099 2

A PSEUDOCODE

We provide pseudocode for the relevant algorithms described in the paper. [Algorithm 1](#) presents the pseudocode for the recursive MinHash. [Algorithm 2](#) presents the pseudocode for the modified recursive MinHash for merge trees. [Algorithm 3](#) presents the pseudocode for storing subpaths using modified DFS. [Algorithm 4](#) presents the pseudocode for generating subpath signatures for merge trees.

Algorithm 1: RMH

Data: Nested set $W^{(r)}$ at r th level, k hash functions at r th level, $\{\sigma_i^{(r)}\}_{r=1, i=1}^{R, k}$

Result: $\mathbf{h}^{(r)}$, fingerprint of $W^{(r)}$

```

1 begin
2   // Level 1
3   if  $r = 1$  then
4      $\mathbf{h}^{(r)} \leftarrow \text{MinHash}(W^{(r)}, \{(\sigma_i^{(r)})\}_{i=1}^k)$ 
5   end
6   else
7     for  $W_{(*)}^{(r-1)} \in W^{(r)}$  do
8       // Higher levels
9        $\mathbf{h}_*^{(r)} \leftarrow \text{RMH}(W_{(*)}^{(r-1)}, \{(\sigma_i^{(r)})\}_{r=1, i=1}^{R, k})$ 
10    end
11     $M \leftarrow |W^{(r)}|$ 
12     $L = \text{dim}(\mathbf{h}_*^{(r-1)})$ 
13    // Reorganize
14    for  $l \leftarrow 1$  to  $L$  do
15       $W_{(l)}^{(r)} = \{h_{m, (l)}^{r-1}\}_{m=1}^M$ 
16       $\mathbf{h}_{(l)}^{(r)} \leftarrow \text{MinHash}(W_{(l)}^{(r)}, \{(\sigma_i^{(r)})\}_{i=1}^k)$ 
17    end
18  end
19  // Concatenate
20   $\mathbf{h}^{(r)} = [\mathbf{h}_{(1)}^{(r)}; \mathbf{h}_{(2)}^{(r)}; \dots; \mathbf{h}_{(L)}^{(r)}]$ 
21 end

```

B HASH FUNCTIONS

We provide a brief description of how hashes used in [Fig. 3](#) are generated using random permutations. The illustrative example in [Fig. 3](#) shows that RMH can be constructed as follows.

Assume a universal set $U = \{a, b, c, \dots, j\}$ which can be indexed by a set $I = \{1, 2, 3, \dots, 10\}$. Now the sets $S_1 = \{a, b, c\}$, $S_2 = \{b, e\}$, $S_3 = \{d, e, a\}$ can be represented using a 0-1 encoding as shown in [Tab. 4](#).

We generate four random permutations $\pi_1, \pi_2, \pi_3, \pi_4$ on the index set I as follows,

$$\begin{aligned} \pi_1\{I\} &= \{3, 2, 4, 1, 5, 6, 7, 8, 9, 10\}, \\ \pi_2\{I\} &= \{4, 5, 3, 6, 8, 9, 10, 2, 1, 7\}, \\ \pi_3\{I\} &= \{1, 3, 5, 2, 4, 7, 9, 6, 8, 10\}, \\ \pi_4\{I\} &= \{1, 4, 7, 6, 5, 3, 8, 2, 9, 10\}. \end{aligned}$$

These permutations provide us four hash functions h_1, h_2, h_3 , and h_4 . We show results for h_1 , the rest of the results are similar.

MinHash stores the index of the first occurrence of 1 in the particular permutation for each of the sets. [Tab. 5](#) shows the permutation π_1 which is used to derive h_1 . For S_1 we see that the first 1 occurs for the index 3. Therefore, we have

$$\begin{aligned} h_1(S_1) &= 3, h_1(S_2) = 2, h_1(S_3) = 4, \\ h_2(S_1) &= 3, h_2(S_2) = 5, h_2(S_3) = 1, \\ h_3(S_1) &= 1, h_3(S_2) = 5, h_3(S_3) = 1, \\ h_4(S_1) &= 1, h_4(S_2) = 5, h_4(S_3) = 1. \end{aligned}$$

Algorithm 2: RMH Signatures for Merge Trees

Data: Set of nodes at r th level i.e $T^{(r)}$, K hash functions at r th level, i.e., $\{\sigma_k^{(r)}\}_{r=1, k=1}^{R, K}$

Result: $\mathbf{h}^{(r)}$, fingerprint of $T^{(r)}$

```

1 begin
2   for  $T_{(*)}^{(r)} \in T^{(r)}$  do
3     // Extrema
4     if  $\text{deg}(T_{(*)}^{(r)}) = 1$  then
5        $\mathbf{h}_*^{(r)} \leftarrow \text{MinHash}(T_{(*)}^{(r)}, \{(\sigma_k^{(r)})\}_{k=1}^K)$ 
6     end
7     else
8       // Saddles recursive call
9        $\mathbf{h}_*^{(r)} \leftarrow \text{RMH}(T_{(*)}^{(r)}, \{(\sigma_k^{(r)})\}_{r=1, k=1}^{R, K})$ 
10    end
11  end
12   $M \leftarrow |T^{(r)}|$ 
13   $L = \text{dim}(\mathbf{h}_*^{(r-1)})$ 
14  // Reorganize
15  for  $l \leftarrow 1$  to  $L$  do
16     $T_{(l)}^{(r)} = \{h_{m, (l)}^{r-1}\}_{m=1}^M$ 
17     $\mathbf{h}_{(l)}^{(r)} \leftarrow \text{MinHash}(T_{(l)}^{(r)}, \{(\sigma_k^{(r)})\}_{k=1}^K)$ 
18  end
19  // Concatenate
20   $\mathbf{h}^{(r)} = [\mathbf{h}_{(1)}^{(r)}; \mathbf{h}_{(2)}^{(r)}; \dots; \mathbf{h}_{(L)}^{(r)}]$ 
21 end

```

After reorganization, we get four sets given by

$$S_4 = \{3, 2, 4\}, S_5 = \{3, 5, 1\}, S_6 = \{1, 5, 1\}, S_7 = \{1, 5, 1\}.$$

Now [Tab. 6](#) shows 0-1 encoding for sets S_4, S_5, S_6 , and S_7 . We again apply the same hash functions for these sets to obtain the following hashes:

$$\begin{aligned} h_1(S_4) &= 3, h_1(S_5) = 3, h_1(S_6) = 1, h_1(S_7) = 1, \\ h_2(S_4) &= 4, h_2(S_5) = 5, h_2(S_6) = 5, h_2(S_7) = 5, \\ h_3(S_4) &= 3, h_3(S_5) = 1, h_3(S_6) = 1, h_3(S_7) = 1, \\ h_4(S_4) &= 4, h_4(S_5) = 1, h_4(S_6) = 1, h_4(S_7) = 1. \end{aligned}$$

We finally concatenate them to get the recursive MinHash signature of the three input sets S_1, S_2 , and S_3 :

$$[3, 4, 3, 4, 3, 5, 1, 1, 1, 5, 1, 1, 1, 5, 1, 1].$$

Algorithm 3: SS-VISIT

Data: $u \in T$, multiset SP to store subpaths

Result: Builds SP

```

1 begin
2   PUSH(S, u)
3   color[u] ← gray
4   for  $v \in \text{children}(u)$  do
5     if color[v] = white then
6       SS-VISIT(v)
7     end
8   end
9   color[u] ← black
10  sp ← POP-PUSH(t, t-1)
11  sp ← REVERSE(sp)
12  SP ← SP + sp
13 end

```

Algorithm 4: SS Signatures for Merge Trees

Data: Tree T , subpath length t

Result: Multiset SP containing all subpaths of length t .

```

1 begin
2    $D \leftarrow \{d_1, d_2, \dots, d_{t-1}\}$ 
    $T' \leftarrow \text{ConcatenateDummyNodes}(D, T)$ 
3    $S \leftarrow \emptyset$ 
4   for  $u \in V(T')$  do
5     |  $\text{color}[u] \leftarrow \text{white}$ 
6   end
7   SS – VISIT( $\text{root}(T')$ )
8   return  $SP$ 
9 end

```

U	I	S_1	S_2	S_3
a	1	1	0	1
b	2	1	1	0
c	3	1	0	0
d	4	0	0	1
e	5	0	1	1
f	6	0	0	0
g	7	0	0	0
h	8	0	0	0
i	9	0	0	0
j	10	0	0	0

Table 4: Sets S_1, S_2, S_3 represented using a 0-1 encoding along with the universal set U and the index set I .

While we use permutations to generate hashes here, permutations are costly operations to be used in practice. Instead, other hashes such as MD5 can be used in practice. In addition, the permutations used in this example were not generated randomly. They are used to define and illustrate the techniques.

C QUANTITATIVE EVALUATION

We provide additional quantitative evaluation results, including a comparison to a clustering algorithm, a comparison of precision and recall to runtime, and the runtime analysis as a function of LSH parameters r and k .

C.1 Comparison to Clustering Algorithm

In this section, we provide additional quantitative evaluation on comparison to a clustering algorithm, k -medoids. Unlike k -means, which requires a well-defined notion of the mean, the k -medoids algorithm uses actual data points as the cluster centers. We use the TOSCA shape dataset to compare the precision and recall of each cluster with our results obtained using the LSH methods. We use the merge tree edit distance matrix as input to the k -medoids algorithm. Fig. 17 shows an evaluation of the number of clusters, n , and the precision and recall scores using k -medoids.

The precision generally increases as n increases because we would get a higher ratio of correctly predicted labels to all the labels in the cluster. Conversely, the recall generally decreases as n increases because the ratio of correctly predicted labels in the same cluster becomes smaller. When $n = 10$, which is the ground truth clustering number in the TOSCA dataset (since we have 10 different shapes), both precision and recall are below 0.5. These values are at the same level as both d_S and d_R when $r = 2$, and they are below the scores obtained when $r = 1$.

C.2 Runtime Analysis

We perform additional analysis on how runtime directly affects the performance of the LSH methods and the relationship between runtime and LSH parameters, r and k .

We investigate the relationship between k and runtime, demonstrating that k can be a representation of runtime, and showing that the

U	I	S_1	S_2	S_3
c	3	1	0	0
b	2	1	1	0
d	4	0	0	1
a	1	1	0	1
e	5	0	1	1
f	6	0	0	0
g	7	0	0	0
h	8	0	0	0
i	9	0	0	0
j	10	0	0	0

Table 5: permutation π_1 resulting in hash function h_1 .

U	S_4	S_5	S_6	S_7
1	0	1	1	1
2	1	0	0	0
3	1	1	0	0
4	1	0	0	0
5	0	1	1	1
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0

Table 6: Sets S_4, S_5, S_6, S_7 represented using 0 – 1 encoding, we take the universal set U and the index set I to be the same since they are numbers

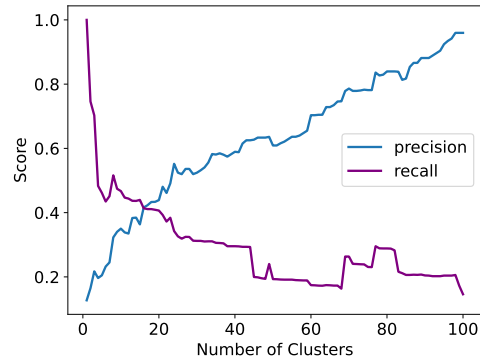


Fig. 17: TOSCA shape dataset: precision and recall using K-Medoids clustering. Number of clusters vary from 1 to 100.

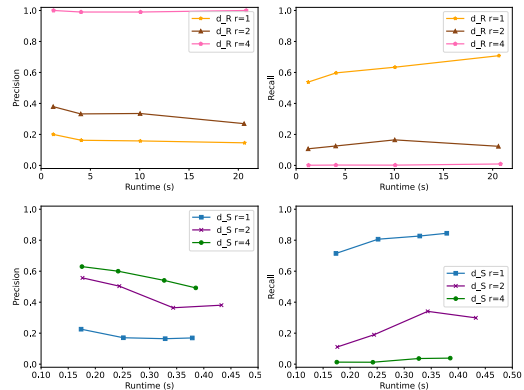


Fig. 18: TOSCA shape dataset: runtime vs. precision and recall for both d_S and d_R .

trends of runtime vs. precision and recall are similar to those of k vs. precision and recall. In this section, we explicitly provide an evaluation of runtime vs. precision and recall for both d_S and d_R (see Fig. 18).

The primary computational cost is computing the number of hash functions of each data object, but there are other lower-order terms that do not depend on the size of each object. For instance, the number of rows, r , is a lower-order term that does not noticeably affect the runtime. We provide Tab. 7 and Tab. 8 to demonstrate that changing r has little affect on the runtime with the same k , for both d_S and d_R , respectively. The runtime remains dominated by k .

r	$k = 20$	$k = 40$	$k = 60$	$k = 80$
1	0.17375625	0.25166304	0.32877218	0.37872514
2	0.17633292	0.24426496	0.34348022	0.43249424
4	0.17529604	0.24204648	0.32700760	0.38517716

Table 7: d_S for TOSCA shape dataset: r vs. All times are in seconds.

r	$k = 20$	$k = 40$	$k = 60$	$k = 80$
1	1.24099221	4.07319212	10.06167383	20.63537765
2	1.22652745	4.02467799	10.03928399	20.58727741
4	1.26209531	4.05044847	10.08340383	20.80203819

Table 8: d_R for TOSCA shape dataset: r vs. Runtime. All times are in seconds.

C.3 Additional Visualization Results

We provide additional visualization results of the Heated Flow dataset, showing the original binary matrices without any annotation in Fig. 19. The clusters are still visible without the annotated boxes.

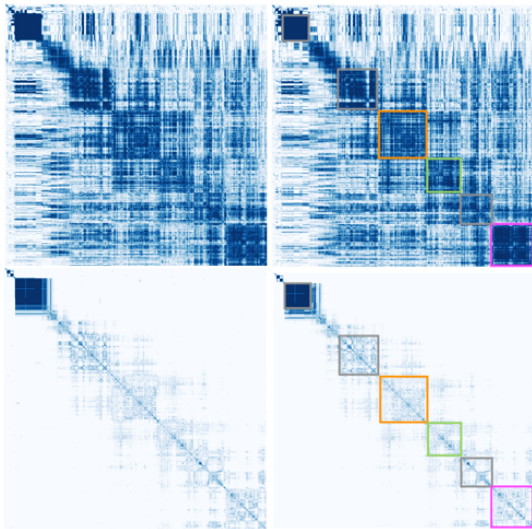


Fig. 19: Heated Flow dataset: d_S (top) and d_R (bottom) binary matrices with (right) and without (left) annotated boxes surrounding the clusters.