

# I Want 'Em All (At Once) – Ultrametric Cluster Hierarchies

Andrew Draganov <sup>\*1</sup> Pascal Weber <sup>\*234</sup> Rasmus Skibdahl Melanchton Jørgensen <sup>1</sup> Anna Beer <sup>2</sup>  
 Claudia Plant <sup>24</sup> Ira Assent <sup>1</sup>

## Abstract

Hierarchical clustering is a powerful tool for exploratory data analysis, organizing data into a tree of clusterings from which a partition can be chosen. This paper generalizes these ideas by proving that, for any reasonable hierarchy, one can optimally solve any center-based clustering objective over it (such as  $k$ -means). Moreover, these solutions can be found exceedingly quickly and are *themselves* necessarily hierarchical. Thus, given a cluster tree, we show that one can quickly access a plethora of new, equally meaningful hierarchies. Just as in standard hierarchical clustering, one can then choose any desired partition from these new hierarchies. We conclude by verifying the utility of our proposed techniques across datasets, hierarchies, and partitioning schemes.

## 1. Introduction

Hierarchical clustering is a fundamental technique for exploratory data analysis (Ward Jr, 1963; Johnson, 1967). The key idea is that having a tree of clusterings over a given dataset allows users to choose any partition from this hierarchy that best suits the users' needs. These notions go beyond standard agglomerative clustering algorithms to also include density-based clustering techniques like DBSCAN (Ester et al., 1996) and HDBSCAN (Campello et al., 2013). Even non-hierarchical clustering algorithms like  $k$ -means can be solved efficiently by leveraging hierarchical representations (Mettu & Plaxton, 2003; Cohen-Addad et al., 2020; 2021).

The central underlying concept of hierarchical clustering methods is that they can be modeled via ultrametrics: dis-

<sup>\*</sup>First authors with equal contribution in alphabetical order. Each first author reserves the right to put his name first on CVs or other documents. <sup>1</sup>Department of Computer Science, Aarhus University, Aarhus, Denmark <sup>2</sup>Data Mining and Machine Learning, University of Vienna, Vienna, Austria <sup>3</sup>UniVie Doctoral School Computer Science, University of Vienna, Vienna, Austria <sup>4</sup>Data Science @ Uni Vienna, University of Vienna, Vienna, Austria. Correspondence to: Andrew Draganov <[draganovandrew@cs.au.dk](mailto:draganovandrew@cs.au.dk)>, Pascal Weber <[pascal.weber@univie.ac.at](mailto:pascal.weber@univie.ac.at)>.

tance functions satisfying the strong triangle inequality  $d(x, z) \leq \max(d(x, y), d(y, z))$  for all  $x, y, z$ . Originally described for agglomerative clustering tasks such as single-linkage clustering (Milligan, 1979), the depth of this connection between ultrametrics and hierarchical clustering has inspired multiple subfields of clustering theory focused on finding the best-fitting ultrametric for a given dataset (Bartal, 1998; Indyk, 2004; Dasgupta, 2016; Cohen-Addad et al., 2019; Chierchia & Perret, 2020).

In this paper, we prove an elegant property of ultrametrics: *all* standard center-based clustering tasks, including  $k$ -means,  $k$ -median, and  $k$ -center, can be solved optimally in an ultrametric. This improves on recent work in center-based clustering in trees (Cohen-Addad et al., 2021; Beer et al., 2023; Lang & Schubert, 2023). Remarkably, our algorithm is extremely efficient: finding optimal solutions for all  $k \in \mathbb{N}_n = \{1, \dots, n\}$  on a dataset of  $n$  points requires only the time to sort  $O(n)$  values. Moreover, these partitions are inherently hierarchical: the optimal center-based clustering solutions in an ultrametric themselves form a cluster-tree.

As a result, this significantly expands the versatility of hierarchical clustering. Not only can one choose partitions from a given cluster hierarchy, we show that one can also efficiently obtain *new* hierarchies from it. We showcase this expanded utility by studying our results under two common ultrametrics: hierarchically well-separated trees (Bartal, 1998) and density-connectivity (Beer et al., 2023). The former is instrumental to accelerating countless machine learning problems (Gray & Moore, 2000; March et al., 2010; Van Der Maaten, 2014). The latter is the backbone of density-based clustering algorithms like DBSCAN or HDBSCAN.

We conclude by verifying the utility of these ultrametric, hierarchy, and partition combinations. As an example, our framework can fully reproduce HDBSCAN. However, once this HDBSCAN clustering has been produced, one also has access to *any* additional hierarchy and *any* additional partition, essentially for free. We show that many of our new combinations indeed yield competitive clustering results. Thus, our approach is ideally suited for exploratory data analysis.

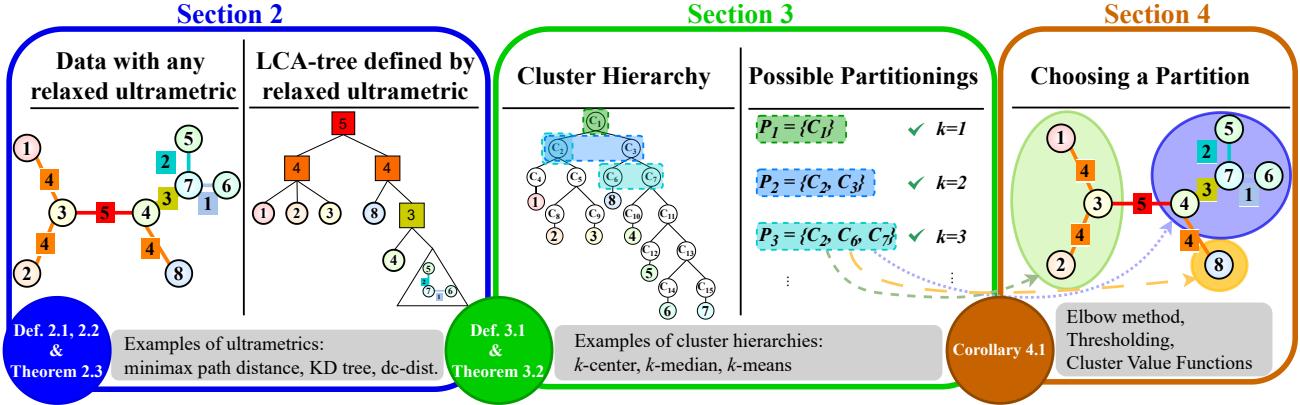


Figure 1. Sec. 2 introduces relaxed ultrametrics and how they correspond to lowest common ancestor (LCA) trees. Based on LCA trees, Sec. 3 shows how various objectives lead to cluster hierarchies. Sec. 4 shows how to partition these hierarchies into desirable clusterings.

Our main contributions are:

1. A theoretical derivation that all center-based clustering objectives can be solved optimally in ultrametrics for all values of  $k$  in the time it takes to sort  $n$  values.
2. A generalization of hierarchical clustering that produces a diverse set of known and novel algorithms essentially simultaneously.
3. A thorough experimental evaluation of our framework across ultrametrics and datasets verifying its utility.

Figure 1 shows the structure of this paper: We introduce ultrametrics and their hierarchical representations in Section 2. Section 3 contains our primary theoretical contribution that center-based clustering tasks like  $k$ -means can be solved optimally in ultrametrics, resulting in new cluster hierarchies. Section 4 shows ways to choose a partition from those hierarchies. We conclude with a thorough experimental analysis highlighting the speed and versatility of our proposed SHIP (Similarity-Hierarchy-Partition) clustering framework.

## 2. Ultrametrics and Tree Representations

We begin by formally introducing the data structure representing ultrametric spaces as it facilitates solving center-based clustering tasks quickly. The upcoming results operate over a generalization of the standard ultrametric:

**Definition 2.1.** Let  $L$  be a set. Then  $d : L \times L \rightarrow \mathbb{R}_{\geq 0}$  is a *relaxed ultrametric* over  $L$  if, for all  $\ell_i, \ell_j, \ell_k \in L$ , the following conditions are satisfied:

$$\begin{aligned} d(\ell_i, \ell_j) &= d(\ell_j, \ell_i) \geq 0 \\ d(\ell_i, \ell_k) &\leq \max(d(\ell_i, \ell_j), d(\ell_j, \ell_k)). \end{aligned}$$

Note that the standard ultrametric is a restriction that additionally requires  $d(\ell_i, \ell_i) = 0$ . Thus, not all relaxed ultrametrics are distances as  $d(\ell_i, \ell_i) \geq 0$  is allowed. Still, we

use the word “distance” for readability. We represent relaxed ultrametric relationships via the following data structure:

**Definition 2.2.** A *lowest-common-ancestor tree (LCA-tree)* is a rooted tree  $T$  such that every node  $\eta \in T$  has value  $d(\eta) \geq 0$  associated with it. We write  $\eta_i \preceq \eta_j$  to indicate that  $\eta_j$  lies on the path from  $\eta_i$  to the root and  $\eta_i \vee \eta_j$  to refer to the LCA of  $\eta_i$  and  $\eta_j$ . We say that the *LCA-distance* between two leaves  $\ell_i, \ell_j \in T$  is given by  $d(\ell_i \vee \ell_j)$ .

An LCA-tree is not necessarily binary: if three or more subtrees are all equidistant, they can all be children of the same node. While similar data structures already exist for standard ultrametrics (Bartal et al., 2003; Hoseini & Haghir Chehreghani, 2022), the following theorem (proof in A.1) states that it can also encode all relaxed ultrametrics:<sup>1</sup>

**Theorem 2.3.** Let  $(L, d')$  be a finite relaxed ultrametric space. Then there exists LCA-tree  $T$  with LCA-distance  $d$  and a bijection  $f : L \leftrightarrow \text{leaves}(T)$  such that, for all  $\ell_i, \ell_j \in L$ ,  $d'(\ell_i, \ell_j) = d(f(\ell_i) \vee f(\ell_j))$ .

This is visualized by the first box of Figure 1. It shows a minimum spanning tree (MST) over some data on the left. In this MST, the minimax ultrametric is given by the weight of the largest edge in the path between two nodes (Fischer et al., 2003). For example, the minimax distance between nodes (1) and (4) is [5]. The right-hand side of the first box of Figure 1 then stores these minimax distances in an LCA-tree where the LCA of nodes (1) and (4) has value 5.

We show in Appendix A.1 that all LCA-trees are relaxed ultrametrics as long as they satisfy the following conditions:

**Corollary 2.4.** Let  $T$  be an LCA-tree. For any leaf  $\ell \in T$ , let  $p(\ell) = [\ell, \eta_a, \dots, \eta_b, r(T)]$  be the path from  $\ell$  to the

<sup>1</sup>Note that *relaxed* ultrametrics cannot be represented via shortest-path distances through a tree as is commonly done for ultrametrics (Page & Holmes, 2009) as  $d(\ell_i, \ell_i) > 0$  is possible.

root of the tree  $r(T)$ . Then the LCA-distances on  $T$  form a relaxed ultrametric if and only if, for all  $\ell \in \text{leaves}(T)$  and  $\eta_i, \eta_j \in p(\ell)$ , the following conditions are satisfied:

$$(1) \quad d(\ell) \geq 0 \quad \text{and} \quad (2) \quad \eta_i \preceq \eta_j \implies d(\eta_i) \leq d(\eta_j).$$

Corollary 2.4 describes a key property: if a tree's node values grow along paths from the leaves to the root, then the corresponding LCA-distances are a relaxed ultrametric. This is the natural representation of a hierarchy: since the subtrees grow in size as we go towards the root, the values corresponding to those subtrees also grow. Thus, relaxed ultrametrics and hierarchies are equivalent notions. Going forward, we assume that a relaxed ultrametric is given in its LCA-tree form, satisfying the properties in Corollary 2.4.

### 3. Center-based clustering in ultrametrics

Our primary theoretical result is that the  $k$ -means,  $k$ -median, and  $k$ -center objectives can be solved optimally in a relaxed ultrametric. Furthermore, these solutions are hierarchical and can all be found in  $\text{Sort}(n)$  time.<sup>2</sup>

**Notation.** We define the  $(k, z)$ -clustering and the  $k$ -center clustering objectives over  $T$  as finding the set of centers  $\mathbf{C} \subseteq \text{leaves}(T)$  with  $|\mathbf{C}| = k$  which minimize, respectively,

$$\begin{aligned} \text{Cost}_z(T, \mathbf{C}) &= \underbrace{\sum_{\ell \in \text{leaves}(T)} \min_{c \in \mathbf{C}} d(\ell, c)^z}_{(k, z)\text{-clustering objective}}, \\ \text{Cost}_\infty(T, \mathbf{C}) &= \underbrace{\max_{\ell \in \text{leaves}(T)} \min_{c \in \mathbf{C}} d(\ell, c)}_{k\text{-center clustering objective}}. \end{aligned}$$

Note that the  $(k, z)$ -clustering task gives us the well-known  $k$ -median and  $k$ -means tasks for  $z = 1$  and  $z = 2$ . Let us now define what it means for a clustering to be hierarchical. Given a set of points  $L$ , we define a *cluster*  $C$  as any subset of  $L$ . We then define a *partition*  $\mathcal{P}_k = \{C_1, \dots, C_k\}$  as any *non-overlapping* set of  $k$  clusters, i.e., for all  $C_i, C_j \in \mathcal{P}$ ,  $C_i \cap C_j = \emptyset$ . We now define cluster hierarchies:

**Definition 3.1.** [Lin et al. (2006)] A *cluster hierarchy*  $\mathcal{H} = \{\mathcal{P}_1, \dots, \mathcal{P}_n\}$  is a set of partitions with

1. for  $k = 1$ ,  $\mathcal{P}_1 \subseteq L$ , and
2. for  $1 < k \leq n$ ,  $\mathcal{P}_k = (\mathcal{P}_{k-1} \setminus C_i) \cup \{C_j, C_l\}$ , such that  $C_i = C_j \cup C_l \in \mathcal{P}_{k-1}$  with  $C_j \cap C_l = \emptyset$  and  $i \neq j \neq l$ .

We say a cluster  $C_i \in \mathcal{H}$  if  $\exists \mathcal{P}_j \in \mathcal{H}$  with  $C_i \in \mathcal{P}_j$ .

<sup>2</sup>I.e., the time required to sort  $O(n)$  distances in the metric. Although sorting is thought of as requiring  $O(n \log n)$  time, in many settings, it only requires  $O(n \log \log n)$  time (Han, 2002).

This means that the partition  $\mathcal{P}_k$  is obtained by splitting a cluster from  $\mathcal{P}_{k-1}$  in two. Thus, all cluster hierarchies can be represented as a rooted tree. We depict this in Figure 1 (middle), where the hierarchy is obtained by subdividing the clusters. We can now state our primary theoretical result:

**Theorem 3.2.** Let  $(L, d)$  be a finite relaxed ultrametric space represented over an LCA-tree  $T$ . Let  $n = |L|$  and  $z \in \mathbb{Z}_{>0}$ . Then, for both the  $k$ -center and  $(k, z)$ -clustering objectives on  $T$ , there exists an algorithm which finds the optimal solutions  $\{\mathbf{C}_1, \dots, \mathbf{C}_n\}$  for all  $k \in \mathbb{N}_n$  in  $\text{Sort}(n)$  time. Furthermore, the respective partitions  $\mathcal{H} = \{\mathcal{P}_1, \dots, \mathcal{P}_n\}$  obtained by assigning all leaves in  $T$  to their closest center satisfy Definition 3.1.

Theorem 3.2 states that given any LCA-tree over a relaxed ultrametric, it takes  $\text{Sort}(n)$  time to find the optimal solutions for  $k$ -means,  $k$ -median or  $k$ -center across all values of  $k$  and that these solutions are hierarchical. While subsets of this are known for specific ultrametrics (Cohen-Addad et al., 2021; Beer et al., 2023), our runtime improves over the previous state of the art, and we are not aware of a comprehensive treatment of this subject over all ultrametrics and center-based clustering objectives.

**Proof Overview.** The  $k$ -center part of this result is an extension of the farthest-first traversal algorithm where one picks each subsequent center as the point that is farthest from the current set of centers (Hochbaum & Shmoys, 1985; Har-peled, 2011). In essence, Appendix B.1 shows that the farthest-first traversal algorithm becomes optimal under the strong triangle inequality. Then, we show in Appendix B.2 that the  $(k, z)$ -clustering objectives in a relaxed ultrametric can be reduced to the  $k$ -center one. Specifically, given an LCA-tree  $T$  on which to do  $(k, z)$ -clustering, there exists a new LCA-tree  $T'$  such that an optimal  $k$ -center solution on  $T'$  is the optimal  $(k, z)$ -clustering solution on  $T$ . Interestingly, even if this reduction is applied in a standard ultrametric, it requires the definition of relaxed ultrametrics to go through. Finally, the runtime bottleneck for  $k$ -center lies in sorting the  $O(n)$  unique distances in the ultrametric so that we may apply farthest-first traversal. The reduction from  $(k, z)$ -clustering only requires  $O(n)$  time. Thus, the bottleneck remains unchanged for  $(k, z)$ -clustering. We verify by a worst-case example that our runtime is tight: one cannot find the optimal centers for all values of  $k \in \mathbb{N}_n$  faster than in  $\text{Sort}(n)$  time (Lemma B.2).

**Takeaway.** An interesting consequence of Corollary 2.4 and Theorem 3.2 is that these center-based cluster hierarchies *themselves* constitute relaxed ultrametrics. That is, let  $\mathcal{H}$  be a hierarchy of optimal  $k$ -center or  $(k, z)$ -clustering solutions from Theorem 3.2. For each cluster  $C$  in this hierarchy, consider the cost  $d_{\text{cost}}(C)$  of assigning all of its points to a single optimal center. As we traverse the

tree towards the root, these costs will necessarily grow (see Lemma B.4 in Appendix B.2). Thus, by Corollary 2.4, the hierarchy  $\mathcal{H}$  with LCA-distances  $d_{cost}$  must be a relaxed ultrametric. This is precisely what facilitates our reduction from the  $(k, z)$ -clustering setting to the  $k$ -center one. Indeed, any LCA-tree satisfying the properties in Corollary 2.4 is its own optimal  $k$ -center hierarchy (non-binary LCA-trees can be binarized while preserving the LCA-distances to recover the equivalence to the  $k$ -center hierarchy).

## 4. Choosing a Partition

Theorem 3.2 gives us the optimal center-based clustering solutions for all values of  $k$  at once. If we simply require the solution for a specific  $k$  then this is sufficient. However, what if we are instead interested in the “best” clustering from this hierarchy? To this end, this section discusses how known techniques for obtaining a partition can be applied to Section 3’s hierarchies in  $O(n)$  time.

### 4.1. Cluster Selection Criteria

#### 4.1.1. FEASIBILITY OF THE ELBOW METHOD

One of the most common methods for choosing a “best” clustering is the elbow method (Thorndike, 1953). Here, one is given a set of values of  $k$ ,  $[k_1, k_2, \dots, k_f]$ , and a set of corresponding partitions  $\mathcal{P} = [\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_f]$ . Each partition incurs a cost  $\mathcal{L} = [\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_f]$  with respect to the clustering objective, where  $\mathcal{L}_i = \sum_{C \in \mathcal{P}_i} \mathcal{L}(C)$ . This gives us a plot of costs over the different values of  $k$ . Informally, the *elbow method* chooses the partition  $\mathcal{P}_i$  whose cost  $\mathcal{L}_i$  looks to be at a sharp point in this curve.

Due to the NP-hardness of  $k$ -means clustering (Dasgupta, 2008), standard elbow plots can only compute approximate solutions, traditionally done sequentially for each  $k$  (Schubert, 2023). However, the elbow method is surprisingly viable in the ultrametric setting: Theorem 3.2 yields optimal clusterings for all  $k$  values at once, eliminating both computational overhead and approximation errors. Moreover, we show that the relaxed ultrametric’s elbow plot for  $(k, z)$ -clustering is guaranteed to be convex:

**Corollary 4.1.** *Let  $\mathcal{P}$  and  $\mathcal{L}$  correspond to the  $n$  partitions and losses obtained in accordance with Theorem 3.2 for the  $(k, z)$ -clustering objective. Let  $\Delta_i = \mathcal{L}_{i+1} - \mathcal{L}_i$ . Then either  $\Delta_i < \Delta_{i+1} \leq 0$  or  $\Delta_i = \Delta_{i+1} = 0$  for all  $i \in [n - 1]$ .*

The idea here is that  $\Delta_i$  represents the elbow plot’s first derivative at  $k = i$ . Thus, Corollary 4.1 states that the elbow plot’s slope is steepest at  $k = 1$  and monotonically levels out to 0 as  $k \rightarrow n$ . We prove this in Appendix C.1, where we also specify how we determine the index of the elbow. We depict relaxed ultrametric elbow plots in the  $(k, z)$ -clustering setting for various values of  $z$  in Figure 2,

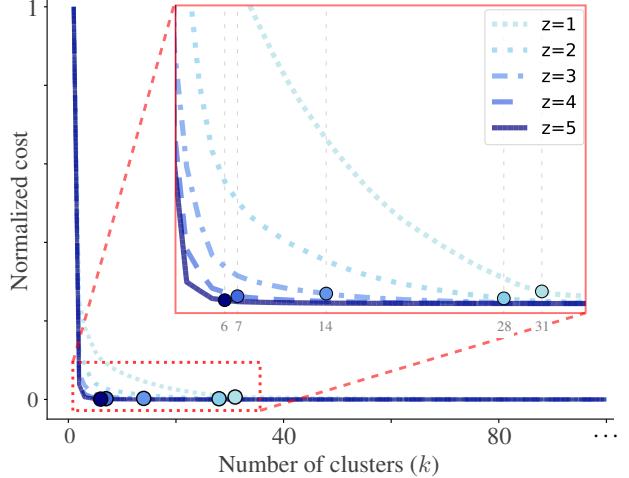


Figure 2. Example elbow plots for the  $z = 1, 2, 3, 4$ , and  $5$  settings under the dc-dist ultrametric on the D31 synthetic dataset. Elbow locations (circles) are determined using all  $k$  from  $1$  to  $n$ .

where we see that the plots are indeed convex.

#### 4.1.2. THRESHOLDING THE TREE

While Corollary 4.1 shows that the elbow method is reasonable in the ultrametric setting, it still has two main drawbacks. *First*, the choice of the elbow is arbitrary, and it is unclear whether the best technique exists (Schubert, 2023). *Second*, the elbow method is restricted to a single partition for each value of  $k$ . Although the cluster hierarchy can be cut in many ways to obtain  $k$  clusters, the elbow method can only produce one of these partitions for any value of  $k$ .

One alternative is to threshold the cluster hierarchy, as done in single-linkage clustering or DBSCAN (see Campello et al. (2013)). Specifically, let  $\varepsilon$  be any user-defined threshold parameter. We now (1) label the nodes of our cluster hierarchy by their costs as discussed at the end of Section 3 and (2) report all clusters whose cost is less than  $\varepsilon$ , both in  $O(n)$  time. We further discuss this in Appendix C.3.

#### 4.1.3. CLUSTER VALUE FUNCTIONS

Rather than picking clusters based on their costs, one can also assign a *new* value function to clusters and choose the partition that maximizes the sum of these values. Importantly, such a partition may not be attainable by choosing a value of  $k$  or thresholding the clusters’ costs in the hierarchy. For example, HDBSCAN uses the *stability* objective to choose the clusters from a hierarchy that persist for the largest range of threshold values. Given any cluster hierarchy obtained via Theorem 3.2 and any reasonable value function, one can find the partition that maximizes this value function in  $O(n)$  time. We detail this in Appendix C.4.

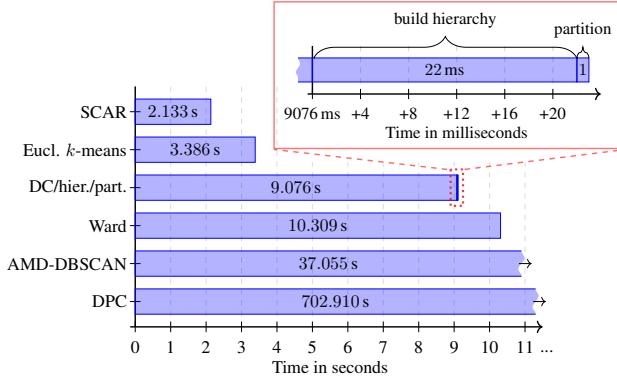


Figure 3. Runtimes of competitors and our framework’s components on letterrec dataset. Computing clusterings on the dc-dist is faster than other density-based methods, but building the hierarchy and partitioning it, is still several orders of magnitude faster.

#### 4.2. Generalizing to Noisy Settings

For a given LCA-tree  $T$  and any node  $\eta \in T$ , one can remove the subtree rooted at  $\eta$  from  $T$  without affecting Corollary 2.4. This allows handling additional noise points consistently, e.g., Lim et al. (2005); Campello et al. (2013).

For example, consider a minimum-cluster-size parameter  $\mu \in \mathbb{Z}_+$  and a cluster hierarchy  $\mathcal{H}$  obtained via Theorem 3.2. Then, for some cluster  $C_i$  with  $|C_i| < \mu$ , we can prune the hierarchy by removing the cluster trees rooted at  $C_i$  in  $O(n)$  time (Campello et al., 2013). We can, therefore, obtain any desired clustering over this pruned hierarchy, guaranteeing that every cluster in the returned partition will have a size greater than or equal to the minimum cluster size.

#### 4.3. Integrating Multiple Partition Methods

Lastly, we note that we can combine results of several partitions without notable impact on runtime. As Section 6 confirms, fitting an ultrametric essentially always requires longer than  $O(n \log n)$  time. Given such an ultrametric, we can find hierarchies of optimal center-based clustering solutions in  $\text{Sort}(n) \leq O(n \log n)$  time (Section 3). Furthermore, we presented several  $O(n) < \text{Sort}(n)$ -time methods for choosing a partition. As Figure 3 highlights, the time it takes to fit the density-connectivity ultrametric dwarfs the time for finding a hierarchy or partition. Consequently, once we have fit an ultrametric, we can calculate multiple hierarchies and partitions and integrate their information in negligible time. To illustrate this, we introduce a method for choosing the number of clusters that we find to work well in practice. Given an ultrametric, our *Median-of-Elbows* (MoE) algorithm starts by fitting the  $(k, z)$ -cluster hierarchies for  $z = 1, 2, 3, 4$ , and 5. Intuitively, each of these hierarchies has a different penalty for large distances. Applying the elbow method to each hierarchy gives us a

set of five  $k$  values, each representing its hierarchy’s ‘best’ number of clusters. The MoE method picks the median  $k$  from this set, which is essentially the  $k$  that gives stable clusterings across values of  $z$ . Figure 2 visualizes this process on the D31 dataset. Here, MoE selects  $k = 14$  (see also the D31  $k$ -median/MoE cell in Figure 4b).

### 5. From Theory to Practice

Sections 2-4 describe what we refer to as the SHIP (Similarity-Hierarchy-Partition) clustering framework: fitting an ultrametric, choosing a hierarchy, and partitioning the data. Table 1 outlines various options for each component. We use the remainder of this paper to demonstrate that the SHIP framework includes diverse, effective clustering strategies and enables quickly finding the many clusterings contained in any ultrametric. We note that this is a particularly valuable feature for exploratory data analysis, where quickly switching between clustering methods is beneficial.

Table 1. Example ultrametric, hierarchy, and partition options. Our SHIP framework allows one to pick any ultrametric, pair it with any clustering hierarchy, and extract any partition.

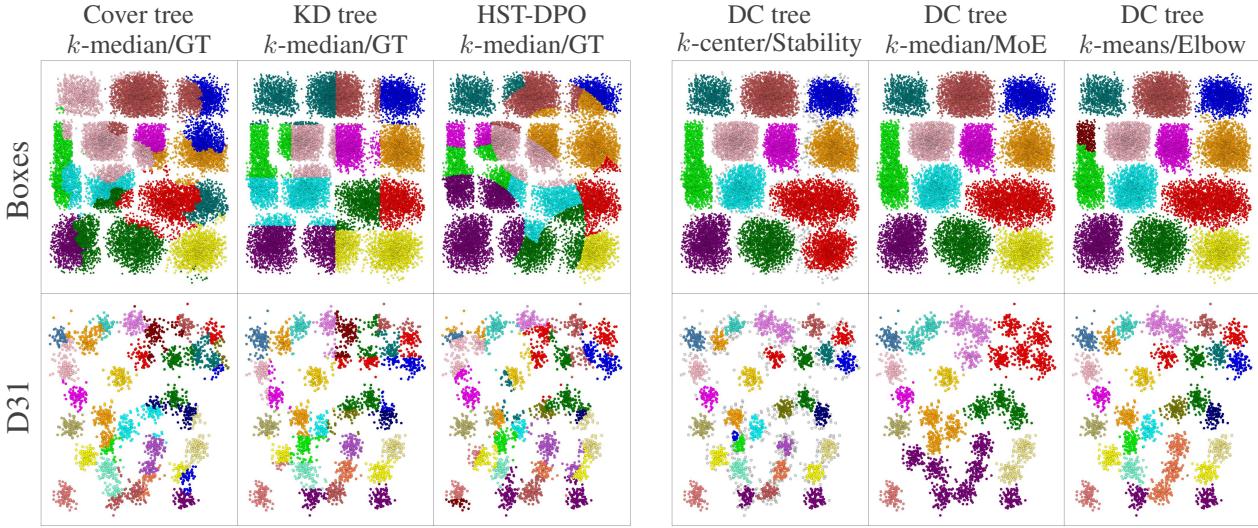
Ultrametrics	Hierarchies	Partitionings
DC tree	$k$ -center	Ground Truth $k$ (GT)
Single Linkage tree	$k$ -median	User-specified $k$
Cover tree	$k$ -means	Elbow
KD tree	...	Median of Elbows (MoE)
HST-DPO	...	Thresholding
...	...	Stability
...	...	...

#### 5.1. Ultrametrics

**HSTs.** Hierarchically well-separated trees (HSTs) are embeddings of a metric space into a tree structure so that distances between points are approximated using the tree’s path distances. There is a spectrum of available HST constructions, from those that minimize the distortion of the original distances (with the best possible distortion being  $O(\log n)$  (Fakcharoenphol et al., 2003)) to those that are quick to construct (Cohen-Addad et al., 2021). These are used regularly within computer graphics (Foley & Sugerman, 2005) and nearest-neighbor search (Ram & Sinha, 2019; Simhadri et al., 2022). Thus, in all these use cases, we provide access to a suite of optimal clusterings.

In HSTs, every internal node is equidistant to all its descendant leaves (Bartal, 1998). Thus, an HST can be transformed to an LCA-tree in  $O(n)$  time<sup>3</sup>. In our experiments, we use the well-known KD trees (Bentley, 1990), Cover

<sup>3</sup>For every internal node  $\eta$  in the HST with leaf  $\ell$  below it,  $d_{HST}(\eta, \ell)$  is known. Then, an LCA-tree can preserve the HST’s path distances by simply assigning  $d(\eta) = 2d_{HST}(\eta, \ell)$ .



*Figure 4.* A visualization of the various ultrametrics and hierarchy/partition combinations on the synthetic datasets *Boxes* and *D31*.

trees (Beygelzimer et al., 2006), and ‘theoretically optimal’ HST-DPOs (Zeng et al., 2021). We use the SHIP clustering framework to evaluate their differences in Figure 4a.

Each of these HSTs works by recursively subdividing the metric space into a set of nested cells. Cover trees are defined so that points  $p \in C_{i-1}$  and  $q \in C_i$  (where  $C_i \subset C_{i-1}$ ) satisfy  $d(p, q) < 2^i$  under ambient metric  $d$ . This can result in non-convex shapes, as evidenced by the optimal  $k$ -median solutions for the ground-truth value of  $k$  in Figure 4a. The KD tree, on the other hand, works by simply subdividing the space into nested hypercubes. Consequently, the KD tree’s  $k$ -median solutions in Figure 4a are the axis-aligned rectangles that best fit the data. These HSTs are fast to construct, requiring  $O(n \log n)$  time assuming constant dimensionality, but have poor distance preservation guarantees. Alternatively, the HST-DPO ultrametric achieves optimal distance preservation at the expense of an  $O(n^2)$  runtime. This works by the principle from Fakcharoenphol et al. (2003), wherein one recursively claims points by placing metric balls. Indeed, one can see this ball structure in Figure 4a’s HST-DPO  $k$ -median solutions on the Boxes dataset. Our remaining experiments use Cover trees, with full comparisons in the appendix.

**Density-Connectivity Distance (dc-dist).** The second ultrametric we use, the dc-dist, is the basis of density-connected clustering algorithms such as DBSCAN (Ester et al., 1996) and HDBSCAN (Campello et al., 2013). The dc-dist builds on mutual reachabilities between points to characterize local density:

**Definition 5.1** (Mutual reachability distance (Ester et al., 1996)). Let  $(L, d')$  be a metric space,  $x$  and  $y$  be any

two points in  $L$ , and  $\mu \in \mathbb{Z}_{>0}$ . Let  $\kappa_\mu(x)$  be the distance from  $x$  to its  $\mu$ -th closest neighbor in  $L$ . Then the mutual reachability between  $x$  and  $y$  is  $m_\mu(x, y) = \max(d'(x, y), \kappa_\mu(x), \kappa_\mu(y))$ .

Note that usually  $m_\mu(x, x) \geq 0$ . The dc-dist is then the minimax distance over the pairwise mutual reachabilities:

**Definition 5.2** (dc-dist (Beer et al., 2023)). Let  $(L, d')$  be a metric space,  $x$  and  $y$  be any two points in  $L$ , and  $\mu \in \mathbb{Z}_{>0}$ . Let  $T$  be an MST over  $L$ ’s pairwise mutual reachabilities. Let  $p(x, y)$  be the path in  $T$  from  $x$  to  $y$  given by edges  $\{e_i, \dots, e_j\}$  in  $T$ . Lastly, let  $|e|$  be the weight of any edge  $e$  in  $T$ . Then the dc-dist between  $x$  and  $y$  is defined as

$$d_{dc}(x, y) = \max_{e \in p(x, y)} |e| \text{ if } x \neq y; \text{ else } m_\mu(x, x).$$

Note that this definition slightly differs from the one in Beer et al. (2023) since they defined  $d_{dc}(x, x) = 0$  in order for it to be an ultrametric. However, our definition allows for the following proposition (proof in Appendix C.5):

**Proposition 5.3.** Let  $(L, d')$  be a metric space. Then, the dc-dist over  $L$  is a relaxed ultrametric.

(without border points)DC treeDC tree (see Appendix C.4). W the

## 5.2. Hierarchies and Partitioning methods

We now give an intuition for several hierarchy/partition combinations. First, recall that the *hierarchies* emphasize different criteria when merging clusters together (e.g., sum of squared distances for  $k$ -means vs. maximum distance

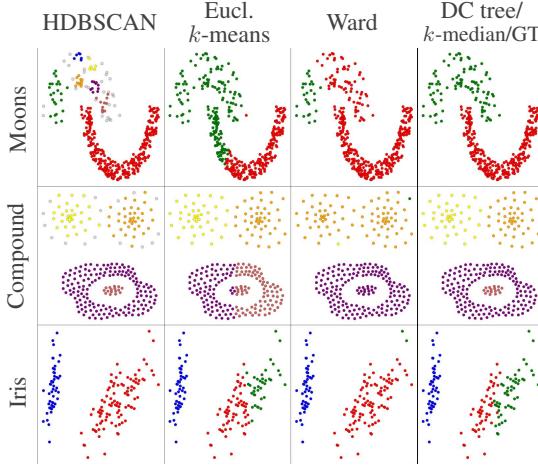


Figure 5. Clusterings on toy datasets indicated by color.

for  $k$ -center). Thus, the different hierarchies offer macro-level control over the available clusterings. From this, the *partitioning* methods then extract the hierarchy’s clustering, which optimizes for a specific goal, such as the number of clusters or the clusters’ stability.

Our experiments focus on the  $k$ -center/stability,  $k$ -median/MoE, and  $k$ -means/elbow combinations as illustrative examples. These are depicted over the dc-dist in Figure 4b. In the first column, the  $k$ -center hierarchy focuses solely on maximal cluster distances, merging clusters based on density-connectivity regardless of their number of points. The stability objective finds the sets of clusters that persist for the longest amount of time while allowing for noise points (a full description is in Appendix C.4). Thus, the

$k$ -center/stability clustering in the left column of Figure 4b separates the D31 dataset but combines the two red boxes as there is a density-connected path between them.

It may initially seem unintuitive to use a  $k$ -median or  $k$ -means objective over the dc-dist. However, it is a very natural notion. Consider that the  $k$ -means objective sums over per-point distances, and distance to a center under the dc-dist is the largest step away from the center’s densely-connected region. Thus, a sum over a cluster’s dc-dists is large if there are many points that are all spread out. To visualize this, notice that  $k$ -median/MoE and  $k$ -means/elbow separate the two red boxes, which were merged under  $k$ -center in the top row of Figure 4b: these two boxes had too many spread-out points to be in a single  $k$ -median or  $k$ -means cluster. Furthermore, the  $k$ -median hierarchy does not penalize large distances as severely as  $k$ -means. Consequently, notice that  $k$ -median/MoE merges some sufficiently density-connected clusters in D31, which  $k$ -means does not.

Lastly, we verify that our introduced hierarchies provide the ability to find previously unattainable clusterings with standard techniques. Specifically, Figure 5 shows several toy datasets where we see that HDBSCAN, Euclidean  $k$ -means, and Ward agglomerative clustering are all unable to correctly partition the data. However, each dataset’s correct clustering exists within the dc-dist’s  $k$ -median hierarchy.

## 6. Experiments

We now show the practical utility of the SHIP clustering framework: our new ultrametric/hierarchy/partition combinations are competitive with standard clustering methods,

Table 2. Runtimes of our SHIP framework’s components (first three groups of columns) and competitors (last column group) in minutes, seconds, and milliseconds [min:sec.ms]. Computation times of the ultrametrics are comparable to the runtimes of our competitors. Computation of the cluster hierarchies and partitioning methods take only *milliseconds* even on large and high-dimensional datasets. Thus, once the ultrametric is computed, we get arbitrarily many hierarchies and partitions in negligible time. Notably, while density-based methods (blue) are sometimes slower than others, building the DC tree is generally faster than its counterparts, AMD-DBSCAN or DPC.

Dataset	ultrametric		hierarchy	partitioning			competitors						
	Cover tree	D <small>DC</small> tree		$k$ -means	Stability	MoE	Elbow	$k$ -means ( $k = \text{GT}$ )	$k$ -means ( $k = 500$ )	SCAR	Ward	AMD- DBSCAN	D <small>DC</small>
Tabular Data	Boxes	00:00.059	00:14.239	00:00.033	00:00.003	00:00.134	00:00.002	00:00.114	00:01.217	00:01.481	00:10.808	01:04.670	12:10.358
	D31	00:00.004	00:00.327	00:00.004	00:00.000	00:00.015	00:00.000	00:00.289	00:00.509	00:00.620	00:00.153	00:00.878	00:12.816
	airway	00:00.027	00:04.997	00:00.021	00:00.001	00:00.085	00:00.001	00:00.122	00:00.392	00:00.651	00:05.243	00:16.822	06:46.726
	lactate	00:00.161	00:47.731	00:00.068	00:00.006	00:00.275	00:00.004	00:00.126	00:01.997	00:02.612	00:59.992	07:43.293	69:34.012
	HAR	00:11.053	00:23.144	00:00.014	00:00.001	00:00.056	00:00.000	00:02.248	00:08.478	00:00.658	00:18.448	00:30.129	03:45.745
	letterrec.	00:00.322	00:09.076	00:00.022	00:00.002	00:00.087	00:00.001	00:03.386	00:03.521	00:02.133	00:10.309	00:37.055	11:42.910
	PenDigits	00:00.067	00:02.566	00:00.013	00:00.001	00:00.053	00:00.000	00:01.001	00:02.725	00:00.490	00:03.281	00:09.254	03:21.409
Image Data	COIL20	00:03.658	00:14.817	00:00.001	00:00.000	00:00.004	00:00.000	00:01.637	00:13.524	00:00.310	00:08.941	00:02.378	00:11.710
	COIL100	03:52.397	14:50.661	00:00.010	00:00.000	00:00.039	00:00.000	00:32.122	01:59.225	00:07.964	12:04.037	02:58.177	14:18.780
	cmu_faces	00:00.046	00:00.238	00:00.001	00:00.000	00:00.002	00:00.000	00:00.206	00:00.661	00:00.116	00:00.064	00:00.346	00:00.515
	OptDigits	00:00.335	00:01.420	00:00.006	00:00.000	00:00.024	00:00.000	00:00.502	00:01.154	00:00.290	00:00.974	00:03.073	00:44.270
	USPS	00:02.924	00:08.670	00:00.011	00:00.001	00:00.045	00:00.000	00:02.709	00:08.493	00:01.433	00:06.092	00:29.259	01:48.607
	MNIST	16:24.220	37:05.095	00:00.120	00:00.015	00:00.491	00:00.008	00:04.320	03:29.969	00:15.352	19:02.498	17:21.183	-

Table 3. ARI values indicate that our framework achieves high clustering quality with an automated selection of  $k$ , matching even competitors' performance that relies on the GT number of clusters ( $k$ -means, SCAR, Ward). (Full version: Table 6 in the Appendix.)

Dataset	DC tree			Cover tree			competitors				AMD-DBSCAN	DPC
	$k$ -center Stability	$k$ -median MoE	$k$ -means Elbow	$k$ -center Stability	$k$ -median MoE	$k$ -means Elbow	Eucl. $k$ -means	SCAR	Ward			
Tabular Data	Boxes	90.1	<b>99.3</b>	<b>97.9</b>	2.6	$42.1 \pm 4.7$	$24.2 \pm 1.6$	$93.5 \pm 4.3$	$0.1 \pm 0.1$	<b>95.8</b>	63.9	25.9
	D31	79.7	42.7	82.9	$46.5 \pm 1.8$	$62.0 \pm 5.4$	$67.7 \pm 3.2$	<b>92.0 ± 2.7</b>	$41.7 \pm 5.4$	<b>92.0</b>	$86.4 \pm 0.1$	18.5
	airway	38.0	<b>65.9</b>	58.8	0.8	$18.2 \pm 2.4$	$12.0 \pm 1.4$	$39.9 \pm 2.0$	$-0.9 \pm 0.5$	43.7	31.7	<b>65.1</b>
	lactate	41.0	41.0	<b>67.5</b>	0.1	$4.1 \pm 0.6$	$1.7 \pm 0.2$	$28.6 \pm 1.1$	$1.5 \pm 1.0$	27.7	<b>71.5</b>	0.0
	HAR	30.0	46.9	<b>52.8</b>	$14.7 \pm 8.8$	$14.2 \pm 4.7$	$9.6 \pm 2.2$	$46.0 \pm 4.5$	$5.5 \pm 3.2$	<b>49.1</b>	0.0	33.2
	letterrec.	12.1	16.6	<b>17.9</b>	$5.8 \pm 0.2$	$7.2 \pm 0.6$	$6.2 \pm 0.3$	$12.9 \pm 0.6$	$0.4 \pm 0.1$	$14.7 \pm 0.9$	7.9	0.0
	PenDigits	66.4	<b>73.1</b>	<b>75.4</b>	$8.0 \pm 0.8$	$12.0 \pm 0.6$	$8.9 \pm 0.5$	$55.3 \pm 3.2$	$0.9 \pm 0.3$	55.2	55.6	$28.8 \pm 1.1$
Image Data	COIL20	<b>81.2</b>	72.8	72.6	$46.4 \pm 4.4$	$46.6 \pm 2.1$	$47.7 \pm 2.0$	$58.2 \pm 2.8$	$33.5 \pm 2.0$	68.6	39.2	$35.9 \pm 0.1$
	COIL100	<b>80.1</b>	66.8	<b>70.0</b>	$44.6 \pm 4.2$	$46.6 \pm 1.5$	$50.1 \pm 1.2$	$56.1 \pm 1.4$	$16.7 \pm 0.8$	61.4	14.2	0.2
	cmu_faces	60.2	56.6	<b>66.5</b>	$8.6 \pm 3.1$	$37.1 \pm 4.1$	$34.2 \pm 2.1$	$53.2 \pm 4.7$	$38.5 \pm 2.9$	<b>61.6</b>	0.7	0.6
	OptDigits	55.3	<b>77.0</b>	<b>77.0</b>	$40.9 \pm 3.5$	$20.9 \pm 2.3$	$18.1 \pm 2.4$	$61.3 \pm 6.6$	$14.4 \pm 4.1$	<b>74.6 ± 2.4</b>	<b>63.2</b>	0.0
	USPS	33.7	29.3	29.3	$12.0 \pm 1.7$	$8.7 \pm 1.0$	$11.2 \pm 1.5$	$52.3 \pm 1.7$	$2.9 \pm 0.9$	<b>63.9</b>	0.0	21.0
	MNIST	19.7	41.7	<b>46.0</b>	$11.1 \pm 1.7$	$5.4 \pm 0.6$	$5.4 \pm 0.6$	$36.9 \pm 1.0$	$1.3 \pm 0.4$	<b>52.7</b>	0.0	-

and once the ultrametric is constructed, one can switch between hierarchies and partitions in essentially no time.

Table 4 gives an overview of the datasets we use. We evaluate the clustering quality with the adjusted rand index (ARI) (Hubert & Arabie, 1985), treating points labeled as noise as singleton clusters. NMI (Estévez et al., 2009) results can be found in Appendix F. Our framework is implemented in C++ and provides a Python interface using pybind11. All experiments were performed with Python 3.12 on a Linux workstation with 2x Intel 6326 with 16 hyperthreaded cores each and 512GB RAM. Our code is available at <https://anonymous.4open.science/r/SHIP-42B1/>.

## 6.1. Runtimes

Clustering is an exploratory data mining task that, in practice, requires several runs of different methods or using different hyperparameter settings. Especially when done sequentially, this requires substantial computation time and resources. In contrast, our SHIP clustering framework requires just a single upfront ultrametric computation, after which we can generate multiple clustering solutions with negligible additional runtime. Table 2 and Figure 3 highlight this efficiency. The first column group in Table 2 shows that computing our ultrametrics (Cover tree and DC tree) has a runtime comparable to that of the corresponding standard clustering algorithms (last column group). Specifically, we show on the right under competitors the time required for Euclidean  $k$ -means with  $k = 10$ , Euclidean  $k$ -means with  $k = 500$ ,<sup>4</sup> agglomerative clustering (Ward Jr, 1963), an adaptive multi-density DBSCAN version (AMD-DBSCAN) (Wang et al., 2022), density peaks clustering (DPC) (Rodriguez & Laio, 2014), and an accelerated version of spectral clustering (SCAR) (Hohma et al., 2022).

<sup>4</sup>We note that the SHIP clustering framework simultaneously obtains the optimal solutions for every value of  $k$ .

Computing the ultrametrics takes time comparable to that of clustering algorithms following similar notions, e.g., density-based methods (shown in blue). However, once the ultrametric has been computed, generating different hierarchies (column group 2) and partitions (column group 3) takes on the order of *milliseconds*. Thus, users can explore many different clustering solutions in essentially the same time it takes to run a traditional clustering algorithm.

## 6.2. Clustering Quality

Our runtime efficiency enables exploring diverse clustering approaches, which is particularly valuable given the results shown in Table 3. Here, we study the quality of  $k$ -center/stability (which is HDBSCAN if done on the dc-dist),  $k$ -median/MoE, and  $k$ -means/elbow combinations on the dc-dist and Cover trees and compare them against the competitors from Table 2. The ARI scores reveal that the SHIP clustering framework on the dc-dist produces clusterings that outperform competitor algorithms in most cases, even without knowing the ground truth number of clusters. Importantly,  $k$ -means, Ward, and SCAR are all given the ground-truth value of  $k$  to maximize their competitiveness. Furthermore, we point out that Ward agglomerative clustering is ultrametric in nature and, therefore, falls under the umbrella of our framework (Milligan, 1979). We also see that Cover tree combinations achieve slightly worse performance than Euclidean  $k$ -means.

Notably, no single hierarchy/partition combination emerges as universally superior, with different combinations excelling on different datasets. This underscores the value of rapidly switching between methods and allowing users to identify the best clustering strategy for their data.

## 7. Conclusions and Future Work

This paper presented a set of fast techniques for hierarchical clustering based on a novel derivation of optimal center-based clustering in ultrametrics. We showed that, after fitting an ultrametric to the data, one has access to a multitude of clustering hierarchies simultaneously. We note that our results also highlight the efficacy of the dc-dist for uncovering a dataset's structure in an unsupervised way: almost any clustering technique seems to be effective over this ultrametric. In future work, we aim to connect our results to hierarchies based on specific hierarchical objective functions (Dasgupta, 2016; Cohen-addad et al., 2019). Similarly, while ultrametrics and spectral clustering have been connected (Little et al., 2020; Beer et al., 2023), their relation to our framework is left for future work.

## Acknowledgements

We thank Chris Schwiegelshohn, who provided advice and insights when developing the theoretical framework surrounding Theorem B.3.

We gratefully acknowledge financial support from the Vienna Science and Technology Fund (WWTF-ICT19-041) and the Austrian funding agency for business-oriented research, development, and innovation (FFG-903641<sup>5</sup>). Andrew Draganov is partially supported by the Independent Research Fund Denmark (DFF) under a Sapere Aude Research Leader grant No 1051-00106B. We also acknowledge funding from Danish Pioneer Centre for AI (aicentre.dk), DNRF grant number P1.

## References

- Bartal, Y. On approximating arbitrary metrics by tree metrics. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pp. 161–168, 1998.
- Bartal, Y., Linial, N., Mendel, M., and Naor, A. On metric ramsey-type phenomena. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pp. 463–472, 2003.
- Barton, T. Clustering benchmark. <https://github.com/deric/clustering-benchmark>, 2019.
- Beer, A., Draganov, A., Hohma, E., Jahn, P., Frey, C. M., and Assent, I. Connecting the dots – density-connectivity distance unifies DBSCAN,  $k$ -center and Spectral Clustering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 80–92, 2023.
- Bentley, J. L. K-d trees for semidynamic point sets. In *Proceedings of the sixth annual symposium on Computational geometry*, pp. 187–197, 1990.
- Beygelzimer, A., Kakade, S., and Langford, J. Cover trees for nearest neighbor. In *Proceedings of the 23rd international conference on Machine learning*, pp. 97–104, 2006.
- Broad Institute. Single cell portal, 2025. URL [https://singlecell.broadinstitute.org/single\\_cell/study/](https://singlecell.broadinstitute.org/single_cell/study/). Accessed: 2025-01-29.
- Campello, R. J. G. B., Moulavi, D., and Sander, J. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pp. 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37456-2. doi: [https://doi.org/10.1007/978-3-642-37456-2\\_14](https://doi.org/10.1007/978-3-642-37456-2_14).
- Campello, R. J. G. B., Moulavi, D., Zimek, A., and Sander, J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1):1–51, July 2015. ISSN 1556-4681, 1556-472X. doi: 10.1145/2733381. URL <https://dl.acm.org/doi/10.1145/2733381>.
- Carlsson, G. E., Mémoli, F., et al. Characterization, stability and convergence of hierarchical clustering methods. *J. Mach. Learn. Res.*, 11(Apr):1425–1470, 2010.
- Chierchia, G. and Perret, B. Ultrametric fitting by gradient descent\*. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124004, dec 2020. doi: 10.1088/1742-5468/abc62d. URL <https://dx.doi.org/10.1088/1742-5468/abc62d>.
- Cohen-addad, V., Kanade, V., Mallmann-trenn, F., and Mathieu, C. Hierarchical clustering: Objective functions and algorithms. *J. ACM*, 66(4), jun 2019. ISSN 0004-5411. doi: 10.1145/3321386. URL <https://doi.org/10.1145/3321386>.
- Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., and Mathieu, C. Hierarchical clustering: Objective functions and algorithms. *Journal of the ACM (JACM)*, 66(4):1–42, 2019.
- Cohen-Addad, V., Lattanzi, S., Norouzi-Fard, A., Sohler, C., and Svensson, O. Fast and accurate  $k$ -means++ via rejection sampling. *Advances in Neural Information Processing Systems*, 33:16235–16245, 2020.
- Cohen-Addad, V., Lattanzi, S., Norouzi-Fard, A., Sohler, C., and Svensson, O. Parallel and efficient hierarchical  $k$ -median clustering. *Advances in Neural Information Processing Systems*, 34:20333–20345, 2021.

<sup>5</sup><https://projekte.ffg.at/projekt/4814676>

- Curtin, R. R., Edel, M., Shrit, O., Agrawal, S., Basak, S., Balamuta, J. J., Birmingham, R., Dutt, K., Eddebuettel, D., Garg, R., Jaiswal, S., Kaushik, A., Kim, S., Mukherjee, A., Sai, N. G., Sharma, N., Parihar, Y. S., Swain, R., and Sanderson, C. mlpack 4: a fast, header-only c++ machine learning library. *Journal of Open Source Software*, 8(82):5026, 2023. doi: 10.21105/joss.05026. URL <https://doi.org/10.21105/joss.05026>.
- Dasgupta, S. The hardness of k-means clustering. 2008.
- Dasgupta, S. A cost function for similarity-based hierarchical clustering. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 118–127, 2016.
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pp. 226–231, 1996.
- Estévez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. Normalized mutual information feature selection. *IEEE Transactions on neural networks*, 20(2):189–201, 2009.
- Fakcharoenphol, J., Rao, S., and Talwar, K. A tight bound on approximating arbitrary metrics by tree metrics. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pp. 448–455, 2003.
- Fischer, B., Roth, V., and Buhmann, J. Clustering with the connectivity kernel. *Advances in neural information processing systems*, 16, 2003.
- Foley, T. and Sugerman, J. Kd-tree acceleration structures for a gpu raytracer. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware*, pp. 15–22, 2005.
- Gray, A. and Moore, A. N-body' problems in statistical learning. *Advances in neural information processing systems*, 13, 2000.
- Han, Y. Deterministic sorting in  $O(n \log \log n)$  time and linear space. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pp. 602–608, 2002.
- Har-peled, S. *Geometric Approximation Algorithms*. American Mathematical Society, USA, 2011. ISBN 0821849115.
- Hochbaum, D. S. and Shmoys, D. B. A best possible heuristic for the k-center problem. *Mathematics of operations research*, 10(2):180–184, 1985.
- Hohma, E., Frey, C. M., Beer, A., and Seidl, T. Scar: spectral clustering accelerated and robustified. *Proceedings of the VLDB Endowment*, 15(11):3031–3044, 2022.
- Hoseini, F. and Haghir Chehreghani, M. Memory-efficient minimax distance measures. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 419–431. Springer, 2022.
- Hubert, L. and Arabie, P. Comparing partitions. *Journal of classification*, 2:193–218, 1985. doi: <https://doi.org/10.1007/BF01908075>.
- Hull, J. J. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.
- Indyk, P. Algorithms for dynamic geometric problems over data streams. In *Proceedings of the thirty-sixth annual ACM Symposium on Theory of Computing*, pp. 373–380, 2004.
- Jahn, P., Frey, C. M., Beer, A., Leiber, C., and Seidl, T. Data with density-based clusters: A generator for systematic evaluation of clustering algorithms. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 3–21. Springer, 2024.
- Jarman, A. M. Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. *Georgia Southern University*, 29, 2020.
- Johnson, S. C. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.
- Lang, A. and Schubert, E. Accelerating k-means clustering with cover trees. In *International Conference on Similarity Search and Applications*, pp. 148–162. Springer, 2023.
- Leclerc, B. Description combinatoire des ultramétriques. *Mathématiques et Sciences humaines*, 73:5–37, 1981.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Lim, A., Rodrigues, B., Wang, F., and Xu, Z. k-center problems with minimum coverage. *Theoretical Computer Science*, 332(1-3):1–17, 2005. doi: <https://doi.org/10.1016/j.tcs.2004.08.010>.
- Lin, G., Nagarajan, C., Rajaraman, R., and Williamson, D. P. A general approach for incremental approximation and hierarchical clustering. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pp. 1147–1156, 2006.

- Little, A. V., Maggioni, M., and Murphy, J. M. Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms. *Journal of machine learning research*, 21, 2020.
- March, W. B., Ram, P., and Gray, A. G. Fast euclidean minimum spanning tree: algorithm, analysis, and applications. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 603–612, 2010.
- Markelle Kelly, Rachel Longjohn, K. N. The uci machine learning repository, 2023. URL <http://archive.ics.uci.edu/ml>.
- McInnes, L. and Healy, J. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 33–42, 2017. doi: 10.1109/ICDMW.2017.12.
- Mettu, R. R. and Plaxton, C. G. The online median problem. *SIAM Journal on Computing*, 32(3):816–832, 2003.
- Milligan, G. W. Ultrametric hierarchical clustering algorithms. *Psychometrika*, 44(3):343–346, 1979.
- Mirzaei, A. and Rahmati, M. A novel hierarchical-clustering-combination scheme based on fuzzy-similarity relations. *IEEE Transactions on Fuzzy Systems*, 18(1): 27–39, 2009.
- Nene, S. A., Nayar, S. K., and Murase, H. Columbia object image library (coil-100). 1996a. URL <https://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php>.
- Nene, S. A., Nayar, S. K., and Murase, H. Columbia object image library (coil-20). 1996b. URL <https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php>.
- Page, R. D. and Holmes, E. C. *Molecular evolution: a phylogenetic approach*. John Wiley & Sons, 2009.
- Ram, P. and Sinha, K. Revisiting kd-tree for nearest neighbor search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pp. 1378–1388, 2019.
- Rodriguez, A. and Laio, A. Clustering by fast search and find of density peaks. *science*, 344(6191):1492–1496, 2014. doi: 10.1126/science.1242072.
- Schubert, E. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *ACM SIGKDD Explorations Newsletter*, 25(1):36–42, 2023.
- Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., and Liu, J. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP journal on wireless communications and networking*, 2021:1–16, 2021.
- Simhadri, H. V., Williams, G., Aumüller, M., Douze, M., Babenko, A., Baranchuk, D., Chen, Q., Hosseini, L., Krishnaswamy, R., Srinivasa, G., et al. Results of the neurips'21 challenge on billion-scale approximate nearest neighbor search. In *NeurIPS 2021 Competitions and Demonstrations Track*, pp. 177–189. PMLR, 2022.
- Thorndike, R. L. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.
- Van Der Maaten, L. Accelerating t-sne using tree-based algorithms. *The journal of machine learning research*, 15 (1):3221–3245, 2014.
- Wang, Z., qin Ye, Z., Du, Y., Mao, Y., Liu, Y., Wu, Z., and Wang, J. AMD-DBSCAN: An adaptive multi-density DBSCAN for datasets of extremely variable density. *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10, 2022. URL <https://api.semanticscholar.org/CorpusID:252918697>.
- Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.
- Yoon, J. Clustering exercises dataset, 2023. URL <https://www.kaggle.com/datasets/joonasyoon/clustering-exercises>. Accessed: 2025-01-29.
- Zeng, Y., Tong, Y., and Chen, L. HST+: An efficient index for embedding arbitrary metric spaces. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 648–659. IEEE, 2021.

## A. Proofs for Section 2 - Ultrametrics and Tree Representations

In this section, we prove Theorem 2.3, Corollary 2.4, and Theorem 3.2 from the main body of the paper. We restate each here:

**Theorem 2.3.** *Let  $(L, d')$  be a finite relaxed ultrametric space. Then there exists LCA-tree  $T$  with LCA-distance  $d$  and a bijection  $f : L \leftrightarrow \text{leaves}(T)$  such that, for all  $\ell_i, \ell_j \in L$ ,  $d'(\ell_i, \ell_j) = d(f(\ell_i) \vee f(\ell_j))$ .*

**Corollary 2.4.** *Let  $T$  be an LCA-tree. For any leaf  $\ell \in T$ , let  $p(\ell) = [\ell, \eta_a, \dots, \eta_b, r(T)]$  be the path from  $\ell$  to the root of the tree  $r(T)$ . Then the LCA-distances on  $T$  form a relaxed ultrametric if and only if, for all  $\ell \in \text{leaves}(T)$  and  $\eta_i, \eta_j \in p(\ell)$ , the following conditions are satisfied:*

$$(1) \quad d(\ell) \geq 0 \quad \text{and} \quad (2) \quad \eta_i \preceq \eta_j \implies d(\eta_i) \leq d(\eta_j).$$

**Theorem 3.2.** *Let  $(L, d)$  be a finite relaxed ultrametric space represented over an LCA-tree  $T$ . Let  $n = |L|$  and  $z \in \mathbb{Z}_{>0}$ . Then, for both the  $k$ -center and  $(k, z)$ -clustering objectives on  $T$ , there exists an algorithm which finds the optimal solutions  $\{\mathbf{C}_1, \dots, \mathbf{C}_n\}$  for all  $k \in \mathbb{N}_n$  in  $\text{Sort}(n)$  time. Furthermore, the respective partitions  $\mathcal{H} = \{\mathcal{P}_1, \dots, \mathcal{P}_n\}$  obtained by assigning all leaves in  $T$  to their closest center satisfy Definition 3.1.*

**More detailed proof outline.** Before diving in, we give a more thorough proof outline to the one that appears in the main body of the paper.

First, we will define a relaxation of ultrametrics and show a few immediate properties of these spaces. Their key property is that there is essentially an equivalence relation induced by any relaxed ultrametric: for any distance value  $d$ , the sets of points that are within distance  $d$  of each other partition the space. This is a generalization of the ideas in (Carlsson et al., 2010).

We then show that all relaxed ultrametrics can be represented as lowest-common-ancestor-trees (LCA-trees). These are rooted trees where every node has a value associated with it. The distance between two leaves in an LCA-tree is the value in their lowest common ancestor. Furthermore, all LCA-trees are (relaxed) ultrametrics if the values are monotonically non-decreasing along the path from any leaf to the root (Corollary 2.4). As a result, there is a bijective relationship between LCA-trees and ultrametrics (Theorem 2.3). We proceed by only considering ultrametrics in the LCA-tree data structure.

We now consider the center-based clustering objectives over these LCA-trees. For  $k$ -center, we will see that the well-known strategy of farthest-first traversal (Har-peled, 2011) (which achieves a 2-approximation in Euclidean space) is actually optimal in ultrametrics. The intuition here is that the farthest-first traversal's standard 2-approximation is a direct consequence of the triangle inequality. Thus, the ultrametric's strong triangle inequality resolves the approximation error. The runtime being  $\text{Sort}(n)$  comes from the fact that the ultrametric allows us to sort these distances quickly. As a result, we will solve  $k$ -center in an ultrametric by sorting the distances from largest to smallest and placing the corresponding centers.

Given this, we will conclude the proof by showing how to reduce the general problems of  $k$ -means and  $k$ -median clustering to this  $k$ -center algorithm. We will consider these through the lens of the *cost-decreases* of placing  $k$ -means or  $k$ -median centers in an ultrametric. That is, if we have an optimal solution for some value of  $k$ , then the optimal solution for  $k + 1$  centers will have a lower cost. Thus, there is a cost-decrease associated with each center. Our key observation is that these cost-decreases *themselves* satisfy the strong triangle inequality. We will also show that greedily choosing centers that maximize these cost-decreases gives optimal  $k$ -means and  $k$ -median solutions in an ultrametric. Putting the pieces together, these results imply that we can simply apply the  $k$ -center algorithm in the LCA-tree of cost-decreases to get optimal  $k$ -means and  $k$ -median solutions.

**Notation.** Throughout this section, we use  $T$  to represent an arbitrary rooted tree with root  $r$ . We use  $\eta$  to represent arbitrary internal nodes and  $\ell$  to represent arbitrary leaves. We also assume that every internal node  $\eta$  in the tree is equipped with a value, which we write by  $d(\eta)$ . We also use the notation  $\eta_i \preceq \eta_j$  to indicate that  $\eta_j$  lies on the path from  $\eta_i$  to  $r$ . We use  $\text{children}(\eta)$  and  $\text{parent}(\eta)$  to indicate the direct children and parent of a node. Lastly, we use the notation from (Dasgupta, 2016) with  $\ell_i \vee \ell_j$  denoting the lowest common ancestor (LCA) of a set of nodes/leaves and  $T[\eta]$  denoting the subtree rooted at  $\eta$ . Thus,  $T[\ell_i \vee \ell_j]$  is the smallest subtree containing both  $\ell_i$  and  $\ell_j$ .

For notation on clustering, we define  $(k, z)$ -clustering as finding the set of centers  $\mathbf{C} \in T$  with  $|\mathbf{C}| = k$  which minimize

$$\text{Cost}_z(T, \mathbf{C}) = \sum_{\ell \in \text{leaves}(T)} \min_{c \in \mathbf{C}} d(\ell, c)^z.$$

This corresponds to  $k$ -median and  $k$ -means for  $z = 1$  and  $z = 2$ , respectively. We also define  $k$ -center clustering as finding the  $\mathbf{C}$  which minimizes

$$\text{Cost}_\infty(T, \mathbf{C}) = \max_{\ell \in \text{leaves}(T)} \min_{c \in \mathbf{C}} d(\ell, c).$$

### A.1. Ultrametrics and LCA-Trees

Throughout the literature, the stand-out candidate for a hierarchical (dis-)similarity measure is the *ultrametric*. We will, however require a looser definition, which we refer to as a *relaxed ultrametric*:

**Definition 2.1.** Let  $L$  be a set. Then  $d : L \times L \rightarrow \mathbb{R}_{\geq 0}$  is a *relaxed ultrametric* over  $L$  if, for all  $\ell_i, \ell_j, \ell_k \in L$ , the following conditions are satisfied:

$$\begin{aligned} d(\ell_i, \ell_j) &= d(\ell_j, \ell_i) \geq 0 \\ d(\ell_i, \ell_k) &\leq \max(d(\ell_i, \ell_j), d(\ell_j, \ell_k)). \end{aligned}$$

We will prove our optimal center-based clustering results over these relaxed ultrametrics. We note that every ultrametric is a relaxed ultrametric with the additional condition that the distance between two points is 0 if and only if they are the same point. Thus, our theoretical results will immediately apply to all ultrametrics.

The *strong triangle inequality* in Definition 2.1 is what allows ultrametrics to capture hierarchical relationships. Specifically, the strong triangle inequality implies that any three points in an ultrametric space must form an isosceles triangle with an angle less than 60 degrees:

**Fact A.1.** Let  $d$  be a dissimilarity measure on a space  $L$ , which satisfies the strong triangle inequality. Then for any  $\ell_i, \ell_j, \ell_k \in L$ , one of the following holds:

1.  $d(\ell_i, \ell_j) \leq d(\ell_i, \ell_k) = d(\ell_j, \ell_k)$
2.  $d(\ell_i, \ell_k) \leq d(\ell_i, \ell_j) = d(\ell_j, \ell_k)$
3.  $d(\ell_j, \ell_k) \leq d(\ell_i, \ell_j) = d(\ell_i, \ell_k)$

*Proof.* We prove this by contradiction. First, assume that all three are unequal, so WLOG  $d(\ell_i, \ell_j) < d(\ell_i, \ell_k) < d(\ell_j, \ell_k)$ . Then the strong triangle inequality does not hold, since  $d(\ell_j, \ell_k) \not\leq \max(d(\ell_i, \ell_j), d(\ell_i, \ell_k))$ .

Similarly, assume that the singleton edge is longer than the two others, i.e.  $d(\ell_i, \ell_k) = d(\ell_j, \ell_k) < d(\ell_i, \ell_j)$ . This also breaks the strong triangle inequality, since  $d(\ell_i, \ell_j) \not\leq \max(d(\ell_i, \ell_j), d(\ell_i, \ell_k))$ .  $\square$

As a result, for any three points in a relaxed ultrametric space, knowing two of the pairwise distances is sufficient to give the ordering of all three. For example, if two of the distances in a triangle are equal, then the third must be equal to this distance or smaller.

This naturally extends to groups of more than 3 points. Consider the case where we have  $n$  points  $\{x_1, x_2, \dots, x_n\}$  so that for any  $x_i, x_j$ , we have ultrametric distance  $d(x_i, x_j) < 1$ . Now let  $y$  be a point with  $d(x_1, y) = 100$ . Since the  $x$ 's are all close to one another, Fact A.1 implies that  $d(x_i, y)$  must also equal 100 for all  $i$ . This has the following fundamental consequences:

**Theorem 2.3.** Let  $(L, d')$  be a finite relaxed ultrametric space. Then there exists LCA-tree  $T$  with LCA-distance  $d$  and a bijection  $f : L \leftrightarrow \text{leaves}(T)$  such that, for all  $\ell_i, \ell_j \in L$ ,  $d'(\ell_i, \ell_j) = d(f(\ell_i) \vee f(\ell_j))$ .

We first put this in context before giving its proof. In other words, Theorem 2.3 states that any ultrametric can be represented over the leaves of a tree, with the property that the distance between two leaves is uniquely determined by the value in their LCA. We note that variants of this theorem have been given elsewhere (Leclerc, 1981; Milligan, 1979; Mirzaei & Rahmati, 2009); nonetheless, the results later in this section require the form given above.

*Proof.* We use Fact A.1 to design an algorithm to construct the tree  $T$  by repeatedly splitting the relaxed ultrametric over its largest distance  $d_{max}$ . First, let us see that  $d_{max}$  induces a partition over  $L$ . Let  $\ell_i \in L$  be any point and let  $L_{\ell_i} = \{\ell_j : d(\ell_i, \ell_j) \leq d_{max}\} \cup \{\ell_i\}$  be the set consisting of  $\ell_i$  and all the points closer to it than  $d_{max}$ . Now let  $\ell_k \in L \setminus L_{\ell_i}$  be any other point not in  $L_{\ell_i}$ . By definition  $d(\ell_i, \ell_k) = d_{max}$ . Furthermore, by Fact A.1,  $d(\ell_j, \ell_k) = d_{max}$  for all  $\ell_j \in L_{\ell_i}$ . Thus,  $d_{max}$  induces a partition on a relaxed ultrametric where, across any two points in distinct clusters, the distance is necessarily  $d_{max}$ . We refer to  $L_{\ell_i}$  as a cluster with center  $\ell_i$ .

We now use this idea to devise an algorithm that embeds the ultrametric in an LCA-tree. Specifically, our algorithm receives a root node  $r$  and a relaxed ultrametric space  $(L, d)$  as input. Let  $d_{max}$  be the largest distance in  $(L, d)$  and let  $\mathcal{P}$  be the partition induced by  $d_{max}$ . Assign  $d(r) = d_{max}$ . For each cluster in the partition, make a node and assign it as a child of the root  $r$ .

We now apply this construction recursively for each of the children. The base case occurs when  $L$  has either one or two points. If there is only one point,  $x_i \in L$ , we simply return a leaf  $\ell_i$ . This leaf is given weight  $d(\ell_i) = d(\ell_i, \ell_i)$  and we define  $f(x_i) = \ell_i$ . If  $L$  has two points,  $x_i$  and  $x_j \in L$ , then we create two leaves  $\ell_i$  and  $\ell_j$  as children of the input node. The mapping is arbitrarily defined as  $f(x_i) = \ell_i$  and  $f(x_j) = \ell_j$ . We assign the input node with weight  $d(\eta) = d(x_i, x_j)$  and give the leaves weight 0, i.e.  $d(\ell_i) = d(\ell_j) = 0$ .

We verify the validity of this construction inductively. In the base case, the space  $(L, d)$  has either one or two points. We respectively represent these as a singleton node with a value of 0 or a rooted tree with two children, such that the value of the root is the distance between the two points. In both settings, all pairwise distances in  $L$  are preserved via the LCA values.

In the inductive step, assume that  $(L, d)$  has more than two points, that the maximal distance is  $d_{max}$ , and that all smaller distances are represented via distinct trees. By the above logic, the distance between any two nodes in separate trees must be  $d_{max}$ . Since the construction described above assigns the value  $d_{max}$  to the root and assigns the existing trees to it as children, this new node is a parent to the already-existing trees. Thus, their internal LCAs are not affected. However, the LCA between any nodes in separate subtrees has value  $d_{max}$ . Therefore, the entire ultrametric  $(L, d)$  is preserved.

□

We refer to this data structure as an *LCA-tree*. The following corollary gives the sufficient conditions for an LCA-tree to correspond to a relaxed ultrametric:

**Corollary 2.4.** *Let  $T$  be an LCA-tree. For any leaf  $\ell \in T$ , let  $p(\ell) = [\ell, \eta_a, \dots, \eta_b, r(T)]$  be the path from  $\ell$  to the root of the tree  $r(T)$ . Then the LCA-distances on  $T$  form a relaxed ultrametric if and only if, for all  $\ell \in \text{leaves}(T)$  and  $\eta_i, \eta_j \in p(\ell)$ , the following conditions are satisfied:*

$$(1) \quad d(\ell) \geq 0 \quad \text{and} \quad (2) \quad \eta_i \preceq \eta_j \implies d(\eta_i) \leq d(\eta_j).$$

*Proof.* First, note that the LCA-distances in the tree satisfy symmetry by the definition of LCA:  $d(\ell_i, \ell_j) = d(\text{LCA}(\ell_i, \ell_j)) = d(\ell_j, \ell_i)$ . Furthermore, conditions (1) and (2) together ensure the non-negativity conditions required for a relaxed ultrametric. Therefore, it remains to show the strong triangle inequality. Let  $\ell_i, \ell_j$  and  $\ell_k$  be three leaves in the LCA-tree. If they all have the same LCA (i.e.,  $\ell_i \vee \ell_j = \ell_j \vee \ell_k = \ell_i \vee \ell_k$ ), then the leaves are equidistant in the LCA-tree, and the strong triangle inequality is satisfied. Thus, assume WLOG that  $\ell_i \vee \ell_j \preceq \ell_i \vee \ell_k$ . This implies that  $\ell_i \vee \ell_k = \ell_j \vee \ell_k$ . This immediately implies that the strong triangle inequality is satisfied:

$$\ell_i \vee \ell_j \preceq \ell_i \vee \ell_k = \ell_j \vee \ell_k \quad \stackrel{\substack{\uparrow \\ \text{By Assumption}}}{\implies} \quad d(\ell_i \vee \ell_j) \leq d(\ell_i \vee \ell_k) = d(\ell_j \vee \ell_k).$$

□

As a result of Corollary 2.4, if we wish to show that a tree's LCA-distances satisfy the strong triangle inequality, we essentially need to show that the tree's values are non-decreasing on paths from the leaves to the root. Going forward, we will rely exclusively on this LCA-distance representation of relaxed ultrametrics: unless stated otherwise, every discussion of ultrametrics will implicitly be through their LCA-tree representation.

## B. Proofs for Section 3 - Center-based Clustering in Ultrametrics

### B.1. $k$ -Center in Ultrametrics

#### B.1.1. STRUCTURE OF A CENTER-BASED SOLUTION

Before delving into how to solve center-based clustering objectives optimally in LCA-trees, we must first describe how cluster memberships are defined. Recall that a cluster is the set of points that are closest to a center. Since many of the center-to-leaf relationships are equidistant in an LCA-tree, we use the “marking” procedure from (Cohen-Addad et al., 2021) to define a consistent notion of cluster attribution:

Let  $\mathbf{C} = [c_1, \dots, c_k]$  be  $k$  arbitrarily ordered centers that correspond to distinct leaves in the LCA-tree. We obtain the cluster memberships  $C_i = \{\ell \in T : c_i = \arg \min_{c \in \mathbf{C}} d(\ell, c)\}$  by adding the centers in the given order and, for each center placed, marking the nodes in the tree from the corresponding leaf to its lowest unmarked ancestor. Thus, if we place center  $c_i$  in a leaf node, we go up the tree and mark every node with “ $C_i$ ” until we hit a previously marked node. Leaves are then assigned to clusters by finding their lowest marked ancestor.

We will often discuss our clustering algorithms through the lens of an LCA-tree’s “most expensive unmarked node”. This represents the node that has the largest value among the LCA-tree’s unmarked nodes. A key insight is that, in a  $k$ -center solution on an LCA-tree, this “most expensive unmarked node” precisely corresponds to the cost of the solution. That is, if  $\eta$  is the most expensive unmarked node for solution  $\mathbf{C}$  on LCA-tree  $T$ , then  $\text{Cost}_\infty(T, \mathbf{C}) = d(\text{parent}(\eta))$ .

Notice also that the number of optimal  $k$ -clusterings in any LCA-tree is exponential in  $k$ . To see this, consider an LCA-tree where there are several leaves below an unmarked node whose parent is marked. Regardless of which leaf the center is placed on, the distances (and therefore the costs) to the rest of the tree are equivalent. Thus, if an optimal clustering had a center on one of these leaves, we could replace it with a center on any of the other leaves without any change to the solution’s optimality. Thus, when describing optimal solutions, we consider them equivalent up to such a permutation. We discuss heuristics for choosing the “best” of the optimal solutions in Appendix D.1.2.

#### B.1.2. $k$ -CENTER IN LCA-TREES

We now move to the properties of center-based clustering objectives in LCA-trees, starting with the  $k$ -center clustering task. Recall that the  $k$ -center objective requires finding  $k$  centers that minimize

$$\text{Cost}_\infty(T, \mathbf{C}) = \max_{\ell \in T} \min_{c \in \mathbf{C}} \text{Dist}(\ell, c).$$

Although the  $k$ -center task is NP-hard in the general metric setting, we will see that it can be solved optimally (and almost trivially) in an LCA-tree. To describe this more formally, however, we must define *hierarchical* clusters:

**Definition 3.1.** [Lin et al. (2006)] A *cluster hierarchy*  $\mathcal{H} = \{\mathcal{P}_1, \dots, \mathcal{P}_n\}$  is a set of partitions with

1. for  $k = 1$ ,  $\mathcal{P}_1 \subseteq L$ , and
2. for  $1 < k \leq n$ ,  $\mathcal{P}_k = (\mathcal{P}_{k-1} \setminus C_i) \cup \{C_j, C_l\}$ , such that  $C_i = C_j \cup C_l \in \mathcal{P}_{k-1}$  with  $C_j \cap C_l = \emptyset$  and  $i \neq j \neq l$ .

We say a cluster  $C_i \in \mathcal{H}$  if  $\exists \mathcal{P}_j \in \mathcal{H}$  with  $C_i \in \mathcal{P}_j$ .

On an intuitive level, a hierarchical set of solutions means that the clustering in  $\mathcal{P}_k$  is the same as the one at  $\mathcal{P}_{k-1}$  except that a single cluster was split apart.

**Farthest-First Traversal.** We will solve  $k$ -center in LCA-trees using the farthest-first traversal algorithm (Hochbaum & Shmoys, 1985; Har-peled, 2011). The naive algorithm works by assigning the first center to a random leaf in the tree. For each subsequent center, we choose it from the subtree that has the highest distance to the current centers. Although this algorithm provides a 2-approximation in standard  $k$ -center (Har-peled, 2011), it turns out that the change from the triangle inequality to the strong triangle inequality makes this method optimal.

Unfortunately, the naive farthest-first algorithm may take  $O(n^2)$  time to obtain all clusterings for  $k \in \{1, \dots, n\}$  since, when placing center  $c_i$ , we may have to search through  $O(n)$  nodes to find the one with the next-highest cost. However, we can improve this to  $\text{Sort}(n)$  time – the time it takes to sort a list of the  $O(n)$  values in the LCA-tree – by noting that the nodes’ values grow as we go up the tree. Thus, it is sufficient to sort the internal nodes by these costs and then place their corresponding centers in that order. The following lemma formalizes this intuition:

**Lemma B.1.** Let  $T$  be an LCA-tree satisfying the conditions in Corollary 2.4. Then there is an algorithm that runs in  $\text{Sort}(n)$  time and finds all the optimal  $k$ -center solutions on  $T$  for  $k \in \{1, \dots, n\}$ . Furthermore, these optimal solutions are hierarchical.

*Proof.* We will first show that the greedy farthest-first traversal is optimal in ultrametrics. After this, we will see a simple algorithm for accomplishing it in  $\text{Sort}(n)$  time. Lastly, we will prove that the optimal  $k$ -center solutions are hierarchical.

We first show that farthest-first traversal is optimal for every choice of  $k$ . Assume for contradiction that the greedy  $k$ -center solution  $\mathbf{C}_g$  is not optimal. Then there must be another clustering  $\mathbf{C}_o$  of  $k$  centers that is *actually* optimal, i.e.  $\text{Cost}_\infty(T, \mathbf{C}_o) < \text{Cost}_\infty(T, \mathbf{C}_g)$ . These two solutions must differ by at least one unmarked node. Of those nodes that are unmarked in  $\mathbf{C}_o$  but marked in  $\mathbf{C}_g$ , let  $\eta$  be the one with the largest value (with ties broken arbitrarily). This means that  $\text{Cost}_\infty(T, \mathbf{C}_o) = d(\text{parent}(\eta))$ . However,  $\mathbf{C}_g$  has marked  $\eta$  and every other node with larger value. Therefore, we must have  $\text{Cost}_\infty(T, \mathbf{C}_g) \leq d(\text{parent}(\eta))$ . This gives the desired contradiction. Interestingly, this correctness proof does not depend on which leaf gets chosen as the center in a subtree – just that one leaf is chosen.

We now show that the farthest-first traversal can be executed in LCA-trees in  $\text{Sort}(n)$  time. The main idea is as follows: for every internal node, we must assign it one leaf as its *corresponding center*. We then sort the internal nodes by their values and place these corresponding centers one at a time. If a set of nodes has equal values, then all of their centers must be placed before the cost can decrease. Thus, ties can be broken arbitrarily.

Algorithm 2 does precisely this. It uses Algorithm 1 as a subroutine to find the corresponding centers for the internal nodes. This is done by depth-first-search: at any given node  $\eta$ , Algorithm 2 assigns  $\eta$ 's corresponding center as the corresponding center of its first child. For  $\eta$ 's remaining children that were not chosen, we store their value in a global dictionary. Algorithm 1 finally returns the dictionary of nodes in the tree and their values. Algorithm 2 concludes by sorting these nodes by their values from largest to smallest and placing the corresponding centers in this order.

The bottleneck of Algorithm 2 is the sorting, which occurs in  $\text{Sort}(n) > O(n)$  time. The other steps run in  $O(n)$  time. Lastly, placing centers in this way must be hierarchical. Every placed center corresponds to an unmarked node  $\eta$  in the tree. Since every leaf in  $T[\eta]$  belonged to the same cluster, placing a new center only splits one cluster at a time.

□

---

**Algorithm 1** CorrespondingCenters

**Input:** node  $\eta$  in an LCA-tree; dict Costs mapping nodes to distances;

```

1: if  $\eta$  is leaf then
2:    $c(\eta) = \eta$ 
3:   Return
4: end if

5: ChildCount = 0
6: for  $\eta' \in \text{children}(\eta)$  do
7:   CorrespondingCenters( $\eta'$ , Costs)
8:   if ChildCount = 0 then
9:      $c(\eta) = c(\eta')$ 
10:   else
11:     Costs{ $\eta'$ } =  $d(\eta)$ 
12:   end if
13:   ChildCount += 1
14: end for
15: Return
```

---

We take a moment to provide context for why the LCA-tree and the  $k$ -center hierarchy are essentially isomorphic. This occurs essentially by combining Lemma B.1 and Corollary 2.4.

Let  $k$  be any integer between 1 and  $n - 1$ , and let  $\mathcal{P}_k$  be the partition corresponding to the optimal clustering for this value of  $k$ . By Lemma B.1, our next center will be placed in the subtree rooted at the unmarked node with the largest value. Let

---

**Algorithm 2** Ultrametric-kCenter

**Input:** LCA-tree  $T$ 

```

1: Costs = { }
2: CorrespondingCenters( $T.\text{root}$ , Costs) // assume pass-by-reference on Costs
3: Costs = OrderedDict(Costs) // sorted from largest to smallest
4: for  $\eta \in \text{Costs}$  do
5:   Place center at  $c(\eta)$ 
6: end for

```

---

this node be  $\eta$ . By Corollary 2.4, this unmarked node's parent must be marked (otherwise, there would be an unmarked node with a larger value than  $d(\eta)$ ). Therefore, there exists a center in  $T[\text{parent}(\eta)]$  but not one in  $T[\eta]$ , implying that our optimal solution had a cluster  $C_i = \text{leaves}(\text{parent}(\eta))$ .

By placing the center in  $T[\eta]$ , we split  $C_i$  into two clusters:  $C_j = \text{leaves}(T[\eta])$  and  $C_l = C_i \setminus C_j$ . Thus, the  $k$ -center hierarchy *directly* follows the hierarchy in the LCA-tree: for every node in  $T$ , there exists a cluster in the optimal  $k$ -center hierarchy.

We also note that our runtime is tight:

**Lemma B.2.** *Let  $T$  be an LCA-tree satisfying the conditions in Corollary 2.4. Then there is a worst-case instance on which one cannot find all the optimal  $k$ -center solutions on  $T$  for  $k \in \{1, \dots, n\}$  in faster than  $\text{Sort}(n)$  time.*

*Proof.* Consider a rooted tree that is complete and perfectly balanced: every leaf is at the same depth, and every internal node has two children. Let the leaves all be at depth  $w$ , so that there are  $2^w$  leaves. Starting at depth  $w$ , assign the leaves' unique values from 1 to  $2^w$ . Then, for the nodes at depth  $w - 1$ , assign them unique values from  $2^w + 1$  to  $2^w + 2^{w-1}$ . Continue this process until we reach the root node, to which we assign value  $2^{w+1}$ . Essentially, we go through the tree's nodes one-by-one from the lowest level to the root and maintain a counter of the number of visited nodes. Each node is assigned the value of the counter when it is visited.

Labeling the nodes by these values necessarily gives us an LCA-tree: all values are non-negative, and values are non-decreasing along paths from the leaves to the root. Furthermore, all internal nodes at depth  $i$  have distinct values. Suppose we are now performing  $k$ -center and have placed centers for all the nodes at depth  $w - 1$  but have not yet for the nodes at depth  $w$ . There are, therefore,  $O(n)$  available nodes on which to place centers. Of these, only one has the maximum leaf-to-center distance and therefore induces the cost. Thus, to place the remaining  $O(n)$  centers, we would require sorting the remaining leaves by their costs.  $\square$

## B.2. $(k, z)$ -clustering in LCA-trees

The result for  $k$ -center essentially boils down to the optimality of the classic 2-approximation when applied in an ultrametric. We now turn to the more interesting result: that the optimal  $(k, z)$ -clustering solutions behave very similarly to the optimal  $k$ -center ones. This may be surprising given that the former's cost depends on the number of points in a cluster while the latter's does not. Nonetheless, in this section, we will see that the optimal solutions to the  $(k, z)$ -clustering problem are hierarchical and can all be found in  $\text{Sort}(n)$  time using the  $k$ -center algorithm as a subroutine. Recall that the  $(k, z)$ -clustering objective has the cost function

$$\text{Cost}_z(T, \mathbf{C}) = \sum_{\ell \in T} \min_{c \in \mathbf{C}} d(\ell, c)^z,$$

and corresponds to  $k$ -median and  $k$ -means for  $z = 1$  and  $z = 2$ , respectively. We now present this section's primary result, which is a complete analog of the  $k$ -center one from Lemma B.1:

**Theorem B.3.** *Let  $T$  be an LCA-tree satisfying the conditions in Corollary 2.4 and let  $z$  be any positive integer. Then there is an algorithm that runs in  $\text{Sort}(n)$  time and finds the optimal  $(k, z)$ -clustering solutions on  $T$  for all  $k \in \{1, \dots, n\}$ . Furthermore, these solutions are hierarchical.*

**Optimal  $(k, z)$  Centers in Subtrees.** To gain some preliminary insight into this, let us consider what happens when we place a  $(k, z)$ -clustering center in an LCA-tree. Placing this center will create a trail of markings from the leaf up until the

first marked node along the path to the root. Then, our claim is that this center is also optimal everywhere along this trail. Specifically, we show this via the following lemma, which states that the first center we place in an LCA-tree must be optimal everywhere along the path from the center's leaf to the root. We will then immediately extend this to the  $k$ -th center being placed.

**Lemma B.4.** *Let  $c_z = OPT_{1,z}(T)$  be an optimal  $(1, z)$ -clustering solution for LCA-tree  $T$ . Then for every subtree  $T' \subset T$  such that  $c_z \in T'$ ,  $c_z$  is an optimal  $(1, z)$ -clustering solution for  $T'$ .*

*Proof.* We show this inductively. The base case is the trivial LCA-tree of one node where, inherently, the only choice of the center is optimal, and there are no subtrees. For the inductive case, consider LCA-tree  $T$  whose root has  $k$  children, such that each child has an optimal center placed within it. Our goal is to show that if we were to have one center for all of  $T$ , the optimum would be one of the  $k$  centers in its subtrees. We, therefore, want to find the center that gives  $\text{cost}_1(T)$  – the cost of optimally placing 1 center in  $T$ .

If we only had one center to place for all of  $T$ , that center must be in one of its subtrees. WLOG, let the optimal center for  $T$  be in the subtree  $T_1$ . Thus, our cost for one center is necessarily of the form  $\text{cost}_1(T) = \text{cost}_1(T_1) + \sum_{i=2}^k |T_i| \cdot d(\text{root}(T))^z$ , where  $|T_i|$  is the number of leaves in the  $i$ -th subtree of  $T$ . By the inductive hypothesis, we already had an optimal center for subtree  $T_1$ , implying that  $\text{cost}_1(T_1)$  is minimized by choosing the optimal center in  $T_1$ . The other term  $\sum_{i=2}^k |T_i| \cdot d(\text{root}(T))^z$  does not depend on where in  $T_1$  the center is placed. Therefore, the optimal center from  $T_1$  remains optimal for  $T$ . □

The key thing to note about Lemma B.4 is that it applies to any LCA-tree. Now consider an LCA-tree  $T$  with some centers, and suppose we place a center in unmarked subtree  $T' \subset T$ . Then Lemma B.4 holds for  $T'$ . Consequently, *every internal node in the LCA-tree has an optimal center (leaf) associated with it*. Furthermore, these centers are optimal along the path from their leaves to their corresponding internal nodes.

This notion proves essential enough that we give it its own definition:

**Definition B.5.** Let  $T$  be an LCA-tree satisfying the conditions in Corollary 2.4, let  $\eta$  be a node in  $T$ , and let  $z$  be a positive integer. Then  $\eta$ 's *corresponding  $z$ -center* is  $c_z(\eta) = OPT_{1,z}(T[\eta])$ .

In essence, Lemma B.4 is the key property of clustering in ultrametric spaces that makes the entire proof go through. To illustrate its effectiveness, consider the Euclidean 1-means setting on a dataset of two clearly separated Gaussian clusters. The Euclidean mean naturally falls *between* the two clusters. What Lemma B.4 says is that, in the ultrametric setting, the optimal 1-means solution for a dataset of two well-separated clusters is *itself* located in one of the clusters. Furthermore, it is optimal for the cluster in which it is located. In practice, this means that after placing an optimal center, we can essentially forget about it – it is guaranteed to be optimal for any set of points it serves.

**Overview of Proof for Theorem B.3.** We now give a simple blueprint illustrating how we use this for fast, optimal  $(k, z)$ -clustering. Assume we have placed the first center and have left a set of nodes unmarked. For each such unmarked node, we can determine its optimal center as well as how much the subtree's  $(k, z)$ -clustering cost would decrease by placing this center. Curiously, an application of Lemma B.4 shows that these cost-decreases themselves satisfy the strong triangle inequality. Another application of Lemma B.4 then allows us to show that greedily choosing the maximum cost-decrease gives an optimal  $(k, z)$ -clustering solution in an LCA-tree. As a result, we can apply the  $k$ -center algorithm to the LCA-tree of cost decreases. This section is devoted to verifying the speed and optimality of the above blueprint.

**Notation.** We will also need some additional notation to simplify the presentation. Let us define the cost of a node as

$$\text{NodeCost}(\eta, z) = |T[\eta]| \cdot d(\text{parent}(\eta))^z,$$

where  $|T[\eta]|$  is the number of leaves in the subtree rooted at  $\eta$ . This represents the cost contributed by  $\eta$ 's leaves when  $\eta$  is unmarked, but its parent is marked. In essence, this is the cost of  $T[\eta]$  in a  $(k, z)$ -clustering solution if there is no center in  $T[\eta]$ . Similarly, we define the cost *decrease* at  $\eta$  as

$$\text{CostDecrease}(\eta, z) = \text{NodeCost}(\eta, z) - \text{Cost}(T[\eta], c_z(\eta)),$$

where  $\text{Cost}(T[\eta], c_z(\eta)) = \sum_{\ell \in \text{leaves}(T[\eta])} d(\ell, c_z(\eta))^z$ . Here,  $c_z(\eta)$  is  $\eta$ 's corresponding  $z$ -center. We define the cost-decrease of the root node to be infinite.

In essence, the cost-decrease quantifies how much placing an optimal center in  $T[\eta]$  would decrease the subtree's total cost. Importantly, the cost-decrease of a node  $\eta$  assumes that  $\text{parent}(\eta)$  – the node directly above  $\eta$  – is marked. We will see in Lemma B.8 that this is a reasonable assumption: every useful center we will place in the  $(k, z)$ -clustering setting will always have a marked parent.

---

**Algorithm 3** GetCostDecreases

**Input:** node  $\eta$  in an LCA-tree; dict Costs mapping nodes to their ; dict CostDecreases mapping nodes to their cost decrease;

```

1: if  $\eta$  is leaf then
2:    $c_z(\eta) = \eta$ 
3:   Costs $\{\eta\} = d(\eta)$ 
4:   Return
5: end if

6: if  $\eta$  is root then
7:   ParentDist =  $d(\eta) + 1$ 
8:   CostDecreases $\{\eta\} = \infty$ 
9: else
10:  ParentDist =  $d(\text{parent}(\eta))$ 
11: end if

12: SumOfCosts =  $\sum_{\eta' \in \text{children}(\eta)} |T[\eta']| \cdot d(\eta)$ 
13: CostIfChosen =  $\{\eta': 0 \text{ for } \eta' \in \text{children}(\eta)\}$ 
14: ChildCostDecreases =  $\{\eta': 0 \text{ for } \eta' \in \text{children}(\eta)\}$ 
15: for  $\eta' \in \text{children}(\eta)$  do
16:   GetCostDecreases( $\eta'$ , Costs, CostDecreases)
17:   CostIfChosen $\{\eta'\} = \text{SumOfCosts} - |T[\eta']| \cdot d(\eta) + \text{Costs}\{\eta'\}$ 
18:   ChildCostDecreases $\{\eta'\} = |T[\eta']| \cdot \text{ParentDist} - \text{CostIfChosen}\{\eta'\}$ 
19: end for

20: ChosenChild = arg max(ChildCostDecreases)
21:  $c_z(\eta) = c_z(\text{ChosenChild})$ 
22: Costs $\{\eta\} = \text{CostIfChosen}\{\text{ChosenChild}\}$ 
23: for  $\eta' \in \text{children}(\eta)$  such that  $\eta'$  is not ChosenChild do
24:   CostDecreases $\{\eta'\} = |T[\eta']| \cdot d(\eta)$ 
25: end for
26: Return
```

---

**Proving Theorem B.3.** We now proceed to the constituent lemmas, which will prove Theorem B.3. First, we see that all of the corresponding  $z$ -centers can be found in  $O(n)$  time on an LCA-tree:

**Lemma B.6.** *Let  $T$  be an LCA-tree satisfying the conditions in Corollary 2.4. Then there exists an algorithm which, for all  $\eta \in T$ , finds  $c_z(\eta)$  and  $\text{Cost}(T[\eta], c_z(\eta))$  in  $O(n)$  time. Furthermore, this algorithm stores the cost-decrease for all nodes  $\eta'$  for which  $c_z(\eta') \neq c_z(\text{parent}(\eta'))$ .*

*Proof.* This is accomplished by Algorithm 3, which is essentially a depth-first implementation of Lemma B.4's proof.

We prove by induction that Algorithm 3 finds the costs for all internal nodes. In the base case, our current node  $\eta$  is a leaf. Thus,  $\eta$ 's corresponding  $z$ -center is  $\eta$  and the cost of  $\eta$  to this center is simply  $d(\eta)$ .

We now essentially reuse the logic from Lemma B.4 to prove the inductive step. We begin the inductive step with a node  $\eta$  along with the costs and corresponding  $z$ -centers of each of  $\eta$ 's children. That is, for all  $\eta' \in \text{children}(\eta)$ , we have access to both  $c_z(\eta)$  and  $\text{Cost}(T[\eta'], c_z(\eta'))$ . We now seek the optimal center for  $\eta$  and what the cost would be in  $T[\eta]$  after placing this center. By Lemma B.4, we know that the optimal center for  $\eta$  is one of its children's corresponding  $z$ -centers. I.e.,  $c_z(\eta)$

must be equal to  $c_z(\eta')$  for one of the children  $\eta'$ . We, therefore, test what the cost would be if we placed the center at each of the children and chose the minimum. By Lemma B.4, this center must be optimal. We therefore record the cost of placing this center in  $\eta$ , concluding the correctness proof.

Since this is done by depth-first search, the algorithm runs in  $O(n)$  time.  $\square$

Rather than thinking about corresponding  $z$ -centers as those which minimize the cost, we will instead think of them through the equivalent notion of the centers which *maximize* the cost-decrease. The next lemma shows the peculiar property that these cost-decreases themselves form a relaxed ultrametric:

**Lemma B.7.** *Let  $T$  be an ultrametric LCA-tree which satisfies the conditions in Corollary 2.4 and let  $T'$  be the LCA-tree obtained by replacing all values in  $T$  by the cost-decreases. I.e., for all  $\eta \in T'$ ,  $d(\eta) = \text{CostDecrease}(\eta, z)$ . Then, the LCA-distances over  $T'$  also satisfy the conditions in Corollary 2.4.*

*Proof.* By Corollary 2.4, showing that  $T'$  satisfies the strong triangle inequality simply requires verifying that the cost-decreases are non-negative and monotonically non-decreasing along any leaf-root path in  $T'$ . We first show that they are monotonically non-decreasing.

Let  $\eta$  be an unmarked node with  $h$  children whose parent is marked. We now place the optimal center  $c_z$  in  $T[\eta]$ . WLOG, this center must land in one of  $\eta$ 's children's subtrees. Call this child  $\eta_c$ , implying that  $c_z(\eta) = c_z(\eta_c)$ . We now show that the cost-decrease of  $\eta$  is greater than or equal to the cost decrease of  $\eta_c$ . After this, we will see that the cost decrease of  $\eta_c$  is, in turn, greater than the cost decrease of any of  $\eta$ 's other children.

First, notice that  $\text{CostDecrease}(\eta, z) \geq \text{CostDecrease}(\eta_c, z)$ . We show this by separating  $\text{CostDecrease}(\eta, z)$  into a sum of terms, of which  $\eta_c$  is a subset:

$$\begin{aligned}
 \text{CostDecrease}(\eta, z) &= \text{NodeCost}(\eta, z) - \text{Cost}(T[\eta], c_z(\eta)) \\
 &= \left( |T[\eta_c]| \cdot d(\text{parent}(\eta))^z + \sum_{\substack{\eta' \in \text{children}(\eta) \\ \eta' \neq \eta_c}} |T[\eta']| \cdot d(\text{parent}(\eta))^z \right) \\
 &\quad - \text{Cost}(T[\eta], c_z(\eta)) \\
 &= \left( |T[\eta_c]| \cdot d(\text{parent}(\eta))^z + \sum_{\substack{\eta' \in \text{children}(\eta) \\ \eta' \neq \eta_c}} |T[\eta']| \cdot d(\text{parent}(\eta))^z \right) \\
 &\quad - \left( \text{Cost}(T[\eta_c], c_z(\eta)) + \sum_{\substack{\eta' \in \text{children}(\eta) \\ \eta' \neq \eta_c}} |T[\eta']| d(\eta)^z \right) \\
 &\geq \left( \text{NodeCost}(\eta_c, z) + \sum_{\substack{\eta' \in \text{children}(\eta) \\ \eta' \neq \eta_c}} |T[\eta']| \cdot d(\text{parent}(\eta))^z \right) - \\
 &\quad \left( \text{Cost}(T[\eta_c], c_z(\eta)) + \sum_{\substack{\eta' \in \text{children}(\eta) \\ \eta' \neq \eta_c}} |T[\eta']| d(\eta)^z \right) \\
 &= (\text{NodeCost}(\eta_c, z) - \text{Cost}(T[\eta_c], c_z(\eta))) \\
 &\quad + \sum_{\substack{\eta' \in \text{children}(\eta) \\ \eta' \neq \eta_c}} |T[\eta']| \cdot (d(\text{parent}(\eta))^z - d(\eta)^z)
 \end{aligned}$$

$$\begin{aligned}
 &= \text{CostDecrease}(\eta_c, z) + \sum_{\substack{\eta' \in \text{children}(\eta) \\ \eta' \neq \eta_c}} |T[\eta']| \cdot (d(\text{parent}(\eta))^z - d(\eta)^z) \\
 &\geq \text{CostDecrease}(\eta_c, z),
 \end{aligned}$$

where both inequalities are due to  $d(\text{parent}(\eta)) \geq d(\eta)$ .

Lastly, we consider the cost decreases of the other children of  $\eta$ . Let  $\eta_o \neq \eta_c$  be any other child of  $\eta$ . Then by Lemma B.4, we have  $\text{CostDecrease}(\eta_c, z) \geq \text{CostDecrease}(\eta_o, z)$ . This concludes by showing that the costs are monotonically non-decreasing along paths to the root.

It remains to be shown that the cost-decreases are necessarily non-negative. For this, we rely on the fact that the original distances in  $T$  are non-negative. Since we already know that they are monotonically non-decreasing, we, therefore only have to show that the cost-decrease of placing a center at a leaf is non-negative. To this end, let  $\ell$  be any leaf. By Corollary 2.4. Then  $\text{CostDecrease}(\ell, z) = \text{NodeCost}(\ell, z) - \text{Cost}(T[\ell], c_z(\ell))$ . However,  $\text{Cost}(T[\ell], c_z(\ell)) = d(\ell)$  while  $\text{NodeCost}(\ell, z) \geq d(\text{parent}(\ell))$ . Thus,  $\text{CostDecrease}(\ell, z) \geq d(\text{parent}(\ell)) - d(\ell) \geq 0$ .

□

We note that the cost-decrease at a leaf  $\ell$  is not necessarily 0. Interpreting the cost-decreases as a relaxed ultrametric means that  $d^{\text{cost-decrease}}(\ell, \ell) \neq 0$ . This is why we required the definition of relaxed ultrametrics rather than standard ultrametrics.

The next lemma shows that not only do these cost-decreases satisfy the conditions in Corollary 2.4, but they are also monotonically *increasing* in all relevant settings. In other words, Lemma B.8 is saying that if  $T[\eta_1]$  has a non-zero cost, then placing a center there decreases this cost by a non-zero amount.

**Lemma B.8.** *Let  $T$  be an ultrametric LCA-tree satisfying the conditions in Corollary 2.4 and let  $z$  be a positive integer. For any leaf  $\ell \in T$ , let  $p(\ell) = [\ell, \eta_i, \dots, \eta_j, r(T)]$  be the path from  $\ell$  to the root of the cost-decrease LCA-tree. Then for all  $\eta_1, \eta_2 \in p(\ell)$  such that  $d(\eta_1) > 0$ ,  $\eta_1 \preceq \eta_2 \implies \text{CostDecrease}(\eta_1) < \text{CostDecrease}(\eta_2)$ .*

*Proof.* By Lemma B.7, we know that the cost decreases are monotonically non-decreasing along paths to the root. Thus, it remains to show that the cost-decrease at a node  $\eta_1$  is non-zero if  $d(\eta_1) > 0$ . To do this, we first decompose the cost-decrease at  $\eta$ :

$$\begin{aligned}
 \text{CostDecrease}(\eta_1) &= \text{NodeCost}(\eta_1) - \text{Cost}(T[\eta_1], c_z(\eta_1)) \\
 &= |T[\eta_1]| \cdot d(\text{parent}(\eta_1)) - \text{Cost}(T[\eta_1], c_z(\eta_1)) \\
 &= d(\text{parent}(\eta_1)) \cdot \left( \sum_{\eta' \in \text{children}(\eta_1)} |T[\eta']| \right) - \text{Cost}(T[\eta_1], c_z(\eta_1)).
 \end{aligned}$$

Now, let  $\eta_c$  be the child of  $\eta_1$  containing  $c_z(\eta_1)$ . Much as in the proof of Lemma B.7, we rewrite the above in terms of the costs in  $\eta_c$  and the costs in the other children:

$$\begin{aligned}
 \text{CostDecrease}(\eta_1) &= d(\text{parent}(\eta_1)) \cdot \left( |T[\eta_c]| + \sum_{\substack{\eta' \in \text{children}(\eta_1) \\ \eta' \neq \eta_c}} |T[\eta']| \right) \\
 &\quad - \text{Cost}(T[\eta_c], c_z(\eta_1)) - d(\eta_1) \cdot \left( \sum_{\substack{\eta' \in \text{children}(\eta_1) \\ \eta' \neq \eta_c}} |T[\eta']| \right) \\
 &= \left( \sum_{\substack{\eta' \in \text{children}(\eta_1) \\ \eta' \neq \eta_c}} |T[\eta']| \right) \cdot (d(\text{parent}(\eta_1)) - d(\eta_1))
 \end{aligned}$$

$$\begin{aligned}
 & + d(\text{parent}(\eta_1)) \cdot |T[\eta_c]| - \text{Cost}(T[\eta_c], c_z(\eta_1)) \\
 & \geq d(\text{parent}(\eta_1)) \cdot |T[\eta_c]| - \text{Cost}(T[\eta_c], c_z(\eta_1)) \\
 & \geq d(\text{parent}(\eta_1)) \cdot |T[\eta_c]| \\
 & \geq d(\eta_1) \cdot |T[\eta_c]| \\
 & > 0
 \end{aligned}$$

where the first inequality is by the fact that  $d(\eta_1) \leq d(\text{parent}(\eta_1))$ , the second is due to the fact that the costs are strictly non-negative, the third is by the fact that  $d(\text{parent}(\eta_1)) > d(\eta_1)$ , and the fourth inequality is by the assumption that  $d(\eta_1) > 0$ .

□

Lemma B.8 allows us to address the fact that the cost-decrease at a node  $\eta$  was defined under the assumption that  $\eta$ 's parent is marked. By Lemma B.8, the cost decreases are either strictly increasing along paths to the root or are 0. Thus, if two nodes have equivalent non-zero cost-decreases, their subtrees must be disjoint.<sup>6</sup> As a result, when we go through the nodes sorted by their cost-decreases, it will always be the case that the next node we place will either have its parent marked or will have a cost-decrease of 0 (in which case it is irrelevant in terms of the optimal solutions).

We finally move to our last lemma, which shows that greedily maximizing cost-decreases results in an optimal  $(k, z)$ -clustering over the LCA-tree.

**Lemma B.9.** *The optimal  $(k, z)$ -clustering solution over an LCA-tree can be obtained by greedily choosing the  $k$  centers that, at each step, maximize the cost-decrease.*

*Proof.* We show this by contradiction. Consider that there is a solution that was obtained greedily and another, different one that is actually optimal. We now map every center in the “optimal” solution to its closest center in the “greedy” one and observe how the optimal solution’s cost changes. There are two cases that can occur: either all of the greedy centers receive one optimal center each, or there is at least one greedy center that receives more than one optimal center and another that receives none.

We start with the case where each greedy center  $c_g$  has one optimal center  $c_o$  mapped to it. By definition,  $c_o$  must be in the set of points that are assigned to  $c_g$ . Thus, it is either (a) in  $T[c_g]$  or (b) in another subtree whose parent was marked by  $c_g$ . In case (a), Lemma B.4 states that the greedy algorithm must have chosen  $c_o$ . In case (b), by the greedy algorithm,  $T[c_g]$  has greater cost minimization than  $T[c_o]$ . Thus, we can decrease the cost of the optimal solution by replacing  $c_o$  with  $c_g$ , giving a contradiction. Therefore, we conclude that if one optimal center was mapped to each greedy center, then the solutions must be equivalent.

It remains to consider the case where more than one optimal center was mapped to a greedy center  $c_g$ . WLOG, let there be two optimal centers  $c_o^1$  and  $c_o^2$  that are mapped to  $c_g$ . By a similar argument as above,  $c_g$  must be the same as one of these optimal centers, i.e.  $c_g = c_o^1$ . By extension,  $c_o^2 \neq c_g$ . Now consider the greedy center elsewhere in the tree that had no optimal center mapped to it. Call this center  $c'_g$ . This means that the greedy algorithm had both  $T[c_o^2]$  and  $T[c'_g]$  available to it but chose  $T[c'_g]$ . Thus,  $\text{CostDecrease}(T[c_o^2]) < \text{CostDecrease}(T[c'_g])$ . We can therefore decrease the cost of the optimal solution by replacing center  $c_o^2$  with center  $c'_g$ . This gives the desired contradiction.

□

This brings us to the primary result of this chapter, restated from before:

**Theorem B.3.** *Let  $T$  be an LCA-tree satisfying the conditions in Corollary 2.4 and let  $z$  be any positive integer. Then there is an algorithm that runs in  $\text{Sort}(n)$  time and finds the optimal  $(k, z)$ -clustering solutions on  $T$  for all  $k \in \{1, \dots, n\}$ . Furthermore, these solutions are hierarchical.*

*Proof.* We use Algorithm 3 from Lemma B.6 to find the cost-decreases for  $(k, z)$ -clustering in the LCA-tree. By Lemma B.7, these satisfy the strong triangle inequality. Furthermore, by Lemma B.9, greedily choosing centers that maximize the cost-decreases gives an optimal  $(k, z)$ -clustering.

<sup>6</sup>Formally, for two nodes  $\eta_1, \eta_2 \in T$  with  $d(\eta_1) > 0$  and  $d(\eta_2) > 0$ , we have that  $d(\eta_1) = d(\eta_2) \implies T[\eta_1] \cap T[\eta_2] = \emptyset$ .

Thus, running farthest-first traversal on the cost-decrease LCA-tree will give the partitions for  $(k, z)$ -clustering solutions. However, we must be a bit careful here: while running a naive farthest-first traversal on the cost-decrease LCA-tree will give the *partitions* of the optimal  $(k, z)$ -clustering solutions, it will not necessarily give the correct *centers*. This is due to the fact that the farthest-first traversal is optimal regardless of which center we pick for every subtree. Luckily, we have already addressed this. When considering the LCA-tree of cost-decreases, we have a mapping between every cost-decrease and the center that induces it. Thus, we will perform the farthest-first traversal by placing the nodes' corresponding  $z$ -centers. This ensures that both the partition *and* the centers align with the optimal  $(k, z)$ -clustering solutions.

Putting this all together, our final algorithm – Algorithm 4 – is quite simple. We first run Algorithm 3 to return a list of nodes and their cost-decreases. Importantly, Algorithm 3 returns only the cost-decreases for those nodes  $\eta'$  with  $c_z(\eta') \neq c_z(\text{parent}(\eta'))$ . We then sort this list in  $\text{Sort}(n)$  time. By the discussion after Lemma B.8's proof, we can safely go through this list and place the nodes' corresponding  $z$ -centers: the nodes with non-zero cost-decrease will always have their parent marked. The nodes with zero cost-decrease come at the end of the sorted list and do not affect the optimality of the solution. By Lemma B.4, each corresponding  $z$ -center is immediately optimal in every node that it marks. Thus, Algorithm 4 optimally solves the  $(k, z)$ -clustering objective in LCA-trees which satisfy the conditions in Corollary 2.4. The bottleneck in this algorithm remains the time to sort the  $O(n)$  internal values in the LCA-tree.

---

**Algorithm 4** Ultrametric-kz

**Input:** LCA-tree  $T$ 

```

1: Costs, CostDecreases = { }, { }
2: GetCostDecreases( $T.\text{root}$ , Costs, CostDecreases) // pass-by-reference
3: CostDecreases = OrderedDict(CostDecreases)
4: for  $\eta \in \text{CostDecreases}$  do
5:   Place center at  $c_z(\eta)$ 
6: end for

```

---

□

Together, Theorem B.3 and Lemma B.1 prove Theorem 3.2 from the main body of the paper.

## C. Further Details for Section 4 - Choosing a Partition

### C.1. Ultrametric Elbow Method

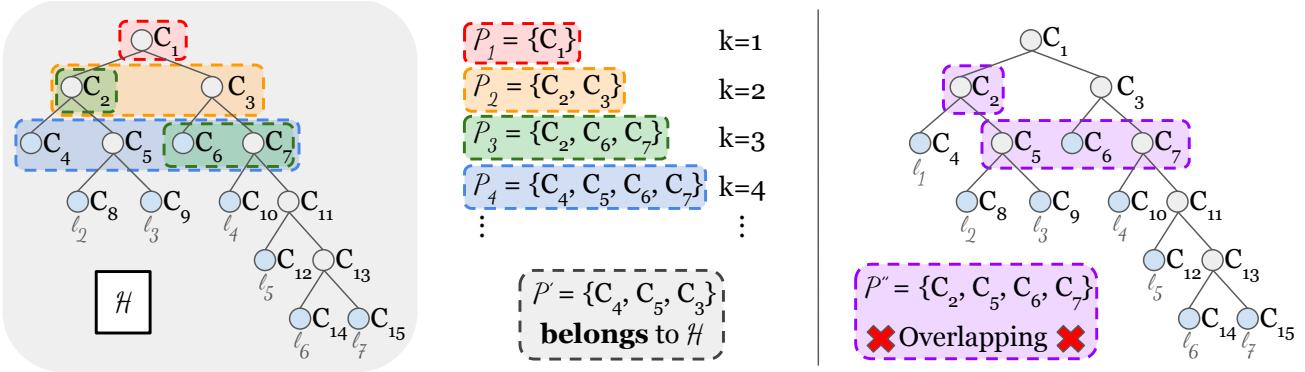


Figure 6. Left: an example hierarchy  $\mathcal{H} = \{\mathcal{P}_1, \mathcal{P}_2, \dots\}$ ;  $\mathcal{P}' \notin \mathcal{H}$  is then an example partition which *belongs to*  $\mathcal{H}$ . Right:  $\mathcal{P}''$  is not a partition since  $\ell_2$  and  $\ell_3$  each belong to clusters  $C_2$  and  $C_5$ .

We begin by quickly showing Corollary 4.1, which holds as a direct consequence of Lemma B.8 in Section B.2:

**Corollary 4.1.** *Let  $\mathcal{P}$  and  $\mathcal{L}$  correspond to the  $n$  partitions and losses obtained in accordance with Theorem 3.2 for the  $(k, z)$ -clustering objective. Let  $\Delta_i = \mathcal{L}_{i+1} - \mathcal{L}_i$ . Then either  $\Delta_i < \Delta_{i+1} \leq 0$  or  $\Delta_i = \Delta_{i+1} = 0$  for all  $i \in [n - 1]$ .*

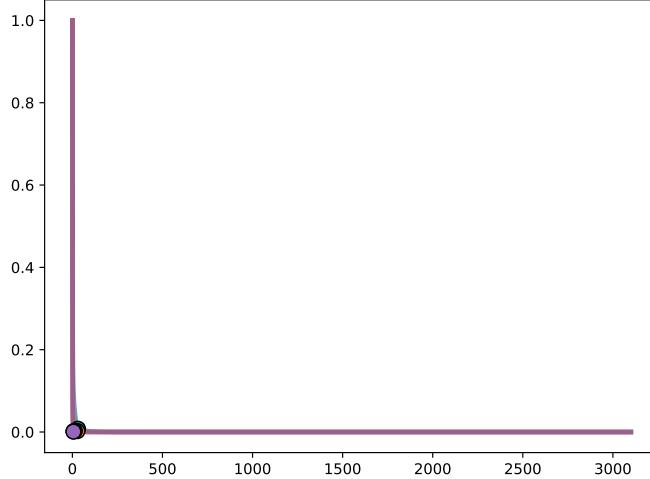


Figure 7. The same elbow plot as in Figure 2 with values of  $k$  from 1 to  $n$ . The chosen elbows are highlighted with circles.

*Proof.* Lemma B.8 states that the cost-decreases associated with nodes in an LCA-tree monotonically increase along paths from any leaf to the root or are all 0. By Lemma B.9, we solve  $(k, z)$ -clustering using farthest-first traversal over the LCA-tree of cost-decreases. I.e., we start at the root and greedily pick the cluster that gives the maximal cost-decrease at that step. Consequently, each cost-decrease we pick must be smaller than the previous one.

Thus, let  $\Delta'_k = \mathcal{L}_k - \mathcal{L}_{k+1} = -\Delta_k$ . By Lemma B.8, we have  $\Delta'_k > \Delta'_{k+1} \geq 0$  or  $\Delta'_k = \Delta_{k+1} = 0$ . Plugging in  $\Delta = -\Delta'$  completes the proof.  $\square$

**Choosing the elbow** Although there are many methods for finding the elbow index, we are in the privileged setting where a single elbow exists and is clearly delineated. Inspired by Shi et al. (2021), we simply define the elbow as the index where there most strongly appears to be a right angle. Namely, let  $\vec{v}_i = (i, \mathcal{L}_i)$  be the  $(x, y)$  position of the  $i$ -th point in the elbow plot. Then, for every index  $k \in \{2, \dots, n-1\}$ , let  $\theta_k$  be the angle induced by the vectors<sup>7</sup>  $(\vec{v}_1 - \vec{v}_k)$  and  $(\vec{v}_k - \vec{v}_n)$ . We define the elbow as being at the index  $k$  where  $\theta_k$  is closest to 90 degrees. Since Theorem 3.2 gives us all of the partitions for  $k \in \{1, \dots, n\}$  simultaneously, this index can be found  $O(n)$  time given the cluster hierarchy.

We note that Figure 2 only plots the elbow curves for values of  $k$  up to 100. This is because plotting until  $k = n$  makes the plot look in practice like a vertical line followed by a horizontal line, as seen in Figure 7. However, we use all values of  $k$  from 1 to  $n$  when choosing the elbow.

## C.2. Agglomerative Clustering Algorithms under our Framework

Corollary 2.4 and Theorem 3.2 imply that a large set of agglomerative clustering algorithms can be interpreted as  $k$ -center over various relaxed ultrametrics. As an example, consider the complete linkage algorithm (Jarman, 2020). Here, one starts with every point in its own cluster and recursively merges those clusters  $C_i, C_j$  which have the smallest merge distance  $\arg \min_{C_i, C_j} \max_{x_i \in C_i, x_j \in C_j} d(x_i, x_j)$ . Importantly, with each subsequent merge, the merge distances are monotonically non-decreasing. Thus, by Corollary 2.4, labeling each cluster in the hierarchy by its merge distance gives us a relaxed ultrametric and, by Theorem 3.2,  $k$ -center over this ultrametric gives us the complete-linkage hierarchy. Indeed, this is true of *any* agglomerative clustering method where the merge distances progressively grow as we approach the root cluster.

## C.3. Thresholding

We quickly explain how one can partition a cluster hierarchy by thresholding. We assume that the cluster hierarchy  $\mathcal{H}$  has every internal node labeled by its cluster's cost. As discussed in the main body of the paper, this constitutes a relaxed ultrametric.

Now let  $\varepsilon$  be any threshold value. We can return the set of clusters that have cost less than  $\varepsilon$  by depth-first search in  $O(n)$

<sup>7</sup>In practice, we normalize the  $k$  values and the costs to be in  $[0, 1]$  so that the scales are comparable.

time. This is precisely what DBSCAN\* does on the dc-dist relaxed ultrametric. Namely, DBSCAN\* returns clusters of core points that are within  $\varepsilon$  of each other under the dc-dist. Under the dc-dist's relaxed ultrametric definition, this is specifically the set of nodes with a value less than  $\varepsilon$ .

#### C.4. Cluster Value Functions

The idea is a generalization of the excess-of-mass measure in HDBSCAN: we have a user-defined function  $v(C)$  which assigns a non-negative value to each cluster in the hierarchy. I.e.,  $v(C) \geq 0$  for all clusters  $C \in \mathcal{H}$ . Then the *best* partition maximizes the sum of these values:

**Definition C.1.** Given a value function  $v$ , we define the **best** partition under  $v$  as the partition that maximizes the sum of valuations. I.e.  $\mathcal{P}_v(\mathcal{H}) = \arg \max_{\mathcal{P} \text{ belonging to } \mathcal{H}} \sum_{C \in \mathcal{P}} v(C)$ .

This brings us to the following result:

**Theorem C.2.** Let  $\mathcal{H}$  be a cluster hierarchy. Let  $v$  be a function such that, for all  $C \in \mathcal{H}$ ,  $v(C)$  can be obtained in  $O(1)$  time. Then there exists an algorithm to find the best partition  $\mathcal{P}_v(\mathcal{H})$  in  $O(n)$  time.

*Proof.* The algorithm for finding the best clustering is essentially Algorithm 3 from [Campello et al. \(2013\)](#). We provide our own version of it in Algorithm 5 for completeness' sake.<sup>8</sup> It works by depth-first search where, at each node, we simply calculate its value and compare it against the sum of the values of its children. Since calculating the value takes  $O(1)$  time and there are  $O(n)$  nodes in the cluster hierarchy, the algorithm therefore runs in  $O(n)$  time.

We prove its correctness inductively. In the base case, we have a single leaf whose best clustering is simply the leaf itself. In the inductive case, we are given a cluster  $C$  in the hierarchy and have the best clusterings of  $C$ 's children. We seek to find  $B_v(C)$ . By the non-overlapping requirement, if we include  $C$  in  $B_v(C)$ , then we cannot include any of its children. Similarly, since the values of the children are non-negative, if we include one child, then we may as well include all of them. Thus,  $B_v(C)$  is either  $\{C\}$  or  $\{C' : C' \in \text{children}(C)\}$ .

---

#### Algorithm 5 BestClustering

---

**Input:** node  $C$  in an cluster hierarchy;

**Output:** the best clustering under this node  $B_v(C)$ , the value of the clustering  $v(C)$ ;

```

1: if  $C$  is leaf then
2:    $B_v(C) = \{C\}$ 
3:   Return  $B_v(C)$ 
4: end if

5: ChildValues = 0
6: ChildClusterings = []
7: for  $C' \in \text{children}(C)$  do
8:    $B_v(C'), v(C') = \text{BestClustering}(C')$ 
9:   ChildValues +=  $v(C')$ 
10:  ChildClusterings.append( $B_v(C')$ )
11: end for
12: if  $v(C) > \text{ChildValues}$  then
13:   Return  $\{C\}, v(C)$ 
14: end if
15: Return ChildClusterings, ChildValues

```

---

□

**The stability cluster value function.** We now introduce the stability cluster value function from [Campello et al. \(2013\)](#). Let  $C$  be any cluster in a hierarchy and let  $C'$  be  $C$ 's parent in that hierarchy. Then the stability objective can be roughly

<sup>8</sup>Note, we write  $B_v(C)$  to refer to the best clustering of the cluster hierarchy rooted at  $C$ .

phrased as the following:

$$v_E(C) = |C| \cdot \left( \frac{1}{\mathcal{L}(C)} - \frac{1}{\mathcal{L}(C')} \right) \quad (1)$$

We note that our description of the stability criterium differs slightly from the original function discussed in (Campello et al., 2013; 2015; McInnes & Healy, 2017). Namely, we omit here the notion that singleton points may fall out of the clustering as it is not easy to represent in our notation and the differences are negligible for the purposes of this discussion. Our implemented stability criterion is the original (correct) one. For a full discussion of the original function, we refer the reader to the referenced literature.

In either case, the stability value function can be interpreted as emphasizing those clusters that have a large number of points and have significantly lower costs than their parent. While we find that the stability criterion performs well in the  $k$ -center hierarchy, it can produce sub-par partitions when applied in the  $(k, z)$ -clustering hierarchies. This is because the per-point cost in the  $(k, z)$ -clustering task is comparable to the per-cluster cost in the  $k$ -center objective.

To be consistent with the literature (Campello et al., 2013), we utilize the stability cluster value function in the noisy setting. Given a user defined parameter  $\mu$  representing the minimum cluster size, we first prune the tree so that all nodes with fewer than  $\mu$  children are removed. We then run the stability value function over the remaining points. This selects a set of internal nodes to represent the clusters. Finally, we re-introduce the points which were pruned away. If the re-introduced points are in the sub-tree of a cluster, we assign them to the cluster. If, instead, they are not below a cluster found by the stability function, we label them as “noise”.

### C.5. Further details on the dc-dist ultrametric

We begin by proving Proposition 5.3:

**Proposition 5.3.** *Let  $(L, d')$  be a metric space. Then, the dc-dist over  $L$  is a relaxed ultrametric.*

*Proof.* Note that it is known that the minimax distances over a space constitute an ultrametric (Fischer et al., 2003). Thus, we have that  $d_{dc}(\ell_i, \ell_j) = d(\ell_j, \ell_i)$  and that  $d_{dc}(\ell_i, \ell_j) \geq 0$  for all  $\ell_i, \ell_j$ .

To show that the dc-dist is a relaxed ultrametric, we will leverage that the only difference between it and the minimax distance is that the minimax distance of a point to itself is 0 while the dc-dist of a point to itself is its mutual reachability. Thus, we will prove that the dc-dist is a relaxed ultrametric by showing that the distance to itself is non-negative and that it inherits the strong-triangle inequality from the minimax ultrametric. Together, these imply the two properties for Corollary 2.4.

First, note that  $d_{dc}(\ell_i, \ell_i) = \max(\|\ell_i - \ell_i\|, \kappa_\mu(\ell_i), \kappa_\mu(\ell_i)) = \max(0, \kappa_\mu(\ell_i)) = \kappa_\mu(\ell_i)$ . Thus, the dc-dist of a point to itself is the distance of  $\ell_i$  to its  $\mu$ -th nearest neighbor in the ambient metric. This is necessarily non-negative. Similarly, the mutual-reachabilities for any other pair of points are also necessarily non-negative. Because the dc-dist is the minimax distance over the pairwise mutual reachabilities and the pairwise mutual reachabilities are all non-negative, the dc-dist must also be.

Let us now show that the dc-dist satisfies the second property of Corollary 2.4. First, notice that the dc-dist of a point to itself is necessarily less than or equal to the dc-dists of that point to any other point in the set, i.e.,  $d_{dc}(\ell_i, \ell_i) \leq \max(d_{dc}(\ell_i, \ell_j), d_{dc}(\ell_j, \ell_i))$  for all  $\ell_j$ . To see this, consider the mutual reachability of  $\ell_i$  to any other point  $\ell_j$  is necessarily greater than  $m_\mu(\ell_i, \ell_i)$ . Namely,  $m_\mu(\ell_i, \ell_j) = \max(d'(\ell_i, \ell_j), \kappa_\mu(\ell_i), \kappa_\mu(\ell_j)) \geq \kappa_\mu(\ell_i)$ . As a result, any step in the mutual reachability MST that originates at  $\ell_i$  will be at least as large as  $m_\mu(\ell_i, \ell_i) = d_{dc}(\ell_i, \ell_i)$ . Put simply, a point’s closet point under the dc-dist is itself. Pairing this with the inheritance of the strong triangle inequality from the minimax distance gives us the second property of Corollary 2.4.

Thus, we have shown the two properties for Corollary 2.4: the dc-dists are non-negative and are non-decreasing as we traverse the DC tree from any leaf to the root. Consequently, it is a relaxed ultrametric.  $\square$

**Relationship to DBSCAN and HDBSCAN.** First, Campello et al. (2013) showed that one can obtain DBSCAN\* partitions using the single-linkage hierarchy over the mutual reachabilities. Rephrasing into this paper’s notation, they showed that one can obtain DBSCAN\* embeddings by thresholding the dc-dist’s LCA-tree at a user-defined value  $\varepsilon$ ; i.e., removing all nodes  $\eta$  in the LCA-tree with  $d(\eta) > \varepsilon$ . Subsequently, Beer et al. (2023) proved that  $k$ -center can be optimally solved over the dc-dist and that these partitions correspond to single-linkage over the mutual reachabilities. In this sense, one can think

of  $k$ -center on the minimax distances as equivalent to single-linkage clustering. Lastly, Campello et al. (2013) showed that, given the dc-dists's LCA-tree, one can choose a partition from it by optimizing the EoM cluster merging criterium over the pruned tree.

## D. Runtime Speedups / Efficient Operations

We now give an overview of how we implement the theory from Section B.1 and B.2 in practice in our codebase. We assume that we have already computed an ultrametric so that queries to the ultrametric require  $O(1)$  time. From this, we describe how we build an LCA-tree, how to transform it to other clustering hierarchies, and how we extract partitions from it. We center this discussion on the dc-dist's dc-tree as a point of reference since we believe that this is the most common use case of our framework. However, we note that the discussion applies immediately to any relaxed ultrametric.

### D.1. Building and utilizing the tree efficiently (theoretical complexities)

The overall structure of our implementation is the following:

#### 1. Building the tree

- (a) Compute annotations over the dc-tree bottom up. These contain information that we will use for creating hierarchies.
- (b) Sort the annotations.
- (c) Create the hierarchy in a way that takes  $O(n)$  time utilizing the parent pointers in the annotations and in-place pointer updating.

#### 2. Get each solution in $O(n)$

- (a) Annotate each node in the tree with the  $k$  that resulted in creating this node, given that the  $k$ -th center is chosen.
- (b) Top-down (or bottom-up) algorithm that cuts all edges in the tree going from  $k$ -annotation  $> k$  to  $k$ -annotation  $\leq k$ . The result is all nodes above the cut.
- (c) Small edge case to deal with for nodes with  $> 2$  children.

#### 3. Initialize smart pointer access to the leaves in the tree so that internal nodes can get their leaves in constant time.

Now, each step in more detail:

##### D.1.1. BUILDING THE TREE

How we build the tree in  $O(n \cdot \log(n))$ .

Two main functions:

- Annotate tree
- Create hierarchy

#### Annotating the tree

Annotate tree creates the annotations for each internal node in the tree. Each annotation is the following:  $A_k = [cost\_decrease, center, parent\_pointer, tree\_node]$ .

**Idea D.1.** *Only maximal annotations for a given center will be picked with the greedy algorithm.*

By maximal annotations, we mean those highest in the tree corresponding to a specific center, i.e., the annotation for that center with the highest annotated cost-decreases. This is very easy to see, as only maximal annotations will be children of marked paths of other, already picked, centers in the tree.

**Idea D.2.** *The cluster corresponding to the parent center  $p'$  of the maximal annotation of  $p$  is exactly the cluster/set of points from which choosing  $p$  will exclusively take points.*

This comes from the fact that each annotation is maximal in its corresponding subtree, which means that the maximal subtree above always will have chosen that center first. Also, any other center  $p''$  chosen within that subtree of the maximal annotation of  $p'$  will have taken disjoint points from that center  $p'$ 's cluster that  $p$  cannot otherwise a contradiction and  $p''$  should have been the parent annotation of  $p$ .

We formally define the parent center as the single center in solution  $k - 1$  that contains all points of the  $k$ 'th center, which will get assigned for  $k$ . We now get the following insights needed to make building the tree efficient:

- 1: We only need to store one annotation for each center, where we just store the highest / best annotation cost-decrease for that center. This can be updated as the annotations are computed.
- 2: Each annotation/center can contain a pointer to the parent center.

Having these, we now have all the information required to build the tree - for each new center, we have exactly the cost-decrease corresponding to picking it and the parent center. We can sort the list of annotations, and each k solution will correspond to the first k centers picked in that order.

### Building the tree from the annotations

First, we sort the annotations. Picking the first annotation corresponds to the root node and cluster of all points with that center. Picking a subsequent new center will always correspond to a split in the tree, splitting up the parent center's cluster. One side of the split will be what is left over from the parent cluster, and the other side will be the new cluster for this  $k$ . The only caveat is that if there are multiple, some number  $a$ , annotations with the same cost-decrease and parent, we only get  $a + 1$  nodes from this multi-split. We get a node for each of the  $a$  new centers and the  $a + 1$ 'st node for what is left over in the partition of the parent node's cluster.

Now, the key is that the cost-decreases are decreasing in the sorted order. This means that any subsequent annotations with a parent that already has been split up will always create a new split at the lowest possible node/cluster corresponding to that parent center, and all nodes above that also correspond to that parent center will never get new splits from them. This means that we can maintain for each center the current lowest/active node, which will be where any new center pointing to it should split from. This can easily be managed within the annotations with a pointer that is updated as the tree is created.

The steps are then generally the following at a given  $k$ 'th annotation for some center  $c_k$  in the traversal of the sorted list:

Case 1: If the parent node  $p$  corresponding to some center  $c_p$  has no offspring, add two new nodes below. One for the new annotation and  $p'$  for the old center representing that nodes have been taken from it. For the new node for  $c_k$ , update the annotation to point to it.

Case 2: If the parent  $p$  already has offspring associated with a higher cost-decrease, then add the new split nodes below  $p'$  that then have to exist as that other offspring will have created that in case 1. Update the parent center's annotation to now point to  $p'$  instead of  $p$ , as we can now be sure everything pointing to that center should never be added to  $p$  but instead to  $p'$  or further below at a later point. For the new node for  $c_k$ , update the annotation to point to it.

Case 3: The other offspring of the parent node has the same cost-decrease associated with the current annotation. This means the  $p'$  has already been created, so just add a singular new node as a child of  $p$  corresponding to center  $c_k$ . For the new node for  $c_k$ , update the annotation to point to it.

**Complexity** Computing the annotations is a single-pass bottom-up algorithm over the tree that does constant work in each tree node. As there are  $O(n)$  nodes in the tree, this step has a worst-case complexity of  $O(n)$ .

Sorting the list of annotations has a worst-case complexity of  $O(n \cdot \log(n))$  as there are  $n$  annotations to sort.

Computing the tree from the sorted annotations does a single scan over the list of annotations, with constant work in any iteration and, therefore, a worst-case complexity of  $O(n)$ .

Therefore, the total worst-case complexity is  $O(n \cdot \log(n))$ .

#### D.1.2. OPTIMIZING THE TREE: RESOLVING TIES BETTER

An important observation is that due to the number of equidistant points under the relaxed ultrametric, many centers will often have tied distances to points between them. For example, consider that we have placed our first center  $c_1$  and it marks every node along the path to the root. Now let there be two subtrees  $T_i$  and  $T_j$  which are *not* marked but whose parent node  $i$  is marked. We now place a center  $c_2$  in  $T_i$ . This creates a new path of markings that stops at the root of  $T_i$ .

Now notice that *every* node in  $T_j$  is equidistant to center  $c_1$  and to center  $c_2$ . Thus, we can leave  $T_j$  assigned to  $c_1$ , or we can re-assign it to  $c_2$ , and the cost remains *unchanged*. Over the course of placing  $k$  clusters, there will inevitably be many such ties, implying that there are an exponential number of optimal solutions!

The previous theory described this setting by simply utilizing a “first come, first served” principle and never re-assigning points to new clusters. However, under the dc-dist and other minimax path distances, this may result in unintuitive clusterings. To see this, consider a set of 10 copies of the same cluster which are equally spaced apart. Let there be two centers assigned to these:  $c_1$  is placed in the first copy of the cluster, and  $c_2$  is placed in the second copy of the cluster. The remaining 8 clusters must now be assigned to either  $c_1$  or  $c_2$ . However, due to them being equally spaced apart, we can assign each of these 8 clusters to either center and the cost will remain the same.

Our recommendation on how to resolve this is to introduce a secondary heuristic for tie-breaking. Namely, if a subtree is equidistant to two centers under the ultrametric, assign it to the center which is closest to in the data’s original metric. I.e., if our ambient distance metric is Euclidean, then we use the Euclidean distance as a tie-breaker. This provides the intuitive clustering that one expects when looking at groups of points.

In practice, we achieve this by giving each internal node in the tree a representative point - the Euclidean mean of its leaves. Each annotation not yet marked (and thus chosen at a later  $k$ ) maintains the closest center based on the Euclidean distance to its representative, which is resolved between any center that marks its parent. At the  $k$  where this annotation itself is chosen, we have found the best annotation, and thus, the internal node that this annotation and sub-tree of nodes itself should become a subtree of. We update all annotation pointers in this way by placing the centers one by one, following the path from the center to the leaf checking distances to unmarked sub-trees, and updating the pointer if the new center is closer than other tied centers that had already been marked.

As this is just a processing step over the annotations before the tree is actually constructed, it can easily be plugged in/out based on a boolean flag. Furthermore, the choice of heuristic can also easily be changed, but here it should be kept in mind that it might influence the running time if any complex function is used.

The choice of using the mean point as a representative and comparing Euclidean distances is based on the following: under the dc-dist, one often gets one cluster which consists of multiple equidistant sub-clusters. These equidistant sub-clusters often follow snaky paths throughout the ambient space. Let us now suppose our large cluster gets split into two, implying that its sub-clusters must get assigned to one of these two new groups. Since the sub-clusters are all equidistant, they can be arbitrarily assigned to either of the two groups. By breaking ties so that the sub-clusters are closest to each other in the ambient space, we dramatically increase the chance of the sub-clusters getting re-assigned in an intuitive manner.

### Complexity

This process requires traversing the tree from the leaf to the root for all centers placed. To construct the full hierarchy/dendrogram/tree of solutions, we do constant time calculations (scales with dimensionality) at each visit of an internal node, as it only requires computing the Euclidean distance of two points. This means that the worst-case complexity becomes  $O(n^2)$  in the case of a fully unbalanced tree, but in practice and expectation, this will be much closer to  $O(n \cdot \log(n))$ .

#### D.1.3. FIND SOLUTION FOR A GIVEN $k$

To find the solution for a given  $k$  efficiently, we start out by annotating each node of the new tree with additional information. Generally, the insight is that going from  $k$  to  $k + 1$  corresponds to splitting up a single node/cluster into two new nodes in the tree. A slight detail is that some nodes will have multiple children, where the “split” is implicit as all the children split from the same node. We can mark at which  $k$  each node is split from the parent in constructing the tree, where  $k$  is just the current iteration. We simply mark all new nodes created in an annotation with that  $k$ .

To then recover a solution for a given  $k$ , all that is required is to cut all edges in the tree going from  $k' \geq k$  to  $k' < k$ . The solution is the clusters associated with all the nodes above the cut. The only note is that if a leaf is reached and no cut is found in that path, the leaf is just part of the solution.

The correctness of this algorithm comes from the fact that any clusters below this cut only existed at a higher  $k''$  than the solution for  $k$  we are recovering.

### Complexity

Marking the tree is done with a constant factor added to the complexity of the list traversal of building the tree, so  $O(n)$ .

Finding the cut can also be done in a single traversal of the tree top-down, simply stopping the traversal of a given branch when a cut is found and returning the nodes. So this is also  $O(n)$ .

Furthermore, the solution for a given  $k$  can be stored, and the algorithm to find another  $k$  can be "resumed" from this solution, making finding similar  $k$  values very fast, almost constant.

#### D.1.4. FIND LEAVES FOR A GIVEN NODE QUICKLY / LABELING A SOLUTION

We create an array of the id's in the leaves, where each id is inserted in postfix order. This means that looking at the tree, each internal node of the tree will correspond to a continuous area of that array. If we store the bounds of those segments in each internal node, the nodes corresponding to a node can be returned in constant time by just returning that continuous segment of data of the new array. Only a single postfix order traversal of the tree is required to set up this array and pointers in the internal node, which then takes  $O(n)$ .

However, if we want to recover the labels in the standard form of each label being in the order that the points were provided, a traversal of the recovered solution is required, putting the corresponding label in the right place in an output array of labels. Simply put, the label value of point  $p$  has to be inserted in place  $p$  of the output array of labels. This requires worst-case  $O(n)$  complexity for a given  $k$  solution, as it is just a linear scan with constant work for each of the  $n$  points.

## E. The Algorithms

We now describe how these algorithms are implemented. We use the terminology  $n$ -ary dc-tree to refer to a non-binary LCA-tree storing dc-dist relationships. Again, this immediately transfers to all relaxed ultrametrics.

Algorithm 6 describes how we find the corresponding  $z$ -centers and the costs associated with them in practice. Algorithm 7 describes how we use this information to obtain the corresponding  $(k, z)$ -clustering hierarchy. Algorithm 8 shows how we extract clusterings for a specific value of  $k$  from a hierarchy. Lastly, Algorithm 9 describes how we implement the tie-breaking heuristic assuming Euclidean distance.

---

#### Algorithm 6 $k$ -centroid-annotation

**Input:** An  $n$ -ary dc-tree  $T$  over the dataset  $\mathbf{X}$ , power  $z$ , array  $A$

```

1: if  $|T| = 1$  then
2:    $A[T.id] = \{(T.parent.dist)^z, T.id, nullPtr\}$ 
3:   return 0,  $T.id$ 
4: else
5:    $bestCost = \infty, bestCenter = -1$ 
6:   for  $C \in T.children$  do
7:      $subCost, subCenter = k\text{-centroid-annotation}(C, z)$ 
8:      $currCost = subCost + (T.dist)^z \cdot (T.size - C.size)$ 
9:     if  $currCost < bestCost$  then
10:        $bestCost = currCost, bestCenter = currCenter$ 
11:     end if
12:   end for
13:   for  $C \in T.children$  do
14:      $A[C.id].parent = bestCenter$ 
15:   end for
16:    $costDecrease = (T.parent.dist^z \cdot T.size - bestCost$ 
17:    $A[T.id] = \{costDecrease, bestCenter, nullPtr\}$ 
18:   return  $bestCost, bestCenter$ 
19: end if

```

---

---

**Algorithm 7**  $k$ -centroid-hierarchy

---

**Input:** An  $n$ -ary dc-tree  $T$  over the dataset  $\mathbf{X}$ , power  $z$ 

```

1:  $A = k$ -centroid-annotation ( $T, z, [T.size]$ )
2: Sort( $A$ )
3: Root = Node(parent = nullPtr, cost = -1, id =  $A[0].center$ )
4:  $A[0].tree = Root$ 
5: for  $a_{cur} \in A[1] \dots A[n-1]$  do
6:    $N_{new} = \text{Node}(\text{parent} = \text{nullPtr}, \text{cost} = -1, \text{id} = a_{cur}.\text{center})$ 
7:    $a_{cur}.\text{tree} = N_{new}$ 
8:    $a_{par} = A[a_{cur}.\text{parent}], c_p = a_{par}.\text{tree}.cost$ 
9:   if  $c_p \neq a_{cur}.\text{cost} \wedge c_p \geq 0$  then                                 $\triangleright$  Parent has added children with higher cost
10:     $N_{par} = a_{par}.\text{tree}.\text{children}[0]$ 
11:     $N_{par}.\text{cost} = a_{cur}.\text{costDecrease}$ 
12:     $N_{new}.\text{parent} = N_{par}$ 
13:     $N'_{par} = \text{Node}(\text{parent} = N_{par}, \text{cost} = -1, \text{id} = N_{par}.\text{id})$ 
14:     $N_{par}.\text{children}.add(N'_{par})$                                           $\triangleright$  Add same center node as first child
15:     $N_{par}.\text{children}.add(N_{new})$ 
16:     $a_{par}.\text{tree} = N_{par}$                                                $\triangleright$  Update annotation to point to lowest corresponding node
17:   else if  $c_p < 0$  then                                               $\triangleright$  Parent has no children added
18:      $N_{new}.\text{parent} = a_{par}.\text{tree}$ 
19:      $N'_{par} = \text{Node}(\text{parent} = a_{par}.\text{tree}, \text{cost} = -1, \text{id} = a_{par}.\text{id})$ 
20:      $a_{par}.\text{tree}.\text{children}.add(N'_{par})$ 
21:      $a_{par}.\text{tree}.\text{children}.add(N_{new})$ 
22:   else                                                                $\triangleright$  Parent has children added with same cost
23:      $a_{par}.\text{tree}.\text{children}.add(N_{new})$ 
24:      $N_{new}.\text{parent} = a_{par}.\text{tree}$ 
25:   end if
26: end for
27: return Root

```

---

---

**Algorithm 8**  $k$ -centroid-cluster

---

**Input:** An annotated  $k$ -centroid hierarchy-tree  $T$  over the dataset  $\mathbf{X}$ ,  $searchK$

```

1: if  $|T| = 1$  then
2:   Output  $T$ 
3: end if
4:  $maxK = \max K(T.\text{children})$ ,  $minK = \min K(T.\text{children})$ 
5: if  $maxK \leq searchK$  then                                 $\triangleright$  Every edge should be cut
6:   for  $C \in T.\text{children}$  do
7:      $k$ -centroid-cluster( $C, searchK$ )
8:   end for
9: else                                                  $\triangleright$  Only some edges should be cut
10:    $A = []$ 
11:   for  $C \in T.\text{children}$  do
12:     if  $C.k > searchK \vee C.\text{is\_orig\_cluster}$  then           $\triangleright$  Non-cut edges merge with original
13:        $A.\text{add}(C)$ 
14:     else
15:       Output  $C$ 
16:     end if
17:   end for
18:   Output Merge( $A$ )
19: end if

```

---

**Algorithm 9** Optimize Annotations

---

**Input:** Sorted list of annotations in decreasing order  $A$ , tree  $T$  annotated with representatives

```

1: for  $a_{cur} \in A$  do                                          $\triangleright$  For each annotation
2:    $N = a_{cur}.\text{leaf}$ 
3:   while  $N \neq \text{null}$  do                                      $\triangleright$  Traverse from center leaf to root
4:      $N.\text{mark}(a_{cur}.\text{center})$ 
5:     for  $C \in N.\text{children}$  do
6:       if  $C.\text{mark} = \text{null}$  then                                $\triangleright$  Any unmarked children of marked path
7:          $C.\text{anno}.bestParent = \min\{C.\text{anno}.bestParent, Euclid(C.\text{rep}, Anno.\text{center})\}$ 
8:       end if
9:     end for
10:     $N = N.\text{parent}$ 
11:  end while
12: end for
13: Return  $A$                                                $\triangleright$  Return list of annotations with updated pointers

```

---

## F. Implementations of the used HSTs

We build KD trees, Cover trees using the C++ library *mlpack4* (Curtin et al., 2023), and build HST-DPO trees using the C++ code from <https://github.com/yzengal/ICDE21-HST>.

## G. Used Datasets

Table 4 lists the datasets on which we validated and compared the proposed framework SHIP to other competitors.

## H. Tables – Runtimes, ARI, and NMI

Table 5 shows the runtimes, Table 6 the ARI values, and Table 7 the NMI values for all datasets.

*Table 4.* Dataset properties. Number of samples (*n*), dimensions (*d*), number of ground truth clusters (*k*), number of noise points (#noise), DISCO score for the ground truth labels (DISCO) including noise, and the source.

	Dataset	<i>n</i>	<i>d</i>	<i>k</i>	#noise	Source
Density-based 2D-Data Tomas Barton Benchmark	Boxes	21,600	2	12	0	(Yoon, 2023)
	D31	3100	2	31	0	(Barton, 2019)
	3-spiral	312	2	3	0	(Barton, 2019)
	aggregation	788	2	7	0	(Barton, 2019)
	chainlink	1,000	3	2	0	(Barton, 2019)
	cluto-t4-8k	8,000	2	6	764	(Barton, 2019)
	cluto-t5-8k	8,000	2	7	1,153	(Barton, 2019)
	cluto-t7-10k	10,000	2	9	792	(Barton, 2019)
	cluto-t8-8k	8,000	2	8	323	(Barton, 2019)
	complex8	2,551	2	8	0	(Barton, 2019)
	complex9	3,031	2	9	0	(Barton, 2019)
	compound	399	2	6	0	(Barton, 2019)
	dartboard1	1,000	2	4	0	(Barton, 2019)
	diamondd9	3,000	2	9	0	(Barton, 2019)
	jain	3,373	2	2	0	(Barton, 2019)
	pathbased	299	2	3	0	(Barton, 2019)
	smile1	1,000	2	4	0	(Barton, 2019)
Real-World Data Tabular Data	Synth_low	5,000	100	10	500	(Jahn et al., 2024)
	Synth_high	5,000	100	10	500	(Jahn et al., 2024)
	Mice	1,077	68	8	0	(Markelle Kelly, 2023)
	adipose	14,947	2	12	0	(Broad Institute, 2025)
	airway	14,163	2	10	0	(Broad Institute, 2025)
	lactate	39,825	2	6	0	(Broad Institute, 2025)
	HAR	10,299	561	6	0	(Markelle Kelly, 2023)
	letterrec.	20,000	16	26	0	(Markelle Kelly, 2023)
	Pendigits	10,992	16	10	0	(Markelle Kelly, 2023)
	COIL20	1,440	16,384	20	0	(Nene et al., 1996b)
Image Data	COIL100	7,200	49,152	100	0	(Nene et al., 1996a)
	cmu_faces	624	960	20	0	(Markelle Kelly, 2023)
	Optdigits	5,620	64	10	0	(Markelle Kelly, 2023)
	USPS	9,298	256	10	0	(Hull, 1994)
	MNIST	70,000	784	10	0	(LeCun et al., 1998)

**Table 5.** Runtimes of various parts of our clustering framework SHIP over ten runs, compared with standard clustering methods. Time is given in minutes, seconds, and milliseconds [min.sec.ms]. The first four columns (“build”) refer to the computation time of the different ultrametric trees. The next two columns state the runtimes of computing one hierarchy ( $k$ -means) on the DC tree. The second column here shows the timing for the heuristic from Section D.1.2. Note that these runtimes are not correlated to the chosen ultrametric. The subsequent three columns state the runtimes of different partitioning methods, i.e., ‘Stability’, ‘MoE’, and the ‘Elbow’ method on various hierarchies.

Dataset	build ultrametric				build hierarchy			partitioning			competitors				
	KD tree	Cover tree	HST-DPO	DC tree	$k$ -means (heut.)	Stability	MoE	Elbow	$k$ -means ( $k = GT$ )	$k$ -means ( $k = 500$ )	SCAR	Ward	AMD-DBSCAN	DPC	
Boxes	00:00:013	00:00:059	00:09:679	00:14:239	00:00:033	00:00:498	00:00:005	00:00:134	00:00:002	00:00:114	00:01:217	00:01:481	00:01:0808	01:04:070	12:10:358
D31	00:00:002	00:00:004	00:00:207	00:00:327	00:00:004	00:00:020	00:00:000	00:00:015	00:00:000	00:00:289	00:00:509	00:00:620	00:00:153	00:00:878	00:12:816
3-spiral	00:00:000	00:00:000	00:00:002	00:00:009	00:00:000	00:00:001	00:00:000	00:00:000	00:00:095	00:00:131	00:00:023	00:00:002	00:00:025	00:00:091	
aggregation	00:00:000	00:00:000	00:00:024	00:00:035	00:00:001	00:00:004	00:00:000	00:00:004	00:00:000	00:00:227	00:00:204	00:00:038	00:00:010	00:00:121	00:00:663
chainlink	00:00:000	00:00:001	00:00:042	00:00:045	00:00:001	00:00:008	00:00:000	00:00:004	00:00:000	00:00:133	00:00:171	00:00:049	00:00:012	00:00:159	00:01:128
cluto-t4-8k	00:00:004	00:00:017	00:01:359	00:01:898	00:00:012	00:00:164	00:00:001	00:00:046	00:00:000	00:01:344	00:01:514	00:00:417	00:01:233	00:06:470	01:32:488
cluto-t5-8k	00:00:005	00:00:013	00:01:319	00:01:836	00:00:010	00:00:198	00:00:001	00:00:042	00:00:000	00:00:682	00:01:758	00:00:378	00:01:256	00:05:942	01:33:225
cluto-t7-10k	00:00:006	00:00:019	00:02:100	00:02:848	00:00:014	00:00:252	00:00:002	00:00:058	00:00:000	00:02:532	00:02:052	00:00:455	00:02:058	00:10:670	02:39:063
cluto-t8-8k	00:00:005	00:00:014	00:01:328	00:01:728	00:00:012	00:00:158	00:00:001	00:00:046	00:00:000	00:01:930	00:01:899	00:00:327	00:01:394	00:09:212	01:39:524
complex8	00:00:001	00:00:003	00:01:129	00:00:244	00:00:003	00:00:011	00:00:000	00:00:012	00:00:000	00:00:565	00:00:592	00:00:221	00:00:120	00:01:295	00:10:900
complex9	00:00:001	00:00:004	00:01:185	00:00:304	00:00:003	00:00:014	00:00:000	00:00:012	00:00:000	00:00:745	00:00:682	00:00:175	00:00:158	00:01:376	00:12:483
compound	00:00:000	00:00:000	00:00:004	00:00:004	00:00:011	00:00:000	00:00:000	00:00:000	00:00:000	00:00:066	00:00:127	00:00:036	00:00:003	00:00:031	00:00:204
dartboard1	00:00:000	00:00:002	00:00:027	00:00:051	00:00:001	00:00:003	00:00:000	00:00:004	00:00:000	00:00:210	00:00:270	00:00:033	00:00:016	00:00:138	00:01:346
diamond9	00:00:001	00:00:004	00:00:183	00:00:310	00:00:003	00:00:017	00:00:000	00:00:012	00:00:000	00:00:751	00:00:194	00:00:153	00:01:107	00:12:540	
jain	00:00:000	00:00:000	00:00:004	00:00:006	00:00:000	00:00:001	00:00:000	00:00:000	00:00:005	00:00:045	00:00:024	00:00:002	00:00:028	00:00:101	
pathbased	00:00:000	00:00:002	00:00:024	00:00:036	00:00:001	00:00:006	00:00:000	00:00:000	00:00:000	00:00:013	00:00:039	00:00:023	00:00:001	00:00:024	00:00:061
smile1	00:00:000	00:00:002	00:00:024	00:00:036	00:00:001	00:00:005	00:00:000	00:00:004	00:00:000	00:00:321	00:00:319	00:00:048	00:00:017	00:00:139	00:01:392
Synth_low	00:00:040	00:00:087	00:04:264	00:01:090	00:00:007	00:00:128	00:00:001	00:00:027	00:00:000	00:00:351	00:03:754	00:00:367	00:00:999	00:02:879	00:33:945
Synth_high	00:00:034	00:00:030	00:04:075	00:00:878	00:00:006	00:00:117	00:00:001	00:00:024	00:00:000	00:00:782	00:04:199	00:00:455	00:00:982	00:04:573	00:33:785
Mice	00:00:003	00:00:006	00:01:126	00:00:052	00:00:001	00:00:013	00:00:000	00:00:004	00:00:000	00:00:173	00:00:876	00:00:054	00:00:073	00:00:482	00:01:423
airway	00:00:008	00:00:027	00:04:054	00:04:997	00:00:021	00:00:592	00:00:003	00:00:085	00:00:001	00:00:122	00:00:392	00:00:651	00:05:243	00:16:822	06:46:726
lactate	00:00:025	00:01:161	00:39:433	00:47:731	00:00:068	00:07:184	00:00:012	00:00:275	00:00:004	00:00:126	00:01:997	00:02:612	00:59:992	07:43:293	69:34:012
HAR	00:00:341	00:11:053	01:55:260	00:23:144	00:00:014	00:01:936	00:00:002	00:00:056	00:00:000	00:02:248	00:08:478	00:06:658	00:18:448	00:30:129	03:45:745
letterrc.	00:00:024	00:00:322	00:14:396	00:09:076	00:00:022	00:01:103	00:00:002	00:00:087	00:00:001	00:03:386	00:03:521	00:02:133	00:10:309	00:37:055	11:42:910
PenDigits	00:00:015	00:00:067	00:04:625	00:02:566	00:00:013	00:00:416	00:00:002	00:00:053	00:00:000	00:01:001	00:02:725	00:00:490	00:03:281	00:09:254	03:21:409
COIL20	00:01:026	00:03:658	01:04:304	00:14:817	00:00:001	00:00:379	00:00:000	00:00:004	00:00:000	00:01:637	00:13:524	00:00:310	00:08:941	00:02:378	00:11:710
COIL100	00:36:361	03:52:397	82:50:274	14:50:661	00:00:010	00:22:517	00:00:001	00:00:039	00:00:000	00:32:122	01:59:225	00:07:964	12:04:037	02:58:177	14:18:780
cmu faces	00:00:021	00:00:046	00:00:682	00:00:238	00:00:001	00:00:010	00:00:000	00:00:002	00:00:000	00:00:206	00:00:661	00:00:116	00:00:064	00:03:346	00:00:515
OptDigits	00:00:019	00:00:335	00:03:442	00:01:420	00:00:006	00:00:162	00:00:001	00:00:024	00:00:000	00:00:502	00:01:154	00:00:290	00:00:974	00:03:073	00:44:270
USPS	00:00:117	00:02:924	00:45:038	00:08:670	00:00:011	00:01:843	00:00:002	00:00:045	00:00:000	00:02:709	00:08:493	00:01:433	00:06:092	00:29:259	01:48:607
MNIST	00:10:649	16:24:220	131:03:099	37:05:95	00:00:120	03:21:012	00:00:030	00:00:491	00:00:008	00:04:320	03:29:969	00:15:352	19:02:498	17:21:183	-

**Table 6.** ARI values for various combinations within our clustering framework SHIP over ten runs, compared with standard clustering methods. Noise points were handled as singleton clusters. This shows that the results yield high-quality clusterings when using the DC tree as an ultrametric.

Dataset	DC tree				HST-DPO				Cover tree				KD tree				competitors			
	k-center		k-median		k-center		k-median		k-center		k-median		k-means		Eucl.-k-means		SCAR		Ward	
	GT	GT	GT	GT	GT	GT	MoE	Elbow	GT	MoE	Elbow	GT	MoE	Elbow	GT	MoE	Elbow	GT	MoE	Elbow
Boxes	06.5	<b>99.3</b>	<b>99.3</b>	90.1	<b>99.3</b>	<b>97.9</b>	2.9 ± 0.1	<b>45.5 ± 7.8</b>	35.5 ± 9.3	2.6	<b>42.4 ± 4.7</b>	24.2 ± 1.6	2.3	36.9	21.7	<b>93.5 ± 4.3</b>	0.1 ± 0.1	<b>95.8</b>	63.9	25.9
D31	65.9	<b>93.7</b>	<b>93.7</b>	73.7	<b>42.7</b>	82.9	41.4 ± 11.7	<b>33.9 ± 11.1</b>	46.5 ± 1.8	<b>62.0 ± 5.4</b>	<b>67.7 ± 3.2</b>	38.7	59.2	74.9	<b>92.0 ± 2.7</b>	41.7 ± 5.4	<b>92.0</b>	86.4 ± 0.1	18.5	
3-spiral	55.2	<b>98.1</b>	<b>98.1</b>	66.6	<b>66.6</b>	58.4 ± 4.2	6.2 ± 3.0	13.8 ± 2.0	14.7 ± 4.1	16.3 ± 1.9	9.6	5.9	7.6	-0.6	-0.3 ± 0.1	0.0 ± 0.1	30.6	<b>97.4 ± 7.8</b>	73.4	-
aggregation	99.2	<b>99.2</b>	80.9	80.9	<b>99.2</b>	<b>100.0</b>	24.8 ± 2.7	52.1 ± 13.3	<b>51.7 ± 12.5</b>	18.6 ± 1.6	<b>46.4 ± 1.3</b>	43.8	29.8	<b>74.7 ± 1.6</b>	20.4 ± 6.3	<b>80.6 ± 0.9</b>	<b>97.8</b>	-	-	
chainlink	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	8.0 ± 5.1	12.9 ± 5.6	6.1 ± 0.4	12.1 ± 2.9	8.2 ± 0.5	5.2	7.2	7.0	9.5 ± 0.4	2.4 ± 1.5	<b>28.0</b>	12.4	8.1	
cluto+4-8k	0.4	79.8	<b>87.0</b>	87.2	75.5	61.5	3.4 ± 0.3	32.2 ± 6.1	27.6 ± 8.4	2.8 ± 0.1	22.9 ± 2.1	19.3 ± 3.6	3.1	23.7	14.1	<b>45.4 ± 2.0</b>	0.6 ± 0.5	49.8	80.1	43.6
cluto+5-8k	0.1	74.8	77.9	80.8	65.0	66.7	3.9 ± 0.3	45.2 ± 10.1	4.3 ± 0.1	42.0 ± 7.5	29.0 ± 6.6	3.2	37.5	22.3	74.2 ± 0.3	0.3 ± 0.2	<b>73.9</b>	<b>87.4</b>	42.2 ± 2.7	
cluto+7-10k	1.0	68.8	<b>88.9</b>	71.4	36.9	28.0	1.1 ± 0.1	25.1 ± 4.3	19.4 ± 4.4	2.8 ± 0.2	18.9 ± 2.4	13.5 ± 0.5	2.3	17.8	10.9	33.5 ± 1.5	-0.1 ± 0.9	41.2	<b>87.8</b>	31.0
cluto+8-8k	63.4	79.9	<b>80.1</b>	63.4	72.4	36.1	3.8 ± 0.3	26.0 ± 4.5	3.0 ± 0.4	3.9 ± 0.2	21.3 ± 3.7	15.5 ± 0.8	3.0	23.0	13.4	35.3 ± 1.7	0.3 ± 0.2	33.5	<b>76.9 ± 0.1</b>	30.2
complex8	90.9	90.2	<b>99.0</b>	<b>91.2</b>	33.2	28.0	9.6 ± 1.4	28.8 ± 7.9	27.9 ± 5.3	9.9 ± 0.7	28.7 ± 1.7	23.0 ± 3.4	8.8	27.6	24.3	37.2 ± 1.6	4.2 ± 1.6	36.7	53.3	29.7
complex9	<b>93.1</b>	<b>83.3</b>	83.3	93.1	<b>93.1</b>	61.4	8.0 ± 0.9	26.7 ± 6.8	24.3 ± 4.5	7.5 ± 0.7	25.2 ± 2.5	17.8 ± 3.4	7.2	20.5	16.6	37.6 ± 2.2	-1.6 ± 1.9	42.2	<b>81.3 ± 0.3</b>	21.2
compound	78.4	65.2	80.7	74.0	85.3	44.0	6.1 ± 6.8	44.0 ± 9.4	42.8 ± 6.0	32.1 ± 5.0	44.0 ± 7.4	47.3 ± 4.8	30.6	52.4	38.6	<b>53.4 ± 0.9</b>	27.7 ± 7.2	<b>53.4</b>	59.1	59.1
dartboard	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	5.3 ± 1.3	87.6 ± 9.0	76.8 ± 9.0	16.2 ± 0.4	16.2 ± 1.9	17.6 ± 0.1	8.9	14.3	13.9	2.8 ± 0.2	14.8 ± 1.5	0.0	<b>0.8 ± 0.1</b>	-
diamond9	08.6	<b>99.7</b>	<b>83.6</b>	70.4	45.1	22.6 ± 8.2	41.9 ± 12.4	44.4 ± 13.1	14.6 ± 0.6	58.8 ± 7.7	45.9 ± 4.1	11.0	44.3	30.0	<b>98.1 ± 5.6</b>	5.2 ± 1.4	<b>99.6</b>	<b>96.6</b>	15.1	
join	1.0	<b>100.0</b>	<b>100.0</b>	<b>92.5</b>	50.7	46.3	9.0 ± 2.5	18.5 ± 10.7	14.5 ± 4.0	7.5 ± 1.9	14.7 ± 2.1	15.1 ± 2.3	7.7	17.4	11.7	31.5 ± 1.2	-5.1 ± 5.7	51.5	<b>83.4</b>	<b>3.2 ± 0.1</b>
pathbased	51.6	54.8	54.8	<b>58.7</b>	54.8	58.3	22.9 ± 4.5	26.6 ± 7.8	35.0 ± 7.0	21.1 ± 5.9	29.1 ± 6.8	26.3 ± 10.0	17.7	24.9	25.4	46.1 ± 0.1	27.6 ± 9.3	48.3	<b>45.7 ± 0.8</b>	38.8
smile1	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	13.3 ± 15.5	28.9 ± 29.0	30.0 ± 31.0	15.5 ± 1.3	66.7 ± 1.0	67.6 ± 1.2	15.3	63.0	66.1	54.6	7.3 ± 1.9	55.5 ± 0.1	29.9	<b>32.5 ± 0.7</b>
Synth_low	94.2	84.0	84.0	<b>95.4</b>	35.9	37.2	34.8 ± 16.4	54.7 ± 14.6	55.2 ± 8.7	13.0 ± 2.3	<b>55.2 ± 8.7</b>	67.6 ± 3.4	12.8	<b>49.2</b>	80.0	<b>55.9 ± 13.2</b>	1.4 ± 1.0	<b>62.5</b>	<b>32.4</b>	
Synth_high	94.1	84.6	84.6	<b>94.4</b>	23.5	70.5	27.3 ± 20.2	36.4 ± 15.0	53.8 ± 10.3	13.0 ± 0.4	<b>24.8 ± 1.6</b>	67.6 ± 3.4	8.0	35.0	46.3	<b>47.9 ± 8.2</b>	2.9 ± 0.9	<b>71.7</b>	0.0	
Mice	1.1	<b>26.7</b>	20.0	0.2	11.6	11.9	10.7 ± 2.0	10.0 ± 2.0	10.8 ± 1.4	9.7 ± 2.7	13.5 ± 1.9	11.8 ± 1.1	8.6	11.0	9.9	14.5 ± 1.1	12.7 ± 3.1	16.4	0.0	3.0
airway	23.5	58.1	58.9	38.0	<b>65.9</b>	58.8	0.9 ± 0.1	28.8 ± 11.9	20.5 ± 8.1	0.8	18.2 ± 2.4	12.0 ± 1.4	0.7	16.9	8.7	39.9 ± 2.0	-0.9 ± 0.5	43.7	<b>31.7</b>	<b>65.1</b>
lactate	4.9	61.7	62.1	41.0	67.5	1.9 ± 1.8	3.6 ± 1.5	3.5 ± 1.0	0.1	4.1 ± 0.6	1.7 ± 0.2	0.1	4.4	2.3	28.6 ± 1.1	1.5 ± 1.0	<b>27.7</b>	<b>71.5</b>	0.0	
HAR	0.7	47.3	47.0	30.0	46.9	<b>52.8</b>	35.1 ± 6.0	31.3 ± 8.5	14.7 ± 8.8	14.2 ± 4.7	9.6 ± 2.2	1.5	2.2	46.0 ± 4.5	5.5 ± 3.2	<b>49.1</b>	0.0	<b>33.2</b>	-	
letterrec.	-0.0	9.6	8.9	12.1	16.6	21.9	4.9 ± 0.5	7.9 ± 0.9	7.6 ± 0.9	5.8 ± 0.6	7.2 ± 0.6	6.2 ± 0.3	3.3	5.4	4.3	12.9 ± 0.6	0.4 ± 0.1	14.7 ± 0.9	7.9	-0.0
PenDigits	0.1	68.3	70.1	66.4	<b>73.1</b>	<b>75.4</b>	8.9 ± 1.5	33.9 ± 7.6	28.6 ± 6.5	8.0 ± 0.8	12.1 ± 0.6	8.9 ± 0.5	4.5	6.2	4.5	56.3 ± 3.2	0.9 ± 0.3	55.2	55.6	<b>28.8 ± 1.1</b>
COIL20	58.4	72.8	<b>75.5</b>	<b>81.2</b>	72.8	72.6	39.9 ± 2.4	38.4 ± 9.0	44.2 ± 4.8	46.4 ± 4.4	46.6 ± 2.1	47.7 ± 2.0	22.5	16.9	18.5	58.2 ± 2.8	33.5 ± 2.0	<b>68.6</b>	39.2	<b>35.9 ± 0.1</b>
COIL100	19.6	<b>70.2</b>	69.9	<b>80.1</b>	66.8	70.0	40.6 ± 4.7	34.3 ± 12.6	43.2 ± 6.2	44.6 ± 4.2	46.6 ± 1.5	50.1 ± 1.2	23.6	22.5	24.4	56.1 ± 1.4	16.7 ± 0.8	61.4	<b>14.2</b>	0.2
cmu_faces	24.2	60.2	61.4	56.6	<b>66.5</b>	71.± 3.2	16.1 ± 10.2	26.6 ± 20.4	8.6 ± 3.1	37.± 4.1	34.2 ± 2.1	11.3	12.2	12.2	53.2 ± 4.7	38.5 ± 2.9	<b>61.6</b>	0.7	0.6	
OptDigits	5.0	74.8	74.8	55.3	<b>77.0</b>	<b>83.4</b>	36.4 ± 4.9	34.3 ± 7.7	34.1 ± 5.4	40.9 ± 3.5	20.9 ± 2.3	18.1 ± 2.4	4.1	2.9	2.9	61.3 ± 6.6	14.4 ± 4.1	<b>74.6 ± 2.4</b>	<b>63.2</b>	0.0
USPS	1.8	55.8	54.0	33.7	29.3	15.0	1.5 ± 2.0	26.8 ± 4.3	23.9 ± 3.5	12.0 ± 1.7	8.7 ± 1.0	11.2 ± 1.5	3.0	2.7	3.9	52.3 ± 1.7	2.9 ± 0.9	<b>63.9</b>	0.0	<b>21.0</b>
MNIST	0.3	50.1	<b>51.0</b>	19.7	41.7	46.0	12.3 ± 2.8	7.3 ± 2.3	7.6 ± 2.0	11.1 ± 1.7	5.4 ± 0.6	5.4 ± 0.6	1.4	0.2	0.3	36.9 ± 1.0	1.3 ± 0.4	<b>52.7</b>	0.0	-

**Table 7.** NMI values for various combinations within our clustering framework SHIP over ten runs, compared with standard clustering methods. Noise points were handled as singleton clusters. This shows that the results yield high-quality clusterings when using the DC tree as an ultrametric.

Dataset	DC tree			HST-DPO			Cover tree			KD tree			competitors									
	$k$ -center	$k$ -means		$k$ -center	$k$ -median		$k$ -means	$k$ -median		$k$ -means	$k$ -center		$k$ -means	$k$ -means		$k$ -means						
		GT	Stability		MoE	Elbow		MoE	Elbow		MoE	Elbow		MoE	Elbow	MoE	Elbow					
Boxes	86.6	99.2	99.2	95.0	99.2	98.3	50.6 ± 0.3	74.0 ± 1.9	71.5 ± 2.7	51.1 ± 0.2	75.0 ± 1.5	69.1 ± 0.5	52.7	73.1	68.0	95.3 ± 1.9	8.8 ± 0.8	97.2	65.6	50.3		
	83.2	95.8	95.8	95.8	87.7	81.8	94.0	75.7 ± 2.4	69.7 ± 6.4	76.4 ± 4.1	78.6 ± 0.5	84.4 ± 2.1	86.1 ± 1.1	79.1	82.3	87.6	95.7 ± 0.9	81.5 ± 1.8	95.1	90.7	60.0	
3-spiral aggregation	71.0	96.7	96.7	97.6	96.7	96.7	24.0 ± 3.6	13.2 ± 8.6	17.0 ± 0.5	35.7 ± 4.1	39.1 ± 7.7	44.8 ± 2.9	26.7	17.0	26.7	0.1	0.2 ± 0.1	0.6 ± 0.1	45.2	95.9 ± 12.4	-	
	86.6	98.8	98.8	98.8	88.9	88.9	59.6 ± 2.2	68.0 ± 7.7	71.3 ± 5.8	59.2 ± 1.3	76.0 ± 2.5	74.9 ± 3.2	62.1	73.8	67.8	85.0 ± 1.5	53.6 ± 4.5	91.9 ± 0.2	96.7	83.6	-	
chainlink	100.0	100.0	100.0	100.0	100.0	100.0	74.0	22.6 ± 9.4	23.2 ± 10.5	25.2 ± 11.8	31.5 ± 0.8	38.9 ± 2.6	35.3 ± 0.5	31.3	27.5	33.5	7.1 ± 0.3	11.4 ± 2.9	36.6	34.7	26.6 ± 0.2	-
cluto-44-8k	2.6	83.6	87.1	81.5	81.0	76.4	44.2 ± 0.3	51.3 ± 2.8	53.3 ± 2.5	44.3 ± 0.2	56.4 ± 1.0	56.0 ± 1.5	46.4	56.3	53.2	58.5 ± 0.4	10.0 ± 1.2	62.8	76.8	48.5	-	
cluto-45-8k	4.7	79.9	82.1	78.3	75.4	74.3	45.0 ± 0.4	65.2 ± 3.2	63.9 ± 2.8	46.8 ± 0.3	63.6 ± 2.9	59.1 ± 2.4	46.2	59.3	56.6	79.6 ± 0.2	10.9 ± 0.8	79.7	77.1	55.7 ± 3.4	-	
cluto-7-10k	5.0	80.5	80.4	83.7	82.4	69.7	45.8 ± 0.4	49.8 ± 3.1	52.0 ± 1.3	47.0 ± 0.4	53.7 ± 1.4	55.3 ± 0.7	47.5	53.7	53.8	57.2 ± 0.7	11.6 ± 0.9	64.1	82.6	43.6	-	
cluto-48-8k	77.3	85.7	85.6	79.3	86.0	80.2	45.9 ± 0.3	50.9 ± 4.5	52.2 ± 3.4	47.9 ± 0.5	54.0 ± 1.6	55.7 ± 0.6	47.6	57.7	54.4	56.5 ± 1.7	12.8 ± 1.2	55.3	70.0 ± 0.1	43.1	-	
complex8	93.8	94.4	98.8	91.3	70.9	68.3	50.8 ± 0.8	50.7 ± 8.3	55.7 ± 1.0	53.2 ± 1.1	58.4 ± 1.1	59.4 ± 1.5	55.1	59.4	61.6	59.4 ± 1.1	28.6 ± 2.6	58.1	65.3 ± 0.2	46.0 ± 0.1	-	
complex9	96.8	93.0	93.0	96.9	96.9	85.4	51.6 ± 0.9	53.1 ± 7.2	56.8 ± 3.3	52.9 ± 1.0	61.9 ± 2.7	62.2 ± 1.3	55.5	55.2	56.5	63.5 ± 1.8	23.6 ± 2.2	70.5	89.5 ± 0.3	41.1	-	
compound	74.5	84.0	84.0	86.4	80.6	91.2	56.4 ± 3.0	58.3 ± 7.5	59.6 ± 5.0	62.6 ± 2.8	68.5 ± 2.1	68.8 ± 2.2	61.3	70.2	65.2	71.8 ± 0.2	51.6 ± 6.3	73.3	77.0	72.9	-	
databoard1	100.0	100.0	100.0	100.0	100.0	100.0	38.5 ± 5.1	26.4 ± 8.3	30.3 ± 10.6	49.8 ± 6.6	42.5 ± 6.4	53.0 ± 1.8	42.2	35.2	37.9	3.5 ± 10.6	37.1 ± 5.8	5.7 ± 1.9	0.0	34.0 ± 0.1	-	
diamond9	86.8	99.6	99.6	83.3	85.7	76.7	58.3 ± 1.1	64.1 ± 8.4	67.9 ± 5.6	61.6 ± 2.6	80.1 ± 2.6	76.0 ± 1.2	59.5	70.2	68.4	99.0 ± 3.1	37.1 ± 3.6	99.6	96.3	-	-	
jain	1.2	100.0	100.0	100.0	78.7	69.7	63.9	28.4 ± 2.5	32.5 ± 6.9	32.4 ± 3.4	30.2 ± 2.8	39.0 ± 3.0	32.2	37.8	36.5	36.4 ± 0.7	7.4 ± 5.1	50.5	49.3	20.2 ± 0.2	-	
pathbased	62.2	66.0	66.0	66.0	66.0	66.4	42.4 ± 3.7	39.5 ± 7.2	46.1 ± 6.0	44.2 ± 3.6	51.9 ± 4.2	50.6 ± 4.9	41.4	44.0	46.1	54.6 ± 0.1	37.3 ± 9.1	43.6	43.1 ± 0.7	43.6	-	
smile1	100.0	100.0	100.0	100.0	100.0	100.0	12.9 ± 9.7	42.5 ± 22.5	44.1 ± 24.2	50.4 ± 1.2	73.0 ± 1.1	73.7 ± 1.0	53.0	66.7	69.7	60.8	26.5 ± 2.8	61.2 ± 0.1	43.5	49.4 ± 4.2	-	
Synth-Low	86.7	89.2	89.2	87.5	59.4	63.2	63.8 ± 8.1	72.5 ± 10.5	72.7 ± 10.0	57.9 ± 1.1	76.2 ± 3.7	79.8 ± 1.6	59.8	76.6	80.7	77.5 ± 7.0	24.0 ± 1.3	80.7	51.8 ± 0.5	0.0	-	
Synth-high	86.6	87.1	87.1	87.5	43.2	81.1	61.6 ± 6.3	64.2 ± 14.2	70.6 ± 1.5	60.9 ± 0.2	66.3 ± 0.6	76.3 ± 0.7	55.5	70.0	69.3	71.4 ± 5.1	21.4 ± 1.1	84.1	0.0	-	-	
Mice	21.2	43.0	39.5	8.0	29.5	25.6	29.6 ± 2.8	19.6 ± 3.6	23.4 ± 2.2	42.2 ± 2.5	36.2 ± 3.6	45.1 ± 2.9	32.5	28.3	25.7	25.7 ± 1.3	29.2 ± 2.2	27.9	0.0	44.9	-	
airway	46.9	68.3	68.3	54.5	70.7	68.3	32.6 ± 0.2	56.4 ± 4.3	53.0 ± 3.8	33.0 ± 0.1	53.6 ± 1.2	48.9 ± 0.7	33.8	52.2	47.3	63.7 ± 1.0	6.7 ± 1.0	63.9	40.9	57.7	-	
lactate	11.8	62.3	62.3	51.9	53.5	63.4	28.7 ± 4.5	38.2 ± 1.5	35.0 ± 1.4	24.2 ± 0.4	37.6 ± 0.8	33.4 ± 0.6	25.0	38.0	35.0	52.2 ± 1.6	3.4 ± 1.0	52.0	38.0	0.0	-	
HAR	11.8	65.3	63.9	43.7	65.9	63.8	40.9 ± 1.0	51.9 ± 5.3	52.1 ± 4.2	39.6 ± 1.9	45.8 ± 1.4	44.3 ± 0.7	29.3	29.6	29.3	60.2 ± 2.2	24.6 ± 6.3	62.9	0.0	55.5	-	
letterrec.	1.4	42.8	44.0	56.6	53.6	54.5	42.2 ± 2.2	33.0 ± 3.9	36.3 ± 4.1	51.9 ± 0.9	51.1 ± 1.3	53.2 ± 0.5	45.2	40.5	43.6	35.3 ± 0.5	19.4 ± 1.6	40.3 ± 0.9	49.3	49.7	-	
PenDigits	7.5	77.4	78.6	70.3	81.7	82.8	50.6 ± 0.6	57.6 ± 2.7	59.3 ± 1.6	51.5 ± 0.5	57.8 ± 0.4	56.2 ± 0.3	49.1	50.2	49.6	68.0 ± 1.0	18.4 ± 1.7	72.7	59.9	46.0 ± 0.6	-	
COIL20	80.7	86.9	88.0	89.1	86.9	86.8	68.8 ± 1.1	67.4 ± 8.4	70.9 ± 3.4	71.6 ± 1.5	75.9 ± 0.8	76.1 ± 0.8	50.0	51.7	53.1	77.8 ± 1.9	64.6 ± 0.7	82.6	65.9	64.5 ± 0.1	-	
COIL100	79.7	89.5	89.6	90.7	88.9	89.6	76.7 ± 0.9	73.5 ± 7.4	75.1 ± 3.3	78.3 ± 0.5	82.2 ± 0.3	82.9 ± 0.3	61.3	62.9	64.1	83.2 ± 0.6	73.1 ± 0.4	86.2	72.1	68.3	-	
cmu_faces	71.2	84.0	85.4	78.9	81.5	87.3	31.5 ± 6.6	43.3 ± 15.2	52.2 ± 23.2	34.2 ± 5.1	78.6 ± 0.8	78.0 ± 0.5	43.4	44.2	45.0	75.7 ± 2.5	64.9 ± 1.6	81.5	22.6	63.5	-	
OptDigits	29.5	83.0	83.0	62.3	84.5	84.5	61.2 ± 1.3	62.2 ± 2.6	61.9 ± 1.6	62.7 ± 1.5	58.8 ± 0.9	57.7 ± 0.9	37.6	38.6	39.3	72.2 ± 3.5	50.1 ± 4.3	83.4 ± 1.2	66.8	42.1	-	
Image USPS	21.0	67.6	67.3	46.9	52.9	52.9	47.0 ± 0.8	47.2 ± 1.9	47.3 ± 0.9	50.1 ± 0.5	50.7 ± 0.5	50.1 ± 0.5	36.0	36.5	37.5	60.8 ± 1.0	31.2 ± 1.9	76.1	0.0	42.0	-	
MNIST	14.8	61.6	62.9	38.2	58.3	62.6	39.1 ± 1.6	39.7 ± 0.5	40.0 ± 0.6	40.9 ± 0.8	41.9 ± 0.4	42.2 ± 0.4	25.6	25.0	25.0	49.1 ± 0.7	19.2 ± 2.7	68.2	0.0	-	-	