Anytime Hierarchical Clustering

Omur Arslan
Department of Electrical and Systems
Engineering, University of Pennsylvania,
Philadelphia, PA 19104
omur@seas.upenn.edu

Daniel E. Koditschek
Department of Electrical and Systems
Engineering, University of Pennsylvania,
Philadelphia, PA 19104
kod@seas.upenn.edu

ABSTRACT

We propose a new anytime hierarchical clustering method that iteratively transforms an arbitrary initial hierarchy on the configuration of measurements along a sequence of trees we prove for a fixed data set must terminate in a chain of nested partitions that satisfies a natural homogeneity requirement. Each recursive step re-edits the tree so as to improve a local measure of cluster homogeneity that is compatible with a number of commonly used (e.g., single, average, complete) linkage functions. As an alternative to the standard batch algorithms, we present numerical evidence to suggest that appropriate adaptations of this method can yield decentralized, scalable algorithms suitable for distributed /parallel computation of clustering hierarchies and online tracking of clustering trees applicable to large, dynamically changing databases and anomaly detection.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*; I.5.3 [**Pattern Recognition**]: Clustering

General Terms

Theory, Algorithms

Keywords

Online Clustering, Homogeneity, Anytime Algorithm, Cluster Tracking, Nearest Neighbor Interchange, Big Data

1. INTRODUCTION

The explosive growth of data sets in recent years is fueling a search for efficient and effective knowledge discovery tools. Cluster analysis [5, 20, 22] is a nearly ubiquitous tool for unsupervised learning aimed at discovering unknown groupings within a given set of unlabelled data points (objects, patterns), presuming that objects in the same group (cluster) are more similar to each other (intra-cluster homogeneity) than objects in other groups (inter-cluster separability). Amongst the many alternative methods, this paper focuses on dissimilarity-based hierarchical clustering as represented by a tree that indexes a chain of successively more finely nested partitions of the dataset. We are motivated to explore this approach to knowledge discovery because clustering can be imposed on arbitrary data types, and hierarchy can relieve the need to commit a priori to a specific partition cardinality or granularity [5]. However, we are gener-

ally interested in online or reactive problem settings and in this regard hierarchical clustering suffers a number of long discussed weaknesses that become particularly acute when scaling up to large, and, particularly, dynamically changing, data sets [5, 20]. Construction of a clustering hierarchy (tree) generally requires $O(n^2)$ time with the number of data points n [32]. Moreover, whenever a data set is changed by insertion, deletion or update of a single data point, a clustering tree must generally be reconstructed in its entirety. This paper addresses the problem of anytime online reclustering.

1.1 Contributions of The Paper

We introduce a new homogeneity criterion applicable to an array of agglomerative ("bottom up") clustering methods through a test involving their "linkage function" — the mechanism by which dissimilarity at the level of individual data entries is "lifted" to the level of the clusters they are assigned. That criterion motivates a "homogenizing" local adjustment of the nesting relationship between proximal clusters in the hierarchy that increases the degree of similitude within them while increasing the dissimilarity between them. We show that iterated application of this local homogenizing operation transforms any initial cluster hierarchy through a succession of increasingly "better sorted" ones along a path in the abstract space of hierarchies that we prove, for a fixed data set and with respect to a family of linkages including the common single, average and complete cases, must converge in a finite number of steps. In particular, for the single linkage function, we prove convergence from any initial condition of any sequence of arbitrarily chosen local homogenizing reassignments to the generically unique[1], globally homogeneous hierarchy that would emerge from application of the standard, one-step "batch" single linkage based agglomerative clustering procedure.

We present evidence to suggest that decentralized algorithms based upon this homogenizing transformation can scale effectively for anytime online hierarchical clustering of large and dynamically changing data sets. Each local homogenizing adjustment entails computation over a proper subset of the entire dataset — and, for some linkages, merely its sufficient statistics (e.g. mean, variance). In these circumstances, given the sufficient statistics of a dataset, such a restructuring decision at a node of a clustering hierarchy can be made in constant time (for further discussion see Section 4.2). Recursively defined ("anytime") algorithms such

---

[1]In the generic case, all pairwise distances of data points are distinct and this guarantees that single linkage clustering yields a unique tree [15, 17].

as this are naturally suited to time varying data sets that arise insertions, deletions or updates of a set of data points must be accommodated. Our particular local restructuring method can also cope with time-varying dissimilarity measures or cluster linkage functions such as might result from the introduction of learning aimed at increasing clustering accuracy [40].

1.2 A Brief Summary of Related Literature

Two common approaches to remediating the limited scaling capacity and static nature of hierarchical clustering methods are data abstraction (summarization) and incremental clustering [5, 23].

Rather than improving algorithmic complexity of a specific clustering method, data abstraction aims to scale down a large data set with minimum loss of information for efficient clustering. The large literature on data abstraction includes (but is not limited to) such widely used methods as: random sampling (e.g., CLARANS [31]); selection of representative points (e.g., CURE [19], data bubble [9]); usage of cluster prototypes(e.g., Stream [18]) and sufficient statistics (e.g., BIRCH [41], scalable k-means [8], CluStream [3], data squashing [12]); grid-based quantization [5, 20] and sparcification of connectivity or distance matrix (e.g., CHAMELEON [24]).

In contrast, incremental approaches to hierarchical clustering generally target algorithmic improvements for efficient handling of large data sets by processing data in sequence, point by point. Typically, incremental clustering proceeds in two stages: first (i) locate a new data point in the currently available clustering hierarchy, and then (ii) perform a set of restructuring operations (cluster merging, splitting or creation), based on a heuristic criterion, to obtain a better clustering model. Unfortunately, this sequential process generally incurs unwelcome sensitivity to the order of presentation [5, 23]. Independent of the efficiency and accuracy of our clustering method, the results we report here may be of interest to those seeking insight into the possible spread of outcomes across the combinatorial explosion of different paths through even a fixed data set. Among the many alternatives (e.g., the widely accepted COBWEB [14] or BIRCH [41] procedures), our anytime method most closely resembles the incremental clustering approach of [39], and relies on analogous structural criteria, using similar concepts ("homogeneity" and "monotonicity"). However, a major advantage afforded by our new homogeneity criterion (Definition 4) relative to that of [39] is that there is now no requirement for a minimum spanning tree over the dataset. Beyond ameliorating the computational burden, this relaxation extends the applicability of our method beyond single-linkage to a subclass of linkages, Definition 5, a family of cluster distance functions that includes single, complete, average, minimax and Ward's linkages [22].

Of course, recursive ("anytime") methods can be adapted to address the general setting of time varying data processing. Beyond the specifics of the data insertion problem handled by incremental clustering methods adapting, we aim for reactive algorithms suited to a range of dynamic settings, including data insertion, deletion, update or perhaps, a processing-induced non-stationarity such time varying dissimilarity measure or linkage function [1]. Hence, as described in the previous section, we propose a partially decentralized, recursive method: a local cluster restructuring

operation yielding a discrete dynamical system in the abstract space of trees guaranteed to improve the hierarchy at each step (relative to a fixed dataset) and to terminate in an appropriately homogenizing cluster hierarchy from any (perhaps even random) initial such structure.

1.3 Organization of The Paper

Section 2 introduces notation and offers a brief summary of the essential background. Section 3 presents our homogeneity criterion and establishes some of its properties. Section 4 introduces a simple anytime hierarchical clustering method that seeks successively to "homogenize" local clusters according to this criterion. We analyze the termination and complexity properties of the method and then illustrate its algorithmic implications by applying it to the specific problem of incremental clustering. Section 5 presents experimental evaluation of the anytime hierarchical clustering method using both synthetic and real datasets. We conclude with a brief discussion of future work in Section 6.

2. BACKGROUND & NOTATION

2.1 Datasets, Patterns, and Statistics

We consider data points (patterns, observations) in $\mathbb{R}^m$ with a dissimilarity measure[2] $d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$, where $m \in \mathbb{N}$ is the dimension of the space containing the dataset and $\mathbb{R}_{\geq 0}$ denotes the set of non-negative real numbers. Note that d need not necessarily be a metric[3], and our results can be easily generalized to qualitative data as well, once some dissimilarity ordering has been defined.

Let $\mathbf{x} = (\mathrm{x}_j)_{j \in J} \in (\mathbb{R}^m)^J$ be a set of data points bijectively labelled by a fixed finite index set J, say $J = [n] := \{1, 2, \ldots, n\}$, and let $\mathbf{x}|I = (\mathrm{x}_i)_{i \in I}$ denote a partial set of observations associated with subset $I \subseteq J$, whose centroid and variance, respectively, are

$$\mathrm{c}(\mathbf{x}|I) := \frac{1}{|I|} \sum_{i \in I} \mathrm{x}_i, \tag{1}$$

$$\mathrm{v}(\mathbf{x}|I) := \frac{1}{|I|} \sum_{i \in I} \|\mathrm{x}_i - \mathrm{c}(\mathbf{x}|I)\|_2^2, \tag{2}$$

where $|I|$ and $\|\mathrm{x}\|_2$ denote the cardinality of set I and the standard Euclidean norm of point $\mathrm{x} \in \mathbb{R}^m$, respectively. Throughout the sequel the term "sufficient cluster statistics" denotes the cardinality, $|I|$, and mean (1) and variance (2) for each cluster in a hierarchy [8].

2.2 Hierarchies

A rooted semi-labelled tree τ over a fixed finite index set J, illustrated in Figure 1, is a directed acyclic graph $G_\tau = (V_\tau, E_\tau)$, whose leaves, vertices of degree one, are bijectively labeled by J and interior vertices all have out-degree at least two; and all of whose edges in E_τ are directed away from a vertex designated to be the *root* [7]. A rooted semi-labelled tree τ uniquely determines (and henceforth will be

---

[2]A dissimilarity measure $d : X \times X \to \mathbb{R}_{\geq 0}$ in X is a symmetric, non-negative and reflexive function, i.e. $d(x, y) = d(y, x)$, $d(x, y) \geq 0$ and $d(x, x) = 0$ for all $x, y \in X$.

[3]A dissimilarity $d : X \times X \to \mathbb{R}_{\geq 0}$ is a metric if it satisfies strong reflexivity and the triangle inequality, i.e. for all $x, y, z \in X$ $d(x, y) = 0 \iff x = y$ and $d(x, y) \leq d(x, z) + d(z, y)$.
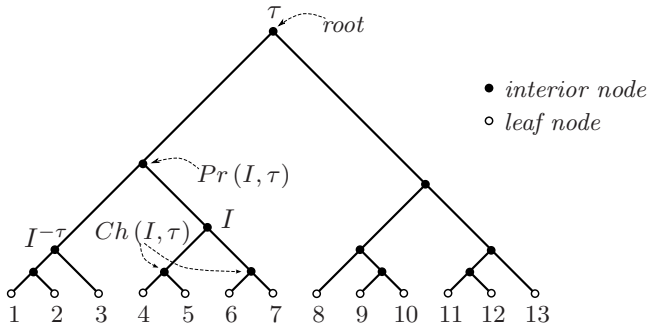
Figure 1: Hierarchical relations: parent - $Pr\,(I,\tau)$, children - $\mathrm{Ch}(I,\tau)$, and local complement (sibling) - $I^{-\tau}$ of cluster I of a rooted binary tree, $\tau \in \mathcal{BT}_{[13]}$. Filled and unfilled circles represent interior and leaf nodes, respectively. An interior node is referred by its cluster, the list of leaves below it; for example, $I = \{4, 5, 6, 7\}$.

interchangeably used with) a *cluster hierarchy* [27]. By definition, all vertices of τ can be reached from the root through a directed path in τ. The *cluster* of a vertex $v \in V_\tau$ is defined to be the set of leaves reachable from v by a directed path in τ. Accordingly, the cluster set $\mathcal{C}\,(\tau)$ of τ is defined to be the set of all its vertex clusters,

$$\mathcal{C}\,(\tau) := \big\{ \mathcal{C}\,(v) \, \big| v \in V_\tau \big\} \subseteq \mathcal{P}(J), \tag{3}$$

where $\mathcal{P}(J)$ denotes the power set of J.

For every cluster $I \in \mathcal{C}\,(\tau)$ we recall the standard notion of *parent* (cluster) $Pr\,(I,\tau)$ and lists of *children* $\mathrm{Ch}(I,\tau)$ of I in τ, illustrated in Figure 1. For the trivial case, we set $Pr\,(J,\tau) = \emptyset$. Additionally, we find it useful to define the *local complement (sibling)* of cluster $I \in \mathcal{C}\,(\tau)$ as $I^{-\tau} := Pr\,(I,\tau) \setminus I$, not to confused with the standard (global) complement $I^C := J \setminus I$. Further, a grandchild in τ is a cluster $G \in \mathcal{C}\,(\tau)$ having a grandparent $\mathrm{Pr}^2(G,\tau) := Pr\,\big(Pr\,(G,\tau),\tau\big)$ in τ. We denote the set of all grandchildren in τ by $\mathcal{G}(\tau)$, the maximal subset of $\mathcal{C}\,(\tau)$ excluding the root J and its children $\mathrm{Ch}(J,\tau)$,

$$\mathcal{G}(\tau) := \big\{ G \in \mathcal{C}\,(\tau) \, \big| \mathrm{Pr}^2(G,\tau) \neq \emptyset \big\}, \tag{4a}$$

$$= \mathcal{C}\,(\tau) \setminus \big(\{J\} \cup \mathrm{Ch}(J,\tau)\big). \tag{4b}$$

A rooted tree with all interior vertices of out-degree two is said to be *binary* or, equivalently, *non-degenerate*, and all other trees are said to be *degenerate*. In this paper $\mathcal{BT}_J$ denotes the set of rooted nondegenerate trees over leaf set J. Note that the number of hierarchies in $\mathcal{BT}_J$ grows super exponentially [7],

$$|\mathcal{BT}_J| = \big(2\,|J| - 3\big)!! = \big(2\,|J| - 3\big)\big(2\,|J| - 5\big)\ldots 3, \tag{5}$$

for $|J| \geq 2$, quickly precluding the possibility of exhaustive search for the "best" hierarchical clustering model in even modest problem settings.

2.3 Nearest Neighbor Interchange (NNI) Moves

Different notions of the neighborhood of a non-degenerate hierarchy in $\mathcal{BT}_J$ can be imposed by recourse to different tree restructuring operations [13] (or *moves*). NNI moves are particularly important for our setting because of their close relation with cluster hierarchy homogeneity (Definition
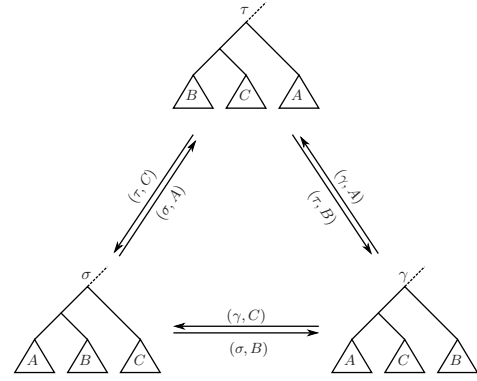
Figure 2: An illustration of NNI moves between binary trees: each arrow is labeled by a source tree and associated cluster defining the move.

4) and their role in the anytime procedure introduced in Section 4.

A convenient restatement of the standard definition of NNI walks [28, 33] for rooted trees, illustrated in Figure 2, is:

DEFINITION 1. *The* Nearest Neighbor Interchange (NNI) *move at a grandchild* $G \in \mathcal{G}(\sigma)$ *on a binary hierarchy* $\sigma \in \mathcal{BT}_J$ *swaps cluster* G *with its parent's sibling* $Pr\,(G,\sigma)^{-\sigma}$ *to yield another binary hierarchy* $\tau \in \mathcal{BT}_J$.

We say that $\sigma, \tau \in \mathcal{BT}_J$ *are* NNI-adjacent *if and only if one can be obtained from the other by a single NNI move.*

More precisely, $\tau \in \mathcal{BT}_J$ is the result of performing the NNI move at grandchild $G \in \mathcal{G}(\sigma)$ on $\sigma \in \mathcal{BT}_J$ if

$$\mathcal{C}\,(\tau) = \big(\mathcal{C}\,(\sigma) \setminus \{Pr\,(G,\sigma)\}\big) \cup \big\{ I^{-\sigma} \cup G^{-\sigma} \big\}. \tag{6}$$

Throughout the sequel we will denote the map of $\mathcal{BT}_J$ into itself defining an NNI move at a grandchild cluster of a tree $G \in \mathcal{G}(\sigma)$ as $\tau = \mathrm{NNI}(\sigma, G)$.

A useful observation for NNI-adjacent hierarchies illustrating their structural difference is:

LEMMA 1. *[4] An ordered pair of hierarchies* (σ, τ) *is NNI-adjacent if and only if there exists one and only one ordered triple* (A, B, C) *of common clusters of* σ *and* τ *such that* $\{A \cup B\} = \mathcal{C}\,(\sigma) \setminus \mathcal{C}\,(\tau)$ *and* $\{B \cup C\} = \mathcal{C}\,(\tau) \setminus \mathcal{C}\,(\sigma)$.

We call (A, B, C) the "*NNI-triplet*" of (σ, τ).

2.4 Hierarchical Agglomerative Clustering

Given a choice of linkage, $\zeta : \big(\mathbb{R}^m\big)^J \times \mathcal{P}(J) \times \mathcal{P}(J) \to \mathbb{R}_{\geq 0}$, Table 1 formalizes the associated Hierarchical Agglomerative ζ-Clustering method [22]. This method yields a sequence of nested partitions of the dataset that can be represented by a tree with root at the coarsest, single cluster partition, leaves at the most refined trivial partition (comprising all singleton sets), and a vertex for each subset appearing in any of the successively coarsened partitions that appears as the mergers of Table 1 are imposed. Because only two clusters are merged at each step, the resulting sequence of nested partitions defines a binary tree, $\tau \in \mathcal{BT}_J$ whose nodes represent the clusters

$$\mathcal{C}\,(\tau) = \bigcup_{k=0}^{|J|-1} \mathcal{J}_k = \{J\} \cup \bigcup_{k=0}^{|J|-2} \{A_k^*, B_k^*\}. \tag{7}$$

Table 1:Hierarchical Agglomerative ζ-Clustering[22]

For any given set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$ and linkage function[4] $\zeta : (\mathbb{R}^m)^J \times \mathcal{P}(J) \times \mathcal{P}(J) \to \mathbb{R}_{\geq 0}$,

- Begin with the finest partition of J, $\mathcal{J}_0 = \big\{ \{i\} \big\}_{i \in J}$.

- For every $k \in [0, |J|-1)$, merge two blocks of $\mathcal{J}_k$ with the minimum linkage value,[5]

$$(A_k^*, B_k^*) = \underset{A \neq B \in \mathcal{J}_k}{\arg\min} \, \zeta(\mathbf{x}; A, B), \qquad (9a)$$

$$\mathcal{J}_{k+1} = \{A_k^* \cup B_k^*\} \cup \mathcal{J}_k \setminus \{A_k^*, B_k^*\}. \qquad (9b)$$

and whose edges represent the nesting relation, again as presented in Table 1. Hence, the set of grandchildren clusters $\mathcal{G}(\tau)$ (4) of τ is given by

$$\mathcal{G}(\tau) = \bigcup_{k=0}^{|J|-3} \mathcal{J}_k = \bigcup_{k=0}^{|J|-3} \{A_k^*, B_k^*\}. \qquad (8)$$

From this discussion it is clear that Table 1 defines a relation from datasets to trees, $\mathtt{HAC}_\zeta \subset (\mathbb{R}^m)^J \times \mathcal{BT}_J$. Note that $\mathtt{HAC}_\zeta$ is in general not a function since there may well be more than one pair of clusters satisfying (9a) at any stage, k. It is, however, a multi-function: in other words, while agglomerative clustering of a dataset always yields some tree, that tree is not necessarily unique to that dataset.

2.4.1 Linkages

A linkage, $\zeta : (\mathbb{R}^m)^J \times \mathcal{P}(J) \times \mathcal{P}(J) \to \mathbb{R}_{\geq 0}$, uses the dissimilarity of observations in the partial datasets, $\mathbf{x}|A$ and $\mathbf{x}|B$, to define dissimilarity between the clusters, $A, B \subseteq J$ and $\mathbf{x} \in (\mathbb{R}^m)^{|J|}$ [2]. Some common examples are

$$\zeta_S(\mathbf{x}; A, B) := \min_{\substack{a \in A \\ b \in B}} d(\mathbf{x}_a, \mathbf{x}_b), \qquad (10a)$$

$$\zeta_C(\mathbf{x}; A, B) := \max_{\substack{a \in A \\ b \in B}} d(\mathbf{x}_a, \mathbf{x}_b), \qquad (10b)$$

$$\zeta_A(\mathbf{x}; A, B) := \frac{1}{|A|\,|B|} \sum_{\substack{a \in A \\ b \in B}} d(\mathbf{x}_a, \mathbf{x}_b), \qquad (10c)$$

$$\zeta_M(\mathbf{x}; A, B) := \min_{a \in A \cup B} \max_{b \in A \cup B} d(\mathbf{x}_a, \mathbf{x}_b), \qquad (10d)$$

$$\zeta_W(\mathbf{x}; A, B) := \frac{|A|\,|B|}{|A| + |B|} \|\mathbf{c}(\mathbf{x}|A) - \mathbf{c}(\mathbf{x}|B)\|_2^2 \qquad (10e)$$

for single, complete, average, minimax and Ward's linkages, respectively [6,22], where d and $\|.\|_2$ are a dissimilarity measure in $\mathbb{R}^m$ and the standard Euclidean norm , respectively.

A common way of characterizing linkages is through their behaviours after merging a set of clusters. For any pairwise disjoint subsets A, B, C of J and dataset $\mathbf{x} \in (\mathbb{R}^m)^J$, a linkage relation between partial observations $\mathbf{x}|A \cup B$ and $\mathbf{x}|C$ after merging $\mathbf{x}|A$ and $\mathbf{x}|B$ are generally described by the recurrence formula of Lance and Williams [25],

$$\zeta(\mathbf{x}; A \cup B, C) = \alpha_A \zeta(\mathbf{x}; A, C) + \alpha_B \zeta(\mathbf{x}; B, C) + \beta \zeta(\mathbf{x}; A, B)$$
$$+ \gamma \left| \zeta(\mathbf{x}; A, C) - \zeta(\mathbf{x}; B, C) \right|, \qquad (11)$$

---

[4]Note that the linkage between any partial observations and the empty set is always defined to be zero, i.e. $\zeta(\mathbf{x}; I, \emptyset) = \zeta(\mathbf{x}; \emptyset, I) = 0$ for all $I \subseteq J$ and $\mathbf{x} \in (\mathbb{R}^m)^J$.

[5]Note that a non-degenerate hierarcy over the leaf set J always has $|J| - 1$ interior nodes [33].

Table 2: Coefficients of the recurrence formula of [25] for some common linkages

Linkage	α_A	α_B	β	γ
Single	0.5	0.5	0	-0.5
Complete	0.5	0.5	0	0.5
Average	$\frac{|A|}{|A|+|B|}$	$\frac{|B|}{|A|+|B|}$	0	0
Ward	$\frac{|A|+|C|}{|A|+|B|+|C|}$	$\frac{|B|+|C|}{|A|+|B|+|C|}$	$-\frac{|C|}{|A|+|B|+|C|}$	0

where $\zeta : (\mathbb{R}^m)^J \times \mathcal{P}(J) \times \mathcal{P}(J) \to \mathbb{R}_{\geq 0}$ is a linkage function and $\alpha_A, \alpha_B, \beta, \gamma \in \mathbb{R}$. Table 2 lists the coefficient of (11) for some common linkages in (10). Although the minimax linkage ζ_M (10d) can not be written in the form of the recurrence formula [6], as many other linkage functions above it satisfies

$$\zeta_M(\mathbf{x}; A \cup B, C) \geq \min\big(\zeta_M(\mathbf{x}; A, C), \zeta_M(\mathbf{x}; B, C)\big), \quad (12)$$

which is known as the strong reducibility property, defined in the following paragraph.

2.4.2 Reducibility & Monotonicity

DEFINITION 2 ([10,29]). *For a fixed finite index set J, a linkage function $\zeta : (\mathbb{R}^m)^J \times \mathcal{P}(J) \times \mathcal{P}(J) \to \mathbb{R}_{\geq 0}$ is said to be* reducible *if for any pairwise disjoint subsets A, B, C of J and set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$*

$$\zeta(\mathbf{x}; A, B) \leq \min\big(\zeta(\mathbf{x}; A, C), \zeta(\mathbf{x}; B, C)\big) \qquad (13a)$$

implies

$$\zeta(\mathbf{x}; A \cup B, C) \geq \min\big(\zeta(\mathbf{x}; A, C), \zeta(\mathbf{x}; B, C)\big). \qquad (13b)$$

Further, say ζ is strongly reducible[6] *if for any pairwise disjoint subsets A, B, C of J and $\mathbf{x} \in (\mathbb{R}^m)^J$ it satisfies*

$$\zeta(\mathbf{x}; A \cup B, C) \geq \min\big(\zeta(\mathbf{x}; A, C), \zeta(\mathbf{x}; B, C)\big). \qquad (14)$$

The well known examples of linkages with the strong reducibility property are single, complete, average and minimax linkages in (10) [6,29]. Even though Ward's linkage is not strongly reducible, it still has the reducibility property.

A property of clustering hierarchies (of great importance in the sequel) consequent upon the reducibility property of linkages is monotonicity:

DEFINITION 3 ([22]). *A non-degenerate hierarchy $\tau \in \mathcal{BT}_J$ associated with a set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$ is said to be ζ-monotone if all grandchildren, $I \in \mathcal{G}(\tau)$, are more similar to their siblings, $I^{-\tau}$, than are their parents, $P = Pr(I, \tau)$, i.e.*

$$\zeta\big(\mathbf{x}; I, I^{-\tau}\big) \leq \zeta\big(\mathbf{x}; P, P^{-\tau}\big). \qquad (15)$$

PROPOSITION 1 ([22]). *If linkage ζ is reducible, then a cluster hierarchy $\tau \in \mathcal{BT}_J$ in the relation $\mathtt{HAC}_\zeta$ (i.e. resulting from procedure of Table 1 applied to some dataset $\mathbf{x} \in (\mathbb{R}^m)^J$) is always ζ-monotone.*

---

[6]Although [16] refers to strong reducibility of linkages as the reducibility property, by definition, strong reducibility is more restrictive than reducibility of linkages.

3. HOMOGENEITY

We now introduce our new notion of homogeneity and explore its relationships to previously developed structural properties of trees.

DEFINITION 4. *(Homogeneity) A binary hierarchy $\tau \in \mathcal{BT}_J$ associated with a set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$ is locally ζ-homogeneous at grandchild cluster $I \in \mathcal{G}(\tau)$ if the siblings, I and $I^{-\tau}$, are closer to each other than to their parent's sibling, $P^{-\tau} = Pr(I, \tau)^{-\tau}$,*

$$\zeta(\mathbf{x}; I, I^{-\tau}) \leq \min(\zeta(\mathbf{x}; I, P^{-\tau}), \zeta(\mathbf{x}; I^{-\tau}, P^{-\tau})). \quad (16)$$

A tree is ζ-homogeneous if it is locally ζ-homogeneous at each grandchild.

A useful observation when we focus attention on reducible linkages is:

PROPOSITION 2. *If a tree, $\tau \in \mathcal{BT}_J$ associated with a set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$, is ζ-homogeneous for a reducible linkage ζ, then it must be ζ-monotone as well.*

PROOF. The result directly follows from homogeneity of τ and reducibility of ζ.
For any grandchild cluster $I \in \mathcal{G}(\tau)$ and its parent $P = Pr(I, \tau)$, using (13) and (16), one can verify the result as

$$\zeta(\mathbf{x}; I, I^{-\tau}) \leq \min(\zeta(\mathbf{x}; I, P^{-\tau}), \zeta(\mathbf{x}; I^{-\tau}, P^{-\tau})) \leq \zeta(\mathbf{x}; P, P^{-\tau}), (17)$$

where $P = I \cup I^{-\tau}$. $\quad \square$

The converse of Proposition 2 only holds for single linkage:

PROPOSITION 3. *A clustering hierarchy $\tau \in \mathcal{BT}_J$ associated with a set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$ is ζ_S-monotone for single linkage ζ_S (10a) if and only if it is ζ_S-homogeneous as well.*

PROOF. The sufficiency of ζ_S-homogeneity of a clustering tree for its ζ_S-monotonicity directly follows from Proposition 2.
The other way of implication is evident from definitions of monotonicity (Definition 3) and single linkage ζ_S (10a), i.e. for any $I \in \mathcal{G}(\tau)$ and $P = Pr(I, \tau)$,

$$\zeta_S(\mathbf{x}; I, I^{-\tau}) \leq \zeta_S(\mathbf{x}; P, P^{-\tau}), \quad (18)$$

$$= \min(\zeta_S(\mathbf{x}; I, P^{-\tau}), \zeta_S(\mathbf{x}; I^{-\tau}, P^{-\tau})), \quad (19)$$

where $P = I \cup I^{-\tau}$. $\quad \square$

A major significance of homogeneity is that it is a common characteristic feature of any clustering hierarchy resulting from agglomerative clustering using any strong reducible linkage:

PROPOSITION 4. *If linkage ζ is strongly reducible then any non-degenerate hierarchy $\tau \in \mathcal{BT}_J$ in the relation $\mathtt{HAC}_\zeta$ (i.e. resulting from the procedure of Table 1 applied to some dataset $\mathbf{x} \in (\mathbb{R}^m)^J$) is ζ-homogeneous.*

PROOF. Let $(\mathcal{J}_k)_{0 \leq k \leq |J|-1}$ be a sequence of nested partitions of J, defining τ as in (7), resulting from agglomerative ζ-clustering of $\mathbf{x}$. Further, for $0 \leq k \leq |J| - 2$ let (A_k^*, B_k^*) be a pair of clusters of $\mathcal{J}_k$ in (9a) with the minimum linkage value.
For $0 \leq k \leq |J| - 3$ and any (grandchild) cluster $C_k \in \mathcal{J}_k \setminus \{A_k^*, B_k^*\}$, from (9a), we have

$$\zeta(\mathbf{x}; A_k^*, B_k^*) \leq \zeta(\mathbf{x}; A_k^*, C_k), \quad (20)$$

$$\zeta(\mathbf{x}; A_k^*, B_k^*) \leq \zeta(\mathbf{x}; B_k^*, C_k). \quad (21)$$

Now, observe that the parent's sibling $(A_k^* \cup B_k^*)^{-\tau}$ of A_k^* and B_k^* can be written as the union of elements of a subset $\mathcal{D}$ of $\mathcal{J}_k \setminus (A_k^*, B_k^*)$,

$$(A_k^* \cup B_k^*)^{-\tau} = \bigcup_{D \in \mathcal{D}} D. \quad (22)$$

That is to say, the elements of $\mathcal{D}$ are merged in a way described by the sequence of nested partitions $(\mathcal{J}_k)_{0 \leq k \leq |J|-1}$ of J such that their union finally yields $(A_k^* \cup B_k^*)^{-\tau}$.
Hence, using strong reducibility of ζ and (20), one can verify that

$$\zeta\left(\mathbf{x}; A_k^*, (A_k^* \cup B_k^*)^{-\tau}\right) = \zeta\left(\mathbf{x}; A_k^*, \bigcup_{D \in \mathcal{D}} D\right), \quad (23)$$

$$\geq \min_{D \in \mathcal{D}} \zeta(\mathbf{x}; A_k^*, D), \quad (24)$$

$$\geq \zeta(\mathbf{x}; A_k^*, B_k^*), \quad (25)$$

which, by symmetry, also holds for B_k^*,

$$\zeta\left(\mathbf{x}; B_k^*, (A_k^* \cup B_k^*)^{-\tau}\right) \geq \zeta(\mathbf{x}; A_k^*, B_k^*). \quad (26)$$

Thus, since $\mathcal{G}(\tau) = \bigcup_{k=0}^{|J|-3} \{A_k^*, B_k^*\}$ (8), the result follows. $\quad \square$

In particular, a critical observation for single linkage is:

THEOREM 1. *A non-degenerate hierarchy $\tau \in \mathcal{BT}_J$ is in the relation $\mathtt{HAC}_{\zeta_S}$ (i.e. results from the procedure of Table 1 applied to some dataset $\mathbf{x} \in (\mathbb{R}^m)^J$ using ζ_S (10a) as the linkage) if and only if τ is ζ_S-homogeneous (or, equivalently, ζ_S-monotone).*

PROOF. The sufficiency, of being a single linkage clustering hierarchy, for homogeneity is evident from Proposition 4.
To see the necessity of homogeneity, we will first prove that if τ is ζ_S-homogeneous, then for any $I \in \mathcal{G}(\tau)$ and nonempty subset $Q \subseteq J \setminus Pr(I, \tau)$ the following holds

$$\zeta_S(\mathbf{x}; I, I^{-\tau}) \leq \zeta_S(\mathbf{x}; I, Q). \quad (27)$$

Observe that (27) states that the cost of merging any one of I and $I^{-\tau}$ with another cluster $Q \subseteq J \setminus Pr(I, \tau)$ is greater and equal to the cost of merging I and $I^{-\tau}$. Then, by induction, we conclude that τ is a possible outcome of agglomerative single linkage clustering of $\mathbf{x}$.

Let $Anc(I, \tau)$ denote the set of ancestors of cluster $I \in \mathcal{C}(\tau)$ of τ, except the root J,

$$Anc(I, \tau) := \{A \in \mathcal{C}(\tau) \setminus \{J\} \mid I \subsetneq A\}. \quad (28)$$

Using the definition of ζ_S (10a) and monotonicity of τ, one can verify that for any grandchild $I \in \mathcal{G}(\tau)$ and its ancestor $A \in Anc(I, \tau)$

$$\zeta_S(\mathbf{x}; I, I^{-\tau}) \leq \zeta_S(\mathbf{x}; A, A^{-\tau}) \leq \zeta_S(\mathbf{x}; I, A^{-\tau}). \quad (29)$$

Now observe that the global complement of $Pr(I, \tau)$ can be written as

$$J \setminus Pr(I, \tau) = \bigcup_{A \in Anc(I, \tau)} A^{-\tau}. \quad (30)$$

As a result, combining (29) and (30) yields

$$\zeta_S(\mathbf{x}; I, J \setminus Pr(I, \tau)) = \zeta_S\left(\mathbf{x}, I, \bigcup_{A \in Anc(I, \tau)} A^{-\tau}\right) \quad (31)$$

$$= \min_{A \in Anc(I, \tau)} \zeta_S\big(\mathbf{x}; I; A^{-\tau}\big) \quad (32)$$

$$\geq \zeta_S\big(\mathbf{x}; I, I^{-\tau}\big), \quad (33)$$

from which one can conclude (27) for single linkage ζ_S.

Finally, using a proof by induction, the result of the theorem can be shown as follows:

- (Base Case) if $I, I^{-\tau} \in \mathcal{C}(\tau)$ are singleton clusters, then, since they satisfy (27), they can be merged at an appropriate step of the agglomerative clustering when the minimum cluster distance is equal to $\zeta_S\big(\mathbf{x}; I, I^{-\tau}\big)$. Note that, due to (27), neither I nor $I^{-\tau}$ can be merged with any other cluster $Q \subseteq J \setminus Pr(I, \tau)$ at a lower linkage value than $\zeta_S\big(\mathbf{x}; I, I^{-\tau}\big)$.

- (Induction) Otherwise, suppose that I and $I^{-\tau}$ are already constructed since their children also satisfy (27) and, by monotonicity, children $\{I_L, I_R\} = \mathrm{Ch}(I, \tau)$ of I satisfies $\zeta_S(\mathbf{x}; I_L, I_R) \leq \zeta_S\big(\mathbf{x}; I, I^{-\tau}\big)$ as do children of $I^{-\tau}$. Thus, since clusters I and $I^{-\tau}$ satisfy (27), they can be directly aggregated when the merging cost, i.e. the value of minimum cluster distance in (9a), reaches $\zeta_S\big(\mathbf{x}; I, I^{-\tau}\big)$.

$\square$

4. ANYTIME HIERARCHICAL CLUSTERING

Given a choice of linkage, $\zeta : (\mathbb{R}^m)^J \times \mathcal{P}(J) \times \mathcal{P}(J) \to \mathbb{R}_{\geq 0}$, Table 3 presents the formal specification of our central contribution, the associated Anytime Hierarchical ζ-Clustering method. Once again, this method defines a new relation from datasets to hierarchies, $\mathtt{AHC}_\zeta \subset (\mathbb{R}^m)^J \times \mathcal{BT}_J$ that is generally not a function but rather a multi-function (i.e. all datasets yield some hierarchy, but not necessarily a unique one).

Because the procedure defining $\mathtt{AHC}_\zeta$ in Table 3 does not entail any obvious gradient-like greedy step as do many previously proposed iterative clustering methods, demonstrating that it terminates requires some analysis that we now present.

4.1 Proof of Convergence

For any non-degenerate hierarchy $\tau \in \mathcal{BT}_J$ associated with a set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$ and a linkage function ζ, we consider the sum of linkage values as an objective function to assess the quality of clustering,

$$H_{\mathbf{x}, \zeta}(\tau) := \frac{1}{2} \sum_{I \in \mathcal{C}(\tau)} \zeta\Big(\mathbf{x}; I, I^{LC}\Big). \quad (36)$$

Intuitively, one might expect that hierarchical agglomerative clustering methods yield clustering hierarchies minimizing (36). However, they are generally known to be step-wise optimal greedy methods [16] with an exception that single linkage clustering always returns a globally optimal clustering tree in the sense of (36) due to its close relation with a minimum spanning tree of the data set [17]. In contrast,

---

Table 3: Anytime Hierarchical ζ-Clustering

For any given clustering hierarchy $\tau \in \mathcal{BT}_J$ associated with a set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$, and linkage function $\zeta : (\mathbb{R}^m)^J \times \mathcal{P}(J) \times \mathcal{P}(J) \to \mathbb{R}_{\geq 0}$,

1. If τ is ζ-homogeneous, then terminate and return τ.

2. Otherwise,

   (a) Find a grandchild cluster $I \in \mathcal{G}(\tau)$ at which τ violates local homogeneity, i.e.

   $$\zeta\big(\mathbf{x}; I, I^{-\tau}\big) > \min\big(\zeta\big(\mathbf{x}; I, P^{-\tau}\big), \zeta\big(\mathbf{x}; I^{-\tau}, P^{-\tau}\big)\big), (34)$$

   where $P = Pr(I, \tau)$.

   (b) Then perform an NNI restructuring on τ at grandchild $G^* \in \mathrm{Ch}(P, \tau)$ with the maximum dissimilarity to $P^{-\tau}$, i.e. swap G^* with $P^{-\tau}$,

   $$G^* = \arg\max_{G \in \mathrm{Ch}(P, \tau)} \zeta\big(\mathbf{x}; G, P^{-\tau}\big), \quad (35\mathrm{a})$$

   $$\tau \leftarrow \mathrm{NNI}(\tau, G^*), \quad (35\mathrm{b})$$

   and go to Step 1.

---

for example, as witness to the general sub-optimality of agglomerative clustering relative to (36), for Ward's linkage ζ_W (10e) $H_{\mathbf{x}, \zeta_W}$ is constant and equal to the sum of squared error of $\mathbf{x}$ (see Appendix A), i.e. for any $\tau \in \mathcal{BT}_J$

$$H_{\mathbf{x}, \zeta_W}(\tau) = \sum_{i \in J} \|\mathrm{x}_i - \mathrm{c}(\mathbf{x}|J)\|_2^2, \quad (37)$$

where $\mathrm{c}(\mathbf{x}|J)$ (1) denotes the centroid of $\mathbf{x}|J$.

Let (σ, τ) be a pair of NNI-adjacent (Definition 1) hierarchies in $\mathcal{BT}_J$ and (A, B, C) be the NNI-triplet (Lemma 1) of common clusters of σ and τ. Recall that $A \cup B \in \mathcal{C}(\sigma) \setminus \mathcal{C}(\tau)$ and $B \cup C \in \mathcal{C}(\tau) \setminus \mathcal{C}(\sigma)$ are only unshared clusters of σ and τ, respectively. Hence, one can write the change in the objective function $H_{\mathbf{x}, \zeta}$ (36) after the NNI transition from σ to τ as

$$H_{\mathbf{x}, \zeta}(\tau) - H_{\mathbf{x}, \zeta}(\sigma) = \zeta(\mathbf{x}; B, C) + \zeta(\mathbf{x}; B \cup C, A) \\ - \zeta(\mathbf{x}; A, B) - \zeta(\mathbf{x}; A \cup B, C). \quad (38)$$

Here we find it useful to define a new class of linkages:

DEFINITION 5. *A linkage* $\zeta : (\mathbb{R}^m)^J \times \mathcal{P}(J) \times \mathcal{P}(J) \to \mathbb{R}_{\geq 0}$ *is* NNI-reducible *if for any set of data points* $\mathbf{x} \in (\mathbb{R}^m)^J$ *and pairwise disjoint subsets* A, B, C *of* J

$$\zeta(\mathbf{x}; B, C) \leq \min\big(\zeta(\mathbf{x}; A, B), \zeta(\mathbf{x}; A, C)\big). \quad (39\mathrm{a})$$

implies

$$\min\begin{pmatrix} \zeta(\mathbf{x}; A, B) + \zeta(\mathbf{x}; A \cup B, C) \\ \zeta(\mathbf{x}; A, C) + \zeta(\mathbf{x}; A \cup C, B) \end{pmatrix} \geq \zeta(\mathbf{x}; B, C) + \zeta(\mathbf{x}; B \cup C, A). (39\mathrm{b})$$

Using (10), (11) and Table 2 one can verify that single, complete, minimax and Ward's linkages are examples of NNI-reducible linkages. Note that a reducible linkage is not necessarily NNI-reducible; for instance, average linkage ζ_A (10c).

We now proceed to investigate the termination of anytime hierarchical clustering for NNI-reducible linkages:

LEMMA 2. *For any set of data points* $\mathbf{x} \in (\mathbb{R}^m)^J$ *and NNI-reducible linkage* ζ, *the value of objective function* $H_{\mathbf{x}, \zeta}$

(36) is nonincreasing at each iteration of anytime hierarchical clustering in Table 3 from any initial hierarchy $\sigma \in \mathcal{BT}_J$ towards $\tau \in \mathcal{BT}_J$,

$$H_{\mathbf{x},\zeta}(\tau) - H_{\mathbf{x},\zeta}(\sigma) \leq 0. \tag{40}$$

PROOF. If σ is ζ-homogeneous, then $\tau = \sigma$, and so the result directly follows.

Otherwise, let (A, B, C) be the NNI-triplet (Lemma 1) associated with (σ, τ). Recall that $A \cup B \in \mathcal{C}(\sigma) \setminus \mathcal{C}(\tau)$ and $B \cup C \in \mathcal{C}(\tau) \setminus \mathcal{C}(\sigma)$. To put it another way, anytime hierarchical clustering performs an NNI move on σ at grandchild $A \in \mathcal{G}(\tau)$ towards τ, and so

$$\zeta(\mathbf{x}; A, B) > \zeta(\mathbf{x}; B, C), \tag{41}$$
$$\zeta(\mathbf{x}; A, C) \geq \zeta(\mathbf{x}; B, C). \tag{42}$$

Therefore, since ζ is NNI-reducible (Definition 5), the change in the objective function $H_{\mathbf{x},\zeta}$ (38) is nonnegative,

$$H_{\mathbf{x},\zeta}(\tau) - H_{\mathbf{x},\zeta}(\sigma) \leq 0, \tag{43}$$

which completes the proof. $\square$

THEOREM 2. *If ζ is an NNI-reducible linkage, then iterated application of the Anytime Hierarchical ζ-Clustering procedure of Table 3 initiated from any hierarchy in $\mathcal{BT}_J$ for a fixed set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$ must terminate in finite time at a tree in $\mathcal{BT}_J$, that is ζ-homogeneous.*

PROOF. For a fixed finite index set J, the number of non-degenerate hierarchies in $\mathcal{BT}_J$ (5) is finite. Hence, for the proof of theorem, we shall show that the anytime clustering procedure in Table 3 can not yield any cycle in $\mathcal{BT}_J$.

Let $\tau^k \in \mathcal{BT}_J$ denote a clustering hierarchy visited at k-th iteration of anytime clustering method, where $k \geq 0$. Since τ^k and τ^{k+1} are NNI-adjacent, let (A^k, B^k, C^k) be the associated NNI-triplet (Lemma 1) of the pair (τ^k, τ^{k+1}) satisfying $A^k \cup B^k \in \mathcal{C}(\tau^k) \setminus \mathcal{C}(\tau^{k+1})$ and $B^k \cup C^k \in \mathcal{C}(\tau^{k+1}) \setminus \mathcal{C}(\tau^k)$. Further, recall from Lemma 2 that for any NNI-reducible linkage ζ, $H_{\mathbf{x};\zeta}(\tau^{k+1}) - H_{\mathbf{x};\zeta}(\tau^k) \leq 0$.

If $H_{\mathbf{x};\zeta}(\tau^{k+1}) - H_{\mathbf{x};\zeta}(\tau^k) < 0$, it is clear that anytime clustering method never revisits any previously visited clustering hierarchy.

Otherwise, $H_{\mathbf{x};\zeta}(\tau^{k+1}) = H_{\mathbf{x};\zeta}(\tau^k)$, we have

$$\zeta(\mathbf{x}; A^k, B^k) + \zeta(\mathbf{x}; A^k \cup B^k, C^k) = \zeta(\mathbf{x}; B^k, C^k) + \zeta(\mathbf{x}; B^k \cup C^k, A^k), \tag{44}$$

$$\zeta(\mathbf{x}; A^k, B^k) > \zeta(\mathbf{x}; B^k, C^k), \tag{45}$$

where the later is due to the anytime clustering rule in Table 3. Hence, the construction cost of (grand)parent $P^k = A^k \cup B^k \cup C^k$ increases after the NNI move,

$$\zeta(\mathbf{x}; A^k \cup B^k, C^k) < \zeta(\mathbf{x}; B^k \cup C^k, A^k). \tag{46}$$

Now, let $\ell_\tau(I)$ denote the level of cluster $I \in \mathcal{C}(\tau)$ of $\tau \in \mathcal{BT}_J$ which is equal to the number of ancestors of I in τ,

$$\ell_\tau(I) := \left| \{ A \in \mathcal{C}(\tau) \mid I \subseteq A \} \right|, \tag{47}$$

and define $L(\tau)$ to be an ordered $(|J|-1)$-tuple of sum of linkages of τ at each level,

$$L(\tau) := (f_\tau(t))_{1 \leq t \leq |J|-1},$$
$$= (f_\tau(1), f_\tau(2), \ldots, f_\tau(|J|-1)), \tag{48}$$

where a binary hierarchy over leaf set J might have at most $|J| - 1$ levels, and

$$f_\tau(t) := \frac{1}{2} \sum_{\substack{I \in \mathcal{C}(\tau) \\ \ell_\tau(I) = t}} \zeta(\mathbf{x}; I, I^{-\tau}). \tag{49}$$

Note that if there is no cluster at level t of τ, then we set $f_\tau(t) = 0$.

Have $(|J|-1)$-tuples of real numbers ordered lexicographically according to the standard order of reals. Then, since NNI transition from τ^k to τ^{k+1} might only change linkages between clusters below (grand)parent cluster $P^k = A^k \cup B^k \cup C^k$, using (46), one can conclude that

$$L(\tau^k) < L(\tau^{k+1}). \tag{50}$$

Thus, it is also impossible to visit the same clustering hierarchy at the same level of objective function $H_{\mathbf{x},\zeta}$, which completes the proof. $\square$

Even though average linkage ζ_A (10c) is not NNI-reducible, as shown in Appendix B, anytime hierarchical clustering based on average linkage still has the finite time termination property.

4.2 A Brief Discussion of Computational Properties

Complexity analysis of any recursive algorithm will necessarily engage two logically independent questions: (i) how many iterations are required to convergence; and (ii) what computational cost is incurred by application of the recursive function at each step along the way? Accordingly, in this section we address this pair of question in the context of the anytime hierarchical clustering algorithm of Table 3. Specifically we : (i) discuss (but defer to a subsequent paper a complete treatment) the problem of determining bounds on the number of iterations of anytime clustering; and (ii) present explicit bounds on the computational cost of checking whether a cluster hierarchy violates local homogeneity at a given cluster (node) of tree or not. Prior work on discriminative comparison of non-degenerate hierarchies [4] and the results of experimental evaluation in Section 5 hint at a bound on the number of iterations (i) that is $O(n^2)$ with the dataset cardinality, n. We leave a comprehensive detailed study of algorithmic complexity of anytime hierarchical clustering to a future discussion of specific implementations. However, we still find it useful to give a brief idea of the computational cost incurred by the determination of tree homogeneity with respect to a number of commonly used linkages.

A straightforward implementation to check (ii) local homogeneity of a clustering hierarchy at any cluster with respect to any linkage function in (10) generally has time complexity of $O(n^2)$ with the dataset size, n, with an exception that local homogeneity of a clustering tree relative to Ward's linkage can be computed in linear, $O(n)$, time.

Alternatively, following the CF(Clustering Feature) tree of BIRCH [41], a simple tree data structure can be used to store sufficient statistics, such as cluster sizes, means and variances, of a clustering hierarchy associated with a dataset. Such a data structure can be constructed in linear time, with the dataset size, using a post-order traversal of a clustering tree and the following recursion of cluster sizes, means and variances. For any $\mathbf{x} \in (\mathbb{R}^m)^J$ and disjoint subsets $A, B \subseteq J$ of a finite index set J, the sufficient statistics of $\mathbf{x}|A \cup B$ can

Table 4: Determining Local Homogeneity of a Clustering Hierarchy at any Cluster

Linkage	Single	Complete	Average	Ward[9]	Minimax	Average[9] with (52)or(54)
Complexity	$O(n^2)$	$O(n^2)$	$O(n^2)$	$O(1)$	$O(n^2)$	$O(1)$

n: the number of data points

be written in terms of the sufficient statistics of $\mathbf{x}|A$ and $\mathbf{x}|B$ as follows:[7]

$$|A \cup B| = |A| + |B|, \tag{51a}$$

$$c(\mathbf{x}|A \cup B) = \frac{|A|}{|A|+|B|}c(\mathbf{x}|A) + \frac{|B|}{|A|+|B|}c(\mathbf{x}|B), \tag{51b}$$

$$v(\mathbf{x}|A \cup B) = \frac{|A|}{|A|+|B|}v(\mathbf{x}|A) + \frac{|B|}{|A|+|B|}v(\mathbf{x}|B)$$
$$+ \frac{|A||B|}{(|A|+|B|)^2}\|c(\mathbf{x}|A) - c(\mathbf{x}|B)\|_2^2, \tag{51c}$$

where for any $I \subseteq J$ $c(\mathbf{x}|I)$(1) and $v(\mathbf{x}|I)$(2) denote the centroid and variance of $\mathbf{x}|I$, respectively. Note that any singleton cluster $i \in \mathcal{C}(\tau)$ has $|i| = 1$, $c(\mathbf{x}|i) = x_i$ and $v(\mathbf{x}|i) = 0$. Also, note that after an NNI restructuring of a clustering tree, the data structure keeping cluster sizes, means and variances can be updated in constant time using (51). Therefore, given the sufficient statistics, local homogeneity of a clustering hierarchy at any cluster with respect to Ward's linkage can be determined in constant time.

To demonstrate another computationally efficient setting of anytime hierarchical clustering, consider the squared Euclidean distance as a dissimilarity measure, i.e. for any $x, y \in \mathbb{R}^m$

$$d(x, y) = \|x - y\|_2^2. \tag{52}$$

As shown in Appendix C, for any $\mathbf{x} \in (\mathbb{R}^m)^J$ and disjoint subsets $A, B \subseteq J$, the average linkage ζ_A (10c) between partial patterns $\mathbf{x}|A$ and $\mathbf{x}|B$, based on the squared Euclidean distance, can be rewritten in terms of sufficient statistics of $\mathbf{x}|A$ and $\mathbf{x}|B$ as [8]

$$\zeta_A(\mathbf{x};A,B) = v(\mathbf{x}|A) + v(\mathbf{x}|B) + \|c(\mathbf{x}|A) - c(\mathbf{x}|B)\|_2^2. \tag{53}$$

Therefore, as in the case of Ward's linkage, given the sufficient statistics of a clustering hierarchy its local homogeneity at any cluster with respect to average linkage with the squared Euclidean distance (52) can be determined in constant time.

A similar computational improvement for average linkage is also possible with the cosine dissimilarity — another commonly used dissimilarity, in information retrieval and text mining [34]: for any $x, y \in \mathbb{R}^m$,

$$d(x, y) = 1 - \frac{x \cdot y}{\|x\|_2 \|y\|_2}, \tag{54}$$

where $\cdot$ denote the Euclidean dot product. For any dataset of unit length vectors $\mathbf{x} \in (\mathbb{S}^{m-1})^J$ and disjoint subsets $A, B \subseteq J$, the average linkage ζ_A (10c) between $\mathbf{x}|A$ and $\mathbf{x}|B$, based on the cosine dissimilarity, is given by

$$\zeta_A(\mathbf{x}; A, B) = 1 - c(\mathbf{x}|A) \cdot c(\mathbf{x}|B). \tag{55}$$

---

[7]A slightly different form of (51) is known as the additivity theorem of CF trees of [41].

[8]This is generally known as the "bias-variance" decomposition of squared Euclidean distance [21].

[9]Assuming the availability of sufficient statistics.

Table 4 briefly summaries the discussion on computational complexity of the determination of local homogeneity of a clustering hierarchy at any cluster.

4.3 Application: Incremental Clustering

As an application of anytime clustering, given a choice of linkage ζ, we now propose an incremental hierarchical clustering method consisting of the following steps: (i) insert a new data point to existing clustering hierarchy based on a specific tree traversal and local homogeneity criterion as described in Table 5, and then (ii) apply anytime clustering of Table 3 to obtain a homogeneous clustering hierarchy of the updated data set with respect to ζ.

Table 5: Incremental Hierarchical ζ-Clustering: Data Insertion Using Local Homogeneity

Let $\tau \in \mathcal{BT}_J$ be a clustering hierarchy associated with a set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$ and ζ be a linkage function. Let $i \notin J$ denote the label of a new data point $x_i \in \mathbb{R}^m$ to be inserted, and $\hat{J} = J \cup \{i\}$ and $\hat{\mathbf{x}} = (x_j)_{j \in \hat{J}}$ be the updated index and data sets after data insertion, respectively.

To insert x_i into the existing clustering hierarchy τ associated with $\mathbf{x}$:

1. Start with $K = J$.

2. For $\{K_L, K_R\} = \text{Ch}(K, \tau)$,

   (a) If $\zeta(\hat{\mathbf{x}}; K_L, K_R) \leq \min(\zeta(\hat{\mathbf{x}}; K_L, \{i\}), \zeta(\hat{\mathbf{x}}; K_R, \{i\}))$, then[5] attach leaf i as the sibling of K in the new clustering tree $\hat{\tau} \in \mathcal{BT}_{\hat{J}}$.

   (b) Otherwise, set $K = \arg\min_{D \in \text{Ch}(K,\tau)} \zeta(\hat{\mathbf{x}}; D, \{i\})$, and go to step 2.

Note that, given the linkage values of a clustering hierarchy, for any linkage function satisfying the recurrence formula (11) of Lance and Williams [25] a data insertion, described in Table 5, can be performed in linear, $O(n)$, time with the dataset size, n. This follows because the linkage distance between the new data point and clusters of an existing hierarchy can be efficiently computed in linear time using a post-order traversal of the clustering hierarchy and (11).

5. EXPERIMENTAL EVALUATION

This section presents a preliminary comparative numerical study of three different hierarchical clustering methods using both simulated and real datasets. We compare: (a) the standard agglomerative batch method (HAC$_\zeta$, Table 1); with (b) the new anytime method (AHC$_\zeta$, Table 3) and (c) its specialization to the incremental "data insertion" problem setting (IHC$_\zeta$, Table 5).

5.1 Datasets

Very high dimensional and sparse data sets generally have simple structure and, specifically, are known to tend toward ultrametricity[11] with the increasing dimensionality and/or

---

[11]A metric $d : X \times X \to \mathbb{R}_{\geq 0}$ is said to be a *ultrametric* if it satisfies the strong triangle inequality, i.e. for any $x, y, z \in X$, $d(x, y) \leq \max(d(x, z), d(z, y))$.
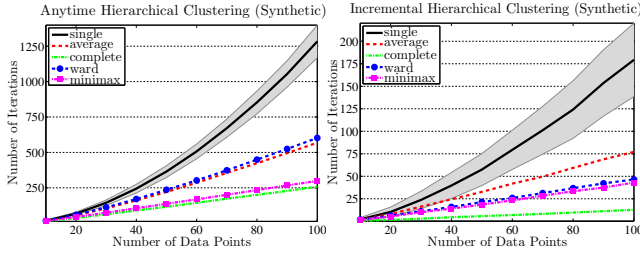
Figure 3: **Average number of iterations of anytime (left) and incremental (right) hierarchical clusterings of simulated data points. Shaded regions illustrate the sample variance for the single linkage.**[13]

Figure 4: **Average number of iterations of anytime (left) and incremental (right) hierarchical clusterings for the MNIST dataset. Shaded regions illustrate the sample variance for the single linkage.**[13]

sparsity [30]. In this context, the fact that a monotone clustering hierarchy associated with a dataset defines an ultrametric between data points (as we will briefly review in the next section) [11], motivates the intuition that hierarchical methods may enjoy particular efficacy in clustering problems involving high dimensional and sparse data. In the following preliminary study we will compare the results of hierarchical clustering on a low dimensional synthetic dataset and a higher dimensional dataset of physical origin. In both cases we will use a validation measure (introduced below) that quantifies the loss of information incurred by approximating the underlying pairwise dissimilarities between points with the coarsened measure arising from the ultrametric induced by the resulting cluster hierarchy.

A challenging dataset for any hierarchical clustering method consists of uniformly distributed low dimensional data points. We generate our synthetic data by uniformly sampling the planar unit cell, $[0,1] \times [0,1]$, thereby generating similar populations of varied cardinality. In contrast, for real data points, we use the MNIST collection of handwritten digits, where each data sample is a black and white 28×28 image of a human produced numeral [26]. We generate test datasets of varied size by randomly sampling an equal number of images for each digit in the MNIST dataset.
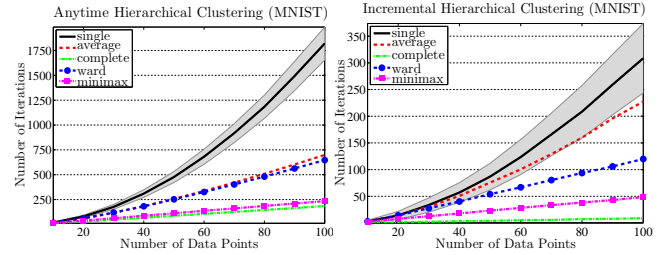
5.2 Validation Measure

To evaluate the accuracy and effectiveness of different hierarchical clustering methods, we use the cophenetic correlation coefficient — a widely accepted validation criterion that measures how well a clustering hierarchy preserves the underlying pairwise dissimilarities between points in a dataset [37]. In order to interpret this criterion we find it helpful to briefly review the manner in which a monotone hierarchy induces an ultrametric between points in a dataset [11].

For any set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$ and a clustering hierarchy $\tau \in \mathcal{BT}_J$ associated with $\mathbf{x}$ and a linkage ζ, let $\mathbf{D}(\mathbf{x}) \in \mathbb{R}^J \times \mathbb{R}^J$ and $\mathbf{U}_\tau(\mathbf{x}) \in \mathbb{R}^J \times \mathbb{R}^J$ denote the original distance matrix of $\mathbf{x}$ and induced ultrametric of τ, respectively. Namely, for any $I \in \mathcal{C}(\tau)$, $i \in I$ and $j \in I^{-\tau}$

$$\mathbf{D}(\mathbf{x})_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2 \,, \tag{56}$$

$$\mathbf{U}_\tau(\mathbf{x})_{ij} = \zeta(\mathbf{x}; I, I^{-\tau}) \,, \tag{57}$$

and for any $i \in J$ set $\mathbf{D}(\mathbf{x})_{ii} = \mathbf{U}(\tau)_{ii} = 0$. The cophenetic correlation coefficient between $\mathbf{D}(\mathbf{x})$ and $\mathbf{U}_\tau(\mathbf{x})$ is defined as

$$\rho_c = \frac{\sum_{i,j \in J} (\mathbf{D}_{ij} - \overline{\mathbf{D}})(\mathbf{U}_{ij} - \overline{\mathbf{U}})}{\sqrt{\sum_{i,j \in J} (\mathbf{D}_{ij} - \overline{\mathbf{D}})^2} \sqrt{\sum_{i,j \in J} (\mathbf{U}_{ij} - \overline{\mathbf{U}})^2}}, \tag{58}$$

where $\overline{\mathbf{D}}$ and $\overline{\mathbf{U}}$ denote the average of the elements of $\mathbf{D}$ and $\mathbf{U}$, respectively, i.e. $\overline{\mathbf{D}} = \frac{1}{|J|^2} \sum_{i,j \in J} \mathbf{D}_{ij}$.
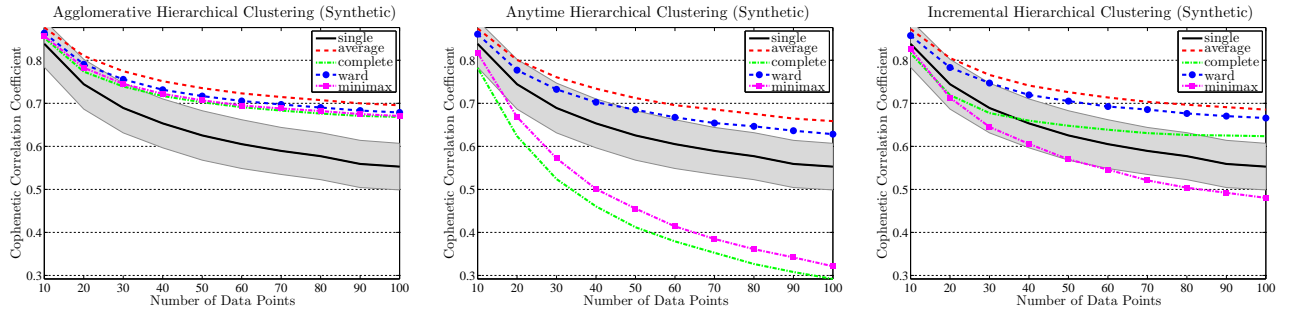
Finally, it is useful to note that Ward's linkage ζ_W (10e) quantifies the change in the sum of squared error after merging clusters [38], and so ζ_W and the standard Euclidean norm do not have the same units. To resolve this unit mismatch, for any clustering hierarchy τ resulting from hierarchical clustering of $\mathbf{x}$ based on Ward's linkage ζ_W we find it convenient to use average linkage ζ_A (10c) to define the induced dissimilarity $\mathbf{U}(\tau)$ of τ.

## 5.3 Preliminary Numerical Results

Using an empirical evaluation of anytime and incremental hierarchical clustering methods we aim to statistically explore: (1) the number of iterations of anytime and incremental clusterings, and (2) their effectiveness compared to the traditional agglomerative clustering methods.

As expected, the number of iterations to homogeneous termination of any anytime clustering depends strongly on the initial conditions (i.e, the initial pair of dataset and tree). To challenge the proposed clustering method, we always start anytime clustering of a dataset at a random initial clustering hierarchy uniformly sampled from the space of non-degenerate hierarchies [36]. To give some preliminary idea of performance as a function of data size we run the various methods on datasets of cardinality $10, 20, \ldots, 100$, and report statistics from the results of 1000 different randomly selected pairings of initial data set and tree using both synthetic and real data collections generated as described in Section 5.1.

For each linkage function discussed in Section 2.4.1, Figure 3 and Figure 4 present the average number of iterations of anytime and incremental clusterings versus the dataset size. Regardless of linkage and size, incremental clustering generally requires an order of magnitude fewer iterations than does anytime clustering. This is a consequence of our experimental design whereby anytime clustering is always initialized from a random clustering tree while incremental clustering takes the advantage of local homogeneity for effective insertion of a new datum into an existing clustering

---

[13]To prevent cluttered figures and give some idea of how the sample variance changes with the cardinality of dataset, we only include the results for single linkage .

Figure 5: Average cophenetic correlation coefficient for agglomerative (left), anytime (middle) and incremental (right) hierarchical clusterings of simulated data points. Shaded regions illustrate the sample variance for the single linkage.[13]
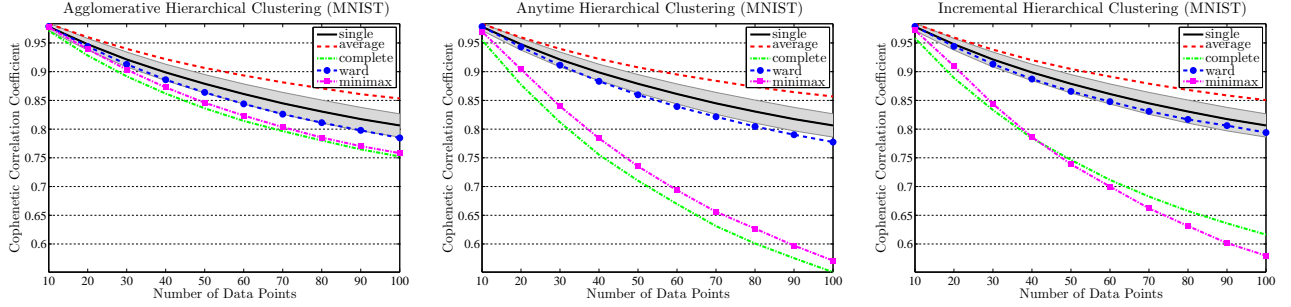
Figure 6: Average cophenetic correlation coefficient for agglomerative (left), anytime (middle) and incremental (right) hierarchical clusterings for the MNIST dataset. Shaded regions illustrate the sample variance for the single linkage.[13]

tree. The next clearest pattern that emerges from these figures is that the number of iterations for both anytime and incremental clustering methods seems to grow quite differently for single linkage (where it appears quadratic in the size of the dataset) than for other familiar linkages (relative to which these preliminary statistics are not inconsistent with linear growth) - but more exploration with larger cardinality datasets will be required before more specific conjectures are possible.

Figure 5 and Figure 6 illustrate how cophenetic correlation coefficient (58) changes with the dataset size for agglomerative, anytime and incremental hierarchical clustering methods. Recall from Theorem 1 that a clustering hierarchy resulting from agglomerative single linkage clustering of a dataset is uniquely characterised by its homogeneity relative to single linkage. Hence, it must be the case (as is indeed reflected in these figures) that the clustering performance of single linkage clustering is the same for all agglomerative, anytime and incremental methods. Here, assuming the results of agglomerative clustering as the ground truth, anytime and incremental clusterings using complete and minimax linkages is observed to perform relatively poorly, which is probably due to their overestimation of cluster dissimilarities. On the other hand, for average and Ward's linkages anytime and incremental hierarchical clusterings perform as well as agglomerative clustering. Further, as expected, incremental clustering generally performs better than anytime clustering since it uses local homogeneity to properly insert each new data point and calls anytime clustering starting at a clustering tree which is far better than a random hierar-

chy. Finally, one can notice that the clustering performance of all hierarchical methods is better on real datasets than synthetic datasets, which is likely due to increased dimensionality and sparsity of data as discussed in Section 5.1.

6. CONCLUSIONS

In this paper, we introduce a new homogeneity criterion (Definition 4), for a clustering tree associated with a data set applicable to a reasonably broad subclass of the familiar linkage functions. We show that homogeneity is a characteristic property of trees resulting from any such standard (linkage based) hierarchical clustering methods (Proposition 4). In particular, homogeneity uniquely characterizes the single linkage clustering tree of a data set (Theorem 1).

We propose an anytime hierarchical clustering method in Table 3 that iteratively transforms any initial clustering hierarchy into a homogeneous clustering tree of a dataset relative to a user-specified linkage function. For the subclass of linkages (specified in Definition 5 — including single, complete, minimax and Ward's linkages) we demonstrate that this iterative clustering procedure must terminate in finite time (Theorem 2). Finally , we discuss certain settings for computationally efficient anytime clustering and describe an incremental hierarchical clustering method based on local homogeneity of cluster trees and anytime clustering.

The "anytime" nature of our method enables users to choose between accuracy and efficiency. In contrast to batch methods, any intermediate stage of anytime hierarchical clustering returns a valid clustering tree and incrementally improves the homogeneity at each iteration. Thus, at any time

a user can stop clustering and continue it later with or without updating the data set. Further, since our method is based on local tree restructuring (the familiar NNI-walk, Definition 1 [28, 33]), it provides an opportunity for distributed/parallel implementation and reactive tracking. The experimental evaluation of sample anytime and incremental hierarchical clustering approaches suggested by these new ideas suggests their value relative to the standard "batch" methods.

Work is presently in progress to establish bounds on the number of iterations of anytime and incremental hierarchical clustering methods, and to develop specific implementations for efficient computation of anytime clustering. In the longer term, we believe these ideas will extend to a randomized algorithm for anytime single linkage clustering as well as to settings where simultaneous distance metric learning must take place in parallel with the hierarchical clustering process.

7. ACKNOWLEDGEMENT

8. REFERENCES

[1] E. Achtert, C. Bohm, H.-P. Kriegel, and P. Kroger. Online hierarchical clustering in a data warehouse environment. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005. 2

[2] M. Ackerman, S. Ben-David, and D. Loker. Characterization of linkage-based clustering. In *COLT*, pages 270–281, 2010. 4

[3] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In *Proceedings of the 29th international conference on Very large data bases*, pages 81–92. VLDB Endowment, 2003. 2

[4] O. Arslan, D. Guralnik, and D. E. Koditschek. Discriminative measures for comparison of phylogenetic trees. Technical report, University Of Pennsylvania, 2013. http://arxiv.org/abs/1310.5202. 3, 7

[5] P. Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006. 1, 2

[6] J. Bien and R. Tibshirani. Hierarchical clustering with prototypes via minimax linkage. *Journal of the American Statistical Association*, 106(495):1075–1084, 2011. 4

[7] L. J. Billera, S. P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001. 2, 3

[8] P. S. Bradley, U. M. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In *Proceedings of the fourth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 9–15, 1998. 2

[9] M. M. Breunig, H.-P. Kriegel, P. Kröger, and J. Sander. Data bubbles: Quality preserving performance boosting for hierarchical clustering. In *ACM SIGMOD Record*, volume 30, pages 79–90. ACM, 2001. 2

[10] M. Bruynooghe. Classification ascendante hiérarchique des grands ensembles de données: un algorithme rapide fondé sur la construction des voisinages réductibles. *Cahiers de l'Analyse des Données*, 3(1):7–33, 1978. 4

[11] G. Carlsson and F. Mémoli. Characterization, Stability and Convergence of Hierarchical Clustering methods. *Journal of Machine Learning Research*, 11:1425–1470, 2010. 9

[12] W. DuMouchel, C. Volinsky, T. Johnson, C. Cortes, and D. Pregibon. Squashing flat files flatter. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 6–15. ACM, 1999. 2

[13] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Suderland, USA, 2004. 3

[14] D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine learning*, 2(2):139–172, 1987. 2

[15] R. G. Gallager, P. A. Humblet, and P. M. Spira. A distributed algorithm for minimum-weight spanning trees. *ACM Transactions on Programming Languages and systems (TOPLAS)*, 5(1):66–77, 1983. 1

[16] A. D. Gordon. A review of hierarchical classification. *Journal of the Royal Statistical Society. Series A (General)*, pages 119–137, 1987. 4, 6

[17] J. C. Gower and G. J. S. Ross. Minimum Spanning Trees and Single Linkage Cluster Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1), 1969. 1, 6

[18] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. *Knowledge and Data Engineering, IEEE Transactions on*, 15(3):515–528, May 2003. 2

[19] S. Guha, R. Rastogi, and K. Shim. Cure: an efficient clustering algorithm for large databases. *Information Systems*, 26(1):35–58, 2001. 2

[20] J. Han, M. Kamber, and J. Pei. *Data mining: concepts and techniques*. Morgan Kaufmann, 2006. 1, 2

[21] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*, volume 2. Springer, 2009. 8, 13

[22] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988. 1, 2, 3, 4, 12

[23] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999. 2

[24] G. Karypis, E.-H. Han, and V. Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999. 2

[25] G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies 1. hierarchical systems. *The Computer Journal*, 9(4):373–380, 1967. 4, 8

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 9

[27] B. Mirkin. *Mathematical Classification and Clustering*. Kluwer Academic Publishers, 1996. 3

[28] G. Moore, M. Goodman, and J. Barnabas. An iterative approach from the standpoint of the additive

hypothesis to the dendrogram problem posed by molecular data sets. *Journal of Theoretical Biology*, 38(3):423–457, 1973. 3, 11

[29] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359, 1983. 4

[30] F. Murtagh. The remarkable simplicity of very high dimensional data: Application of model-based clustering. *Journal of Classification*, 26(3):249–277, 2009. 9

[31] R. T. Ng and J. Han. Clarans: A method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 14(5):1003–1016, 2002. 2

[32] C. F. Olson. Parallel algorithms for hierarchical clustering. *Parallel computing*, 21(8):1313–1325, 1995. 1

[33] D. Robinson. Comparison of labeled trees with valency three. *Journal of Combinatorial Theory, Series B*, 11(2):105–119, 1971. 3, 4, 11

[34] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513 – 523, 1988. 8, 13

[35] A. Schrijver. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer, 2003. 13

[36] C. Semple and M. Steel. *Phylogenetics*, volume 24. Oxford University Press, 2003. 9

[37] R. R. Sokal and F. J. Rohlf. The comparison of dendrograms by objective methods. *Taxon*, 11(2):33–40, 1962. 9

[38] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. 9, 12

[39] D. Widyantoro, T. Ioerger, and J. Yen. An incremental approach to building a cluster hierarchy. In *Data Mining, 2002. ICDM 2002. Proceedings. 2002 IEEE International Conference on*, pages 705–708, 2002. 2

[40] E. P. Xing, M. I. Jordan, S. Russell, and A. Ng. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, pages 505–512, 2002. 2

[41] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. In *ACM SIGMOD Record*, volume 25, pages 103–114. ACM, 1996. 2, 7, 8

APPENDIX

A. SUM OF WARD'S LINKAGES & SUM OF SQUARED ERROR

Although sums of linkage values $H_{\mathbf{x},\zeta}$ (36) of distinct clustering hierarchies associated with a set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$ and a linkage function ζ generally differ, $H_{\mathbf{x},\zeta_W}$ is constant for Ward's linkage ζ_W (10e):

LEMMA 3. *The sum of linkage values $H_{\mathbf{x},\zeta_W}$ (36) of any clustering hierarchy associated with a data set $\mathbf{x} \in (\mathbb{R}^m)^J$ and Ward's linkage ζ_W (10e) is constant and equal to the sum of squared errors of $\mathbf{x}$, i.e. for any $\tau \in \mathcal{BT}_J$*

$$H_{\mathbf{x},\zeta_W}(\tau) = \text{SSE}(\mathbf{x}), \tag{59}$$

where

$$\text{SSE}(\mathbf{x}) := \sum_{i \in J} \|\mathbf{x}_i - \mathbf{c}(\mathbf{x}|J)\|_2^2. \tag{60}$$

PROOF. Recall from [22,38] that Ward's linkage quantifies the change in the sum of squared errors after merging a group of data points, i.e. for any $\mathbf{x} \in (\mathbb{R}^m)^J$ and disjoint subsets $A, B \subseteq J$

$$\zeta_W(\mathbf{x}; A, B) = \text{SSE}(\mathbf{x}|A \cup B) - \text{SSE}(\mathbf{x}|A) - \text{SSE}(\mathbf{x}|B). \tag{61}$$

Let $\tau \in \mathcal{BT}_J$ be a binary hierarchy with the root split $\{J_L, J_R\} = \text{Ch}(J, \tau)$. Proof by induction.

- (Base Case) if $|J| = 2$, then there is only one clustering hierarchy $\tau \in \mathcal{BT}_{[2]}$, i.e. $\left|\mathcal{BT}_{[2]}\right| = 1$. Note that τ only has one Ward's linkage joining two data points of $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ whose value equals to the sum of squared errors of $\mathbf{x}$,

$$H_{\mathbf{x},\zeta_W}(\tau) = \zeta_W(\mathbf{x}; \{1\}, \{2\}) = \frac{1}{2} \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2, \tag{62}$$

$$= \sum_{i \in \{1,2\}} \|\mathbf{x}_i - \mathbf{c}(\mathbf{x}|\{1,2\})\|_2^2 = \text{SSE}(\mathbf{x}). \tag{63}$$

- (Induction) Let τ_L and τ_R denote the subtrees of τ rooted at J_L and J_R, respectively. Suppose that

$$H_{\mathbf{x}|J_L,\zeta_W}(\tau_L) = \text{SSE}(\mathbf{x}|J_L), \tag{64}$$
$$H_{\mathbf{x}|J_R,\zeta_W}(\tau_R) = \text{SSE}(\mathbf{x}|J_R). \tag{65}$$

Note that if any of subtrees only has one leaf, e.g. $|J_L| = 1$, then we set the associated sum of linkage values to zero, $H_{\mathbf{x}|J_L,\zeta_W}(\tau_L) = 0$.

Hence, using (61), one can obtain the result as follows:

$$H_{\mathbf{x},\zeta_W}(\tau) = H_{\mathbf{x}|J_L,\zeta_W}(\tau_L) + H_{\mathbf{x}|J_R,\zeta_W}(\tau_R) + \zeta_W(\mathbf{x}; J_L, J_R), \tag{66}$$
$$= \text{SSE}(\mathbf{x}|J_L) + \text{SSE}(\mathbf{x}|J_R) + \zeta_W(\mathbf{x}; J_L, J_R), \tag{67}$$
$$= \text{SSE}(\mathbf{x}). \tag{68}$$

$\square$

B. TERMINATION ANALYSIS FOR AVERAGE LINKAGE

LEMMA 4. *Iterated application of Anytime Hierarchical ζ_A-Clustering procedure in Table 3 initiated from any clustering tree in $\mathcal{BT}_J$ for a fixed set of data points $\mathbf{x} \in (\mathbb{R}^m)^J$ terminates in finite time at a clustering hierarchy in $\mathcal{BT}_J$ that is homogeneous relative to average linkage ζ_A (10c).*

PROOF. As in the proof of more general result in Theorem 2, we shall show that the anytime clustering rule does not cause any cycle in $\mathcal{BT}_J$ before terminating at a structurally homogeneous clustering hierarchy. Consequently, the finite time termination of the anytime clustering method is simply due to finiteness of tree space $\mathcal{BT}_J$ (5).

Let $S(\tau)$ denote the ordered set of linkage values of a binary clustering hierarchy $\tau \in \mathcal{BT}_J$ associated with $\mathbf{x}$ and ζ_A in ascending order, i.e.

$$S(\tau) := \left(\zeta_A(\mathbf{x}; I, I^{-\tau})\right)_{I \in \mathcal{C}(\tau)}, \tag{69}$$

$$S(\tau)_i \leq S(\tau)_j, \quad \forall 1 \leq i \leq j \leq 2|J| - 1, \tag{70}$$

where $S(\tau) = \left(S(\tau)_1, S(\tau)_2, \ldots, S(\tau)_{2|J|-1}\right)$ and note that the number of clusters of a binary tree over leaf set J is $2|J|-1$ [35]. Further, have the set of $(2|J|-1)$-tuple of real numbers ordered lexicographically according to the standard order of reals.

Let $\tau^k \in \mathcal{BT}_J$ be a clustering hierarchy visited at k-th iteration of anytime hierarchical clustering of $\mathbf{x}$, where $k \geq 0$. To prove the result, we shall show the following

$$S\left(\tau^k\right) > S\left(\tau^{k+1}\right). \tag{71}$$

Since τ^k and τ^{k+1} are NNI-adjacent, let $\left(A^k, B^k, C^k\right)$ be the NNI-triplet (Lemma 1) associated with the pair of $\left(\tau^k, \tau^{k+1}\right)$. Recall that $A^k \cup B^k \in \mathcal{C}\left(\tau^k\right) \setminus \mathcal{C}\left(\tau^{k+1}\right)$ and $B^k \cup C^k \in \mathcal{C}\left(\tau^{k+1}\right) \setminus \mathcal{C}\left(\tau^k\right)$.

Note that after the NNI transition from τ^k to τ^{k+1} two elements of $S\left(\tau^k\right)$, $\zeta_A\left(\mathbf{x}; A^k, B^k\right)$ and $\zeta_A\left(\mathbf{x}; A^k \cup B^k, C^k\right)$, are replaced by another two reals, $\zeta_A\left(\mathbf{x}; B^k, C^k\right)$ and $\zeta_A\left(\mathbf{x}; B^k \cup C^k, A^k\right)$, to yield $S\left(\tau^{k+1}\right)$.

By construction of the anytime clustering rule, we have

$$\zeta_A\left(\mathbf{x}; A^k, B^k\right) > \zeta_A\left(\mathbf{x}; B^k, C^k\right), \tag{72}$$

$$\zeta_A\left(\mathbf{x}; A^k, C^k\right) \geq \zeta_A\left(\mathbf{x}; B^k, C^k\right), \tag{73}$$

and, using the definition of average linkage (10c), one can verify that

$$\zeta_A\left(\mathbf{x}; A^k \cup B^k, C^k\right) \geq \zeta\left(\mathbf{x}; B^k, C^k\right). \tag{74}$$

If $\zeta_A\left(\mathbf{x}; A^k \cup B^k, C^k\right) > \zeta_A\left(\mathbf{x}; B^k, C^k\right)$, then it is clear from (72) and (10c) that

$$\min\left(\zeta_A\left(\mathbf{x}; A^k, B^k\right), \zeta_A\left(\mathbf{x}; A^k \cup B^k, C^k\right)\right) > \zeta_A\left(\mathbf{x}; B^k, C^k\right), \tag{75}$$

$$= \min\left(\zeta_A\left(\mathbf{x}; B^k, C^k\right), \zeta_A\left(\mathbf{x}; B^k \cup C^k, A^k\right)\right). \tag{76}$$

Otherwise $\left(\zeta_A\left(\mathbf{x}; A^k \cup B^k, C^k\right) = \zeta_A\left(\mathbf{x}; B^k, C^k\right)\right)$, by definition (10c), we have $\zeta_A\left(\mathbf{x}; B^k, C^k\right) = \zeta_A\left(\mathbf{x}; A^k, C^k\right)$, and so $\zeta_A\left(\mathbf{x}; A^k, B^k\right) > \zeta_A\left(\mathbf{x}; B^k \cup C^k, A^k\right)$.

In overall, the minimum of changed linkage values at each iteration of anytime clustering strictly decreases, which proves (71) and completes the proof. $\quad\square$

C. SPECIAL CASES OF AVERAGE LINKAGE

We now consider certain settings of average linkage ζ_A (10c) that enables efficient computation of linkage values of a clustering hierarchy and its restructuring during online clustering.

Consider the squared Euclidean distance as a dissimilarity measure of a pair of data points, i.e. for any $x, y \in \mathbb{R}^m$

$$d(x, y) = \|x - y\|_2^2. \tag{77}$$

For any $\mathbf{x} \in \left(\mathbb{R}^d\right)^J$ and disjoint subsets $A, B \subseteq J$, the average linkage ζ_A (10c) between partial patterns $\mathbf{x}|A$ and $\mathbf{x}|B$,

based on the squared Euclidean distance, is

$$\zeta_A(\mathbf{x}; A, B) = \frac{1}{|A||B|} \sum_{\substack{a \in A \\ b \in B}} d(x_a, x_b) = \frac{1}{|A||B|} \sum_{\substack{a \in A \\ b \in B}} \|x_a - x_b\|_2^2, \tag{78}$$

$$= \frac{1}{|A||B|} \sum_{\substack{a \in A \\ b \in B}} \|x_a - c(\mathbf{x}|B) + c(\mathbf{x}|B) - x_b\|_2^2. \tag{79}$$

After expanding the norm, (79) simplifies to

$$\zeta_A(\mathbf{x}; A, B) = \frac{1}{|A|} \sum_{a \in A} \|x_a - c(\mathbf{x}|B)\|_2^2 + \frac{1}{|B|} \sum_{b \in B} \|x_b - c(\mathbf{x}|B)\|_2^2, \tag{80}$$

$$= \frac{1}{|A|} \sum_{a \in A} \|x_a - c(\mathbf{x}|A) + c(\mathbf{x}|A) - c(\mathbf{x}|B)\|_2^2 + v(\mathbf{x}|B). \tag{81}$$

Using a similar trick on (81), one can conclude that the value of average linkage between $\mathbf{x}|A$ and $\mathbf{x}|B$ is a simple function of their centroids and variances, [14]

$$\zeta_A(\mathbf{x}; A, B) = v(\mathbf{x}|A) + v(\mathbf{x}|B) + \|c(\mathbf{x}|A) - c(\mathbf{x}|B)\|_2^2. \tag{82}$$

where for any $I \subseteq J$ $c(\mathbf{x}|I)$(1) and $v(\mathbf{x}|I)$(2) denote the centroid and variance of $\mathbf{x}|I$, respectively.

A similar computational improvement for average linkage is also possible with the cosine dissimilarity — another commonly used dissimilarity, in information retrieval and text mining [34]: for any $x, y \in \mathbb{R}^m$,

$$d(x, y) = 1 - \frac{x \cdot y}{\|x\|_2 \|y\|_2}, \tag{83}$$

where $\cdot$ denote the Euclidean dot product. For any dataset of unit length vectors $\mathbf{x} \in \left(\mathbb{S}^{m-1}\right)^J$ and disjoint subsets $A, B \subseteq J$, the average linkage ζ_A (10c) between $\mathbf{x}|A$ and $\mathbf{x}|B$, based on the cosine dissimilarity, can be rewritten as

$$\zeta_A(\mathbf{x}; A, B) = \frac{1}{|A||B|} \sum_{\substack{a \in A \\ b \in B}} d(x_a, x_b) = \frac{1}{|A||B|} \sum_{\substack{a \in A \\ b \in B}} 1 - \underbrace{\frac{x_a \cdot x_b}{\|x_a\|_2 \|x_b\|_2}}_{\|x_a\|_2 = \|x_b\|_2 = 1},$$

$$= 1 - \frac{1}{|A||B|} \sum_{\substack{a \in A \\ b \in B}} x_a \cdot x_b = 1 - c(\mathbf{x}|A) \cdot c(\mathbf{x}|B), \tag{84}$$

which directly follows the linearity of the dot product.

D. SAMPLE VARIANCE AND MEAN AFTER MERGING CLUSTERS

It is well know that for any $\mathbf{x} \in \left(\mathbb{R}^d\right)^J$ and any disjoint subsets $A, B \subseteq J$ of a finite index set J, the centroid of merged patterns $\mathbf{x}|A \cup B$ is simply equal to the weighted average, proportional to the cardinality of sets, of centroids of partial patterns $\mathbf{x}|A$ and $\mathbf{x}|B$,

$$c(\mathbf{x}|A \cup B) = \frac{|A|}{|A| + |B|} c(\mathbf{x}|A) + \frac{|B|}{|A| + |B|} c(\mathbf{x}|B). \tag{85}$$

Similarly, using (61), one can verify that for any disjoint subsets $A, B \subseteq J$ and $\mathbf{x} \in \left(\mathbb{R}^d\right)^J$, the variance of merged patterns $\mathbf{x}|A \cup B$ is given by

$$v(\mathbf{x}|A \cup B) = \frac{|A|}{|A| + |B|} v(\mathbf{x}|A) + \frac{|B|}{|A| + |B|} v(\mathbf{x}|B)$$

$$+ \frac{|A||B|}{(|A| + |B|)^2} \|c(\mathbf{x}|A) - c(\mathbf{x}|B)\|_2^2. \tag{86}$$

---

[14]This is generally referred to the "bias-variance" decomposition of squared Euclidean distance [21].