Foundations of Comparison-Based Hierarchical Clustering

Debarghya Ghoshdastidar[1], Michaël Perrot[2], and Ulrike von Luxburg[1,2]

[1] *University of Tübingen, Department of Computer Science, Sand 14, 72076 Tübingen, Germany*
[2] *Max-Planck-Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany*

Abstract

We address the classical problem of hierarchical clustering, but in a framework where one does not have access to a representation of the objects or their pairwise similarities. Instead we assume that only a set of comparisons between objects are available in terms of statements of the form "objects i and j are more similar than objects k and l". Such a scenario is commonly encountered in crowdsourcing applications. The focus of this work is to develop comparison-based hierarchical clustering algorithms that do not rely on the principles of ordinal embedding. We propose comparison-based variants of average linkage clustering. We provide statistical guarantees for the proposed methods under a planted partition model for hierarchical clustering. We also empirically demonstrate the performance of the proposed methods on several datasets.

1 Introduction

The definition of clustering as *the task of grouping similar objects* emphasizes the importance of assessing similarity scores for the process of clustering. Unfortunately, many applications of data analysis, particularly in crowdsourcing and psychometric problems, do not come with a natural representation of the underlying objects or a well-defined similarity function between pairs of objects. Instead, one only has access to the results of comparisons of similarities, for instance,
Triplet comparison: Object x_i is more similar to object x_j than to object x_k;
Quadruplet comparison: Similarity between x_i and x_j is larger than similarity between x_k and x_l.

The importance and robustness of collecting such ordinal information from human subjects and crowds have been widely discussed in psychometric and crowdsourcing literature (Shepard, 1962; Young, 1987; Borg and Groenen, 2005; Stewart et al., 2005). Subsequently, there has been growing interest in the machine learning and statistics communities to perform data analysis in a comparison-based framework (Agarwal et al., 2007; van der Maaten and Weinberger, 2012; Heikinheimo and Ukkonen, 2013; Kleindessner and Luxburg, 2014; Zhang et al., 2015; Arias-Castro et al., 2017; Haghiri et al., 2018). The traditional approach for learning in an ordinal setup involves a two step procedure — first obtain an Euclidean embedding of the objects from available similarity comparisons, and subsequently learn from the embedded data using standard machine learning techniques (Borg and Groenen, 2005; Agarwal et al., 2007; Jamieson and Nowak, 2011b; Tamuz et al., 2011; van der Maaten and Weinberger, 2012; Terada and von Luxburg, 2014; Zhang et al., 2015; Amid and Ukkonen, 2015). As a consequence, the statistical performance of the resulting comparison-based learning algorithms relies both on the goodness of the embedding and the subsequent statistical consistency of learning from the embedded data. While there exists theoretical guarantees on the accuracy of ordinal embedding from triplet or quadruplet comparisons (Jamieson and Nowak, 2011b; Kleindessner and Luxburg, 2014; Jain et al., 2016; Arias-Castro et al., 2017), there is no theoretical evidence that it is indeed possible to consistently learn from the mutually dependent embedded data points.

An alternative approach, which has become popular in recent years, is to directly learn from the ordinal relations. This approach has been used in data dimension, centroid or density estimation

(Kleindessner and Luxburg, 2015; Heikinheimo and Ukkonen, 2013; Ukkonen et al., 2015), object retrieval and nearest neighbour search (Kazemi et al., 2018; Haghiri et al., 2017), classification and regression (Haghiri et al., 2018), clustering (Kleindessner and von Luxburg, 2017a; Ukkonen, 2017), as well as hierarchical clustering (Vikram and Dasgupta, 2016; Emamjomeh-Zadeh and Kempe, 2018). The theoretical advantage of a direct learning principle over an indirect embedding-based approach is reflected in the fact that some of the above works come with statistical guarantees for learning from ordinal comparisons (Haghiri et al., 2017, 2018; Kazemi et al., 2018).

1.1 Motivation and Contributions

The motivation for the present work arises from the absence of comparison-based clustering algorithms that have strong statistical guarantees, or more generally, the limited theory in the context of comparison-based clustering and hierarchical clustering. While theoretical foundations of standard hierarchical clustering can be found in the literature (Hartigan, 1981; Chaudhuri et al., 2014; Dasgupta, 2016; Moseley and Wang, 2017), corresponding works in the ordinal setup has been limited (Emamjomeh-Zadeh and Kempe, 2018). A naive approach to derive guarantees for comparison-based clustering is to combine the analysis of a classical clustering or hierarchical clustering algorithm with existing guarantees for ordinal embedding (Arias-Castro et al., 2017). Unfortunately, this does not work since the known worst-case error rates for ordinal embedding are too weak to provide any reasonable guarantee for the resulting comparison-based clustering algorithm. The existing guarantees for ordinal hierarchical clustering hold under a triplet framework, where each comparison returns the two most similar among three objects (Emamjomeh-Zadeh and Kempe, 2018). The results show that the underlying hierarchy can be recovered by few adaptively chosen comparisons, but if the comparisons are provided beforehand, which is the case in crowdsourcing, then the number of required comparisons is rather large.

The focus of the present work is to develop provable comparison-based hierarchical clustering algorithms that can find an underlying hierarchy in a set of objects given a non-adaptively chosen set of comparisons. Our contribution is two-fold.

Agglomerative algorithms for comparison-based clustering. The only known hierarchical clustering algorithm in a comparison-based framework employs a divisive approach (Emamjomeh-Zadeh and Kempe, 2018). However, we observe that it is easy to perform agglomerative hierarchical clustering using only quadruplet comparisons since one can reformulate single linkage and complete linkage clustering algorithms in a quadruplet comparison framework. However, it is well known that single and complete linkage algorithms typically have poor worst-case guarantees (Cohen-Addad et al., 2018). While average linkage clustering is known to have stronger theoretical guarantees (Moseley and Wang, 2017; Cohen-Addad et al., 2018), it cannot be directly applied in a comparison-based setup. We propose three variants of average linkage hierarchical clustering that can be applied to a comparison based framework, where we have access to either quadruplet or triplet comparisons. We numerically compare the merits of these methods with single and complete linkage and embedding based methods.

Guarantees for recovering true hierarchy. Dasgupta (2016) provided a new perspective for hierarchical clustering in terms of optimizing a cost function that depends on the pairwise similarities between objects. Subsequently, theoretical research has focused on worst-case analysis of different algorithms with respect to this cost function (Roy and Pokutta, 2016; Moseley and Wang, 2017; Cohen-Addad et al., 2018). However, such an analysis is complicated in an ordinal setup, where the algorithm is oblivious to the pairwise similarities. In this case, one can study a stronger notion of guarantee in terms of exact recovery of the true hierarchy (Emamjomeh-Zadeh and Kempe, 2018). That work, however, considers a simplistic noise model, where the result of each comparison may be randomly flipped independent of other comparisons (Jain et al., 2016). Such an independent noise can be easily tackled by repeatedly querying the same comparison and using a majority vote. However, this cannot account for noise in the underlying objects and their associated similarities.

We present a theoretical model that generates random pairwise similarities with a planted hierarchical structure. This induces considerable dependence among the available triplet or quadruplet comparisons, and makes the analysis challenging. We derive conditions under which different comparison-based agglomerative algorithms can exactly recover the planted hierarchy with high probability. Cohen-Addad et al. (2018) previously studied a planted hierarchical model where a hierarchy is planted in a random graph. Due to the formulation of that model in terms of an unweighted graph, it cannot be applied to the ordinal setting.

Although we focus on hierarchical clustering, the proposed algorithms and theory can be used for the problem of clustering with a fixed number of clusters. Similarly the presented theoretical techniques can be adapted to analyze other ordinal clustering methods.

2 Background

We consider a standard setup for hierarchical clustering. Let $\mathcal{X} = \{x_i\}_{i=1}^N$ be a set of N objects, which may not have a known feature representation. We assume that there exists an underlying symmetric similarity function $w : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $w_{ij} = w(x_i, x_j)$ denotes the similarity score between x_i and x_j. The goal of hierarchical clustering is to group the N objects to form of a binary tree such that x_i and x_j are merged in the bottom of the tree if w_{ij} is high, and vice-versa.

2.1 Agglomerative Hierarchical Clustering with Known Similarity Scores

In this section, we assume that w_{ij} can be accessed by the clustering algorithm, and briefly review the popular agglomerative clustering algorithms (Cohen-Addad et al., 2018). We focus on agglomerative methods since divisive methods typically require a global structure of the data, which can be rarely obtained from a given non-adaptive set of comparisons. Agglomerative clustering algorithms rely on using the similarity score w between two objects to define a similarity function between any two clusters, $W : 2^{\mathcal{X}} \times 2^{\mathcal{X}} \to \mathbb{R}$. Starting from N singleton clusters, each iteration of the algorithm merges the two clusters with largest similarity according to function W. The approach is described in Algorithm 1, where different choices of W lead to different algorithms. For clusters G and G', the three popular choices for W, respectively corresponding to single, complete and average linkage algorithms, are

$$\textbf{Single (SL):} \qquad W(G, G') = \max_{x_i \in G, x_j \in G'} w_{ij}\,; \tag{1}$$

$$\textbf{Complete (CL):} \qquad W(G, G') = \min_{x_i \in G, x_j \in G'} w_{ij}\,; \tag{2}$$

$$\textbf{Average (AL):} \qquad W(G, G') = \sum_{x_i \in G, x_j \in G'} \frac{w_{ij}}{|G||G'|}. \tag{3}$$

2.2 The Comparison-Based Framework

In the rest of this paper, we consider the ordinal setting, where the similarity function w is not available, and information about similarities can only be accessed through comparisons. We assume that one can access either a set of quadruplets or triplets. We also assume that $w_{ii} = \infty$, and thus $w_{ii} > w_{ik}$ for all $k \neq i$.

In the quadruplet setting, we are given a set

$$\mathcal{Q} = \{(x_i, x_j, x_k, x_l) : x_i, x_j, x_k, x_l \in \mathcal{X}, w_{ij} > w_{kl}\},$$

that is, for every ordered tuple $(x_i, x_j, x_k, x_l) \in \mathcal{Q}$, we know that x_i and x_j are more similar than x_k and x_l. There exists a total of $\mathcal{O}\left(N^4\right)$ quadruplets, but in a practical crowdsourcing application, the available set $\mathcal{Q}$ may only be a subset of all possible quadruplets. We later discuss that single linkage (1) and complete linkage (2) can be easily implemented in this setting. However,

input : Set of objects $\mathcal{X} = \{x_1, \ldots, x_N\}$;
Cluster-level similarity $W : 2^{\mathcal{X}} \times 2^{\mathcal{X}} \to \mathbb{R}$
output: Binary tree, or dendrogram, representing hierarchical clustering of $\mathcal{X}$
begin
Let $\mathcal{B}$ be a collection of N singleton trees $\mathcal{C}_1, \ldots, \mathcal{C}_N$ with root nodes $\mathcal{C}_i.root = \{x_i\}$.
while $|\mathcal{B}| > 1$ **do**
Let $\mathcal{C}, \mathcal{C}'$ be the pair of trees in $\mathcal{B}$ for which $W(\mathcal{C}.root, \mathcal{C}'.root)$ is maximum.
Create $\mathcal{C}''$ with $\mathcal{C}''.root = \{\mathcal{C}.root \cup \mathcal{C}'.root\}$, $\mathcal{C}''.left\text{-}subtree = \mathcal{C}$, and $\mathcal{C}''.right\text{-}subtree = \mathcal{C}'$.
Add $\mathcal{C}''$ to the collection $\mathcal{B}$, and remove $\mathcal{C}, \mathcal{C}'$.
end
return The surviving element in $\mathcal{B}$
end

Algorithm 1: Agglomerative Hierarchical Clustering

such implementations work only when all the possible quadruplets are available. We also present an ordinal variant of average linkage, called Quadruplets based Average Linkage (4–AL), that can work with a partial set of quadruplets.

We also consider a triplet setting, where we have a set

$$\mathcal{T} = \{(x_i, x_k, x_l) : x_i, x_k, x_l \in \mathcal{X}, w_{ik} > w_{il}\},$$

that is, for every ordered tuple $(x_i, x_k, x_l) \in \mathcal{T}$, we know that x_i is more similar to x_k than to x_l. There are $\mathcal{O}(N^3)$ possible triplets, but again a real-world application may provide only a subset $\mathcal{T}$ of all the triplets. We propose an ordinal variant of average linkage based on triplet comparisons, called Triplets based Average Linkage (3–AL). We also study a modification of average linkage that uses a kernel function computed from triplet comparisons (Kleindessner and von Luxburg, 2017a). We call this method Triplet Kernel based Average Linkage (3K–AL). Both algorithms can be used when a partial set of triplets are available.

3 Comparison-Base Hierarchical Clustering

We now describe the ordinal variants of the agglomerative clustering algorithms (1)–(3).

3.1 Comparison-Based Single Linkage (4–SL) and Complete Linkage (4–CL)

We begin with a construction of single and complete linkage algorithms in an ordinal framework. To cast these methods to the comparison-based setting, the first step is to notice that the arg max and arg min functions only depend on quadruplet comparisons. Although it is not possible to compute $W(G, G')$ in (1) or (2), one can retrieve, in each cluster, the pair of objects that achieve the maximum or minimum similarity using only quadruple comparisons. The knowledge of these optimal object pairs is sufficient since our primary aim is to find the pair of clusters G, G' that maximizes $W(G, G')$ and this can be easily achieved through quadruplet comparisons between the optimal object pairs of every G, G'.

The above discussion emphasizes the fact that 4–CL and 4–SL are not applicable when only a subset of the quadruplets is available. Indeed, they require a complete ordering of the similarities to compute and compare the optimal objects of every G, G'. However, it is well known that, when they cannot be actively chosen, all the comparisons are necessary to recover such an ordering (Jamieson and Nowak, 2011a).

3.2 Comparison-Based Average Linkage

Average linkage is difficult to cast to the ordinal framework due to the averaging in (3) that cannot be computed via comparisons. The focus in this section is to derive proxies for the similarity function W in (3). Hence, the generic structure of the algorithms proposed here is the same as Algorithm 1 while we use alternative forms of W that can be computed through quadruplet or triplet comparisons.

4–AL: Quadruplets based Average Linkage. We first consider the setting where we have access to a set of quadruplets $\mathcal{Q}$. We observe that although one cannot compute the average pairwise similarity between two clusters, as in (3), one can estimate the relative similarity between two pairs of clusters. For instance, let G_1, G_2, G_3, G_4 be four clusters such that G_1, G_2 are disjoint and so are G_3, G_4, and define

$$\mathbb{W}_{\mathcal{Q}}\left(G_1, G_2 \| G_3, G_4\right) = \sum_{x_i \in G_1} \sum_{x_j \in G_2} \sum_{x_k \in G_3} \sum_{x_l \in G_4} \frac{\mathbb{I}_{(x_i,x_j,x_k,x_l)\in\mathcal{Q}} - \mathbb{I}_{(x_k,x_l,x_i,x_j)\in\mathcal{Q}}}{|G_1|\,|G_2|\,|G_3|\,|G_4|} \tag{4}$$

where $\mathbb{I}$ is the indicator function. The underlying idea is that clusters G_1, G_2 are more similar to each other than G_3, G_4 if their objects are, on average, more similar to each other than the objects of G_3 and G_4. For example, if all four clusters are singletons $\{x_i\}, \{x_j\}, \{x_k\}, \{x_l\}$, then $\mathbb{W}_{\mathcal{Q}}\left(G_1, G_2 \| G_3, G_4\right)$ is $+1$ if the former two objects have higher similarity than the latter two, and -1 otherwise. One may extend this intuition to clusters with several objects to say that if the similarities between objects in G_1, G_2 are typically higher than the similarities between objects in G_3, G_4, then $\mathbb{W}_{\mathcal{Q}}\left(G_1, G_2 \| G_3, G_4\right)$ tends to be larger. In other words, $\mathbb{W}_{\mathcal{Q}}\left(G_1, G_2 \| G_3, G_4\right) > 0$ suggests that an agglomerative clustering should merge G_1, G_2 before G_3, G_4. Also, note that $\mathbb{W}_{\mathcal{Q}}\left(G_1, G_2 \| G_1, G_2\right) = 0$ and $\mathbb{W}_{\mathcal{Q}}\left(G_1, G_2 \| G_3, G_4\right) = -\mathbb{W}_{\mathcal{Q}}\left(G_3, G_4 \| G_1, G_2\right)$, which hints that (4) provides a preference relation between pairs of clusters.

We use the above preference relation $\mathbb{W}_{\mathcal{Q}}$ to define the similarity function W. Suppose that we have a disjoint partition $G_1, \ldots, G_K$ of $\mathcal{X}$ and that we want to know which clusters should be merged next. We define the similarity of clusters G_p, G_q, $p \neq q$, as

$$W\left(G_p, G_q\right) = \sum_{r,s=1, r\neq s}^{K} \frac{\mathbb{W}_{\mathcal{Q}}\left(G_p, G_q \| G_r, G_s\right)}{K(K-1)}. \tag{5}$$

The underlying idea is that two clusters G_p and G_q are similar to each other if, on average, the pair is often preferred over the other possible cluster pairs. The above similarity measure W, in conjunction with Algorithm 1, results in the proposed 4–AL algorithm.

3–AL: Triplets based Average Linkage. We employ a similar strategy to define W in a triplet scenario, where we are given a set of triplets $\mathcal{T}$. We require an alternative for the preference relation $\mathbb{W}_{\mathcal{Q}}$ since every element $(x_i, x_k, x_l) \in \mathcal{T}$ uses one point, x_i, as a reference with respect to which relative similarity of x_k and x_l are compared. We use a similar flavour, where for any three sets G_1, G_2, G_3 with G_2, G_3 both disjoint from G_1, we define the quantity

$$\mathbb{W}_{\mathcal{T}}\left(G_1, G_2 \| G_1, G_3\right) = \sum_{x_i \in G_1} \sum_{x_k \in G_2} \sum_{x_l \in G_3} \frac{\mathbb{I}_{(x_i,x_k,x_l)\in\mathcal{T}} - \mathbb{I}_{(x_i,x_l,x_k)\in\mathcal{T}}}{|G_1|\,|G_2|\,|G_3|}. \tag{6}$$

The idea is that a cluster G_1 is closer to G_2 than to G_3 if its objects are, on average, closer to the ones of G_2 than to the ones of G_3. Hence, G_1 is used as a reference cluster and (6) provides a preference measuring whether G_1 is more similar to G_2 than to G_3. In particular, $\mathbb{W}\left(G_1, G_2 \| G_1, G_3\right) > 0$ when object pairs in G_1, G_2 tends to be more similar than pairs in G_1, G_3. As in the case of 4–AL, we also have $\mathbb{W}_{\mathcal{T}}(G_1, G_2 \| G_1, G_3) = -\mathbb{W}_{\mathcal{T}}(G_1, G_3 \| G_1, G_2)$, and $\mathbb{W}_{\mathcal{T}}(G_1, G_2 \| G_1, G_2) = 0$. We use this preference relation to define the similarity function W as

$$W(G_p, G_q) = \sum_{r=1, r\neq p,q}^{K} \frac{\mathbb{W}_{\mathcal{T}}(G_p, G_q \| G_p, G_r) + \mathbb{W}_{\mathcal{T}}(G_p, G_q \| G_q, G_r)}{2(K-2)}, \tag{7}$$

assuming that the similarity W is computed when $\mathcal{X}$ is partitioned into $G_1, \ldots, G_K$, and $p \neq q$. The underlying idea is that two clusters G_p, G_q are similar if, on average, they are more similar than any other third cluster. We refer to Algorithm 1 with W given by (7) as the 3–AL hierarchical clustering algorithm.

3K–AL: Triplet Kernel based Average Linkage. In the triplet setting, we discuss an alternative approach based on a comparison-based kernel function proposed by Kleindessner and von Luxburg (2017a). Given a set of triplets $\mathcal{T}$, the kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined as

$$K_{ij} = K(x_i, x_j) = \sum_{\substack{k,l=1\\k<l}}^{N} \frac{\left(\mathbb{I}_{(x_i,x_k,x_l)\in\mathcal{T}} - \mathbb{I}_{(x_i,x_l,x_k)\in\mathcal{T}}\right)}{\sqrt{|\{(x,y,z) \in \mathcal{T} : x = x_i\}|}} \times \frac{\left(\mathbb{I}_{(x_j,x_k,x_l)\in\mathcal{T}} - \mathbb{I}_{(x_j,x_l,x_k)\in\mathcal{T}}\right)}{\sqrt{|\{(x,y,z) \in \mathcal{T} : x = x_j\}|}}. \tag{8}$$

The above kernel function is derived from the Kendall's τ correlation coefficient between two rankings, where the denominator is a normalizing factor.

The 3K–AL method, studied in our work, uses Algorithm 1 with the similarity function W given by

$$W(G, G') = \sum_{x_i \in G, x_j \in G'} \frac{K_{ij}}{|G||G'|}. \tag{9}$$

Here the idea is to use the triplet-based kernel K as a proxy for the similarity score w in (3). We note that empirical performance of hierarchical clustering using the above Kendall's τ kernel has been studied before (Kleindessner and von Luxburg, 2017a), but its theoretical merits are still unknown.

4 Theoretical Analysis

In this section, we provide statistical guarantees for the comparison-based hierarchical clustering algorithms considered in this paper. We first present a planted hierarchical model for the similarity scores $\{w_{ij}\}_{i,j=1,\ldots,N}$. We then provide sufficient conditions under which the agglomerative algorithms 4–SL, 4–CL, 4–AL and 3K–AL exactly recover the planted hierarchy with high probability. We do not present any theoretical guarantees for 3–AL, but later empirically show that its performance is similar to 4–AL.

4.1 Planted Hierarchical Model

We consider a theoretical model where the objects have a true underlying hierarchy, but this latent hierarchy may not be evident due to randomness or noise in the observed similarity scores. The model is similar in spirit to stochastic block models that have been widely used in non-hierarchical clustering. Formally, given N objects we assume that the pairwise similarities $\{w_{ij}\}_{1\leq i<j\leq N}$ are random and mutually independent. Each similarity w_{ij} follows a normal distribution, $w_{ij} \sim$

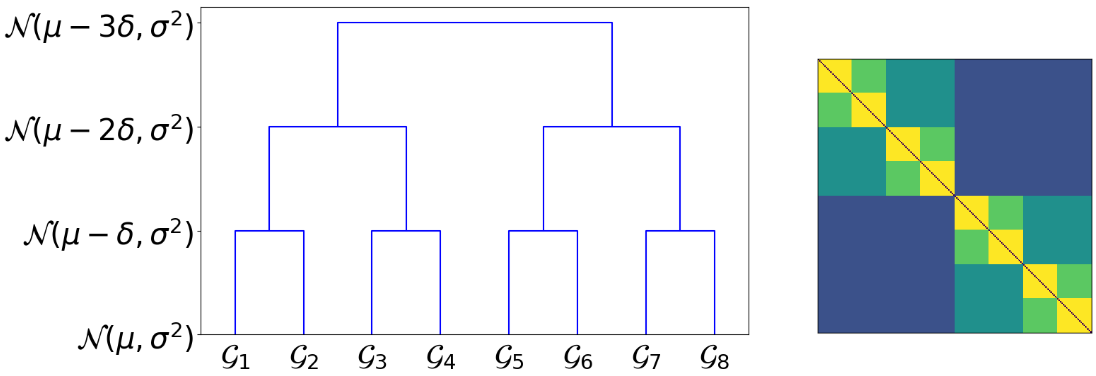


Figure 1: **(Left)** Illustration of the planted hierarchical model for $L = 3$ along with specification of the distributions for similarities at different levels; **(Right)** Hierarchical block structure in expected pairwise similarity matrix, where darker implies smaller similarity.

$\mathcal{N}\left(\mu_{ij}, \sigma^2\right)$, for some fixed variance σ^2 and means μ_{ij} depending on the pair x_i, x_j under consideration. We assume that the similarities are symmetric ($w_{ji} = w_{ij}$ for all $i < j$) and that $w_{ii} = \infty$ for all i.

A hierarchy is introduced among the objects by specifying the means μ_{ij}. Let N be large, and for simplicity be of the form $N = 2^L N_0$ for integers L and N_0. Let μ be a constant and $\delta > 0$. We assume that the expected similarity matrix of μ_{ij}-s has a hierarchical block structure (as depicted in Figure 1) defined as follows:
Step-0**:** $\mathcal{X}$ is divided into two equal sized clusters, and, given x_i and x_j lying in different clusters, their expected similarity is set as $\mu_{ij} = \mu - L\delta$ (dark blue off-diagonal block in Figure 1).
Step-1**:** Each of the two groups is further divided into two sub-groups, and, for each pair x_i, x_j separated due to this sub-group formation, we set $\mu_{ij} = \mu - (L-1)\delta$.
Step-$2, \ldots, L-1$**:** The above process is repeated $L-1$ times, and in step ℓ, the expected similarity across two newly-formed sub-groups is $\mu_{ij} = \mu - (L-\ell)\delta$.
Step-L**:** The above steps result in the formation of 2^L clusters, $\mathcal{G}_1, \ldots, \mathcal{G}_{2^L}$, each of size N_0. The expected similarity between two objects x_i, x_j belonging to the same cluster is $\mu_{ij} = \mu$ (yellow blocks in Figure 1).

The above process may be viewed as a hierarchical tree, where the clusters $\mathcal{G}_1, \ldots, \mathcal{G}_{2^L}$ lie at level-L of the binary tree, and are successively merged at higher levels until we reach level-0 (top level) that corresponds to the entire set $\mathcal{X}$ (see Figure 1). The pairwise similarity gets smaller in expectation, when two objects are merged higher in the true hierarchy. The above model is similar to the general stochastic block model (Amini and Levina, 2018), but is different in terms of:
— Gaussian pairwise similarities instead of Bernoullis, which aid in a comparison-based framework, and
— hierarchical block structure in the expected matrix.

4.2 Analysis of Agglomerative Algorithms

We now analyze different algorithms under the above model assuming that all triplet or quadruplet comparisons are available. We provide sufficient conditions under which the aforementioned algorithms exactly recover the underlying hierarchy with high probability. In this scenario, exact recovery involves two aspects:
— correct recovery of all the pure clusters, $\mathcal{G}_1, \ldots, \mathcal{G}_{2^L}$,
— recovery of the entire hierarchy among the clusters.

The complexity of the model, which governs the guarantees for exact recovery, is characterized by (i) δ, the separation between the expected similarities across different levels, (ii) σ, the standard deviation of the similarities, and (iii) $N_0 = \frac{N}{2^L}$, the size of the pure clusters. In fact, $\frac{\delta}{\sigma}$ is the signal-to-noise ratio that is crucial for recovering a planted model, whereas a large N_0 ensures that, for a given N, the algorithm does not have to estimate very small clusters or a long hierarchy.

We first present sufficient conditions for the single and complete linkage algorithms, 4–SL and 4–CL. We give a sketch of proof here and details in Appendix A.

Theorem 1 (Exact recovery by 4–SL and 4–CL). *Let $\eta \in (0,1)$. If $\frac{\delta}{\sigma} \geq 4\sqrt{\ln(N/\eta)}$, and all quadruplet comparisons are available in $\mathcal{Q}$, then 4–SL and 4–CL exactly recover the planted hierarchy with probability $1-\eta$.*

Proof sketch. A simple way to achieve exact recovery is to ensure that the similarities w_{ij} at different levels of the hierarchy follow the ordering of the respective means. In particular, every pair of objects or clusters are merged only at levels designated by the planted hierarchy if $|w_{ij} - \mu_{ij}| < \frac{\delta}{2}$ for all $i \neq j$. The condition on $\frac{\delta}{\sigma}$ ensures that this holds with high probability. □

Theorem 1 implies that a sufficient condition for exact recovery by single/complete linkage is that the signal-to-noise ratio grows as $\sqrt{\ln N}$ with the number of examples. This is a very strong requirement that raises the question of whether one can do better with average linkage algorithms. It is the case for 4–AL and 3K–AL. For subsequent discussion, let $\beta = 2\Phi\left(\frac{\delta}{\sqrt{2}\sigma}\right) - 1 > 0$, with $\Phi(\cdot)$ the standard normal distribution function.

Theorem 2 (Exact recovery by 4–AL). *Let $\eta \in (0,1)$. There is a constant $C > 0$ such that if:*
1. $N_0 > c \ln N$ for $c = C \max\left\{\ln\left(\frac{1}{\eta}\right), \frac{\sigma^2}{\delta^2} e^{L^2\delta^2/4\sigma^2}\right\}$,
2. an initialization step partitions $\mathcal{X}$ into pure clusters of sizes between $c\ln N$ to $2c\ln N$, and
3. all quadruplet comparisons are available in $\mathcal{Q}$,
then starting from the initial partition, 4–AL exactly recovers the planted hierarchy with probability $1-\eta$.
In particular, if N is sufficiently large, and δ, σ are fixed, 4–AL exactly recovers the planted hierarchy with high probability if $L \ll \sqrt{\ln N}$.

Proof sketch. We observe that if clusters G_1, G_2 must be merged before G_3, G_4 in the planted hierarchy, then $\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4) \geq \beta$ in expectation. Similarly, it is less than $-\beta$ if G_1, G_2 must be merged later, and zero if there is no preference. These effects get aggregated in the expectation of $W(G, G')$ in (5), and one can show that, at any stage, the maximum of $\mathbf{E}[W(G, G')]$ is achieved by a pair of clusters that is merged first under the planted model.

To extend this conclusion to the random instances of $\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4)$, and hence $W(G, G')$, we use a concentration result for sums of dependent random variables based on the notion of equitable coloring (Janson and Ruciński, 2002). The assumptions aid in the concentration by ensuring that sufficiently many terms are summed in (4). Assumption-3 in the theorem can be possibly relaxed to the case of observing an uniformly sampled subset of quadruplets. □

Theorem 2 shows that 4–AL can recover the true hierarchy even when the signal-to-noise ratio is much less than that prescribed for 4–SL and 4–CL in Theorem 1. However, this guarantee comes with an additional requirement of an initial partitioning of $\mathcal{X}$ into $\mathcal{O}(\ln N)$ sized pure clusters. This is not a particularly strong assumption as it is often seen that average linkage methods require an initial clustering step to achieve near optimal results (Cohen-Addad et al., 2018). We finally present the result for the 3K–AL algorithm, which uses the kernel (8) in conjunction with average linkage.

Theorem 3 (Exact recovery by 3K–AL). *Let $\eta \in (0,1)$. There is a constant $C > 0$ such that if $N_0 > C\frac{8^L}{\beta^4}\ln\left(\frac{N}{\eta}\right)$, and $\mathcal{T}$ contains all valid triplets, then 3K–AL exactly recovers the planted hierarchy with probability $1-\eta$.*
If N is sufficiently large, and δ, σ are fixed, the above condition on N_0 holds for $L \ll \ln N$.

Proof sketch. The proof consists of two steps. First, we analyze the expected kernel matrix $\mathbf{E}[K]$. It is obvious that $\mathbf{E}[K]$ has a block structure that corresponds to the underlying hierarchy, and hence, there are exactly $L+1$ distinct off-diagonal entries in $\mathbf{E}[K]$. Each distinct value of $\mathbf{E}[K_{ij}]$ corresponds to the level at which x_i and x_j are merged. We show that the difference between these values is at least $\beta^2 N_0^2/N^2$.

We next show that the random values of K_{ij} concentrate around $\mathbf{E}[K_{ij}]$ for all $i \neq j$. The concentration involves a sum of dependent random variables, and thus, we can again use the techniques of Janson and Ruciński (2002). Finally, the condition on N_0 ensures that the deviation of K_{ij} is smaller than half of the separation between distinct values of $\mathbf{E}[K_{ij}]$. □

Theorem 3 shows that the sufficient criteria for exact recovery by 3K–AL is nearly similar to that of 4–AL, but the initial clustering is no longer needed since the agglomerative stage relies on a kernel function instead of the raw comparisons as in (4)–(5). The results also show that the sufficient conditions for 4–AL and 3K–AL are considerably better than the conditions of single and complete linkage.

Remark. Our discussion has focused on exact recovery of the planted hierarchy. But one can easily extend these results to show that under the planted model and aforementioned assumptions, the above methods provide $(1+o(1))$-optimal solutions for Dasgupta's cost function (Dasgupta, 2016).

5 Experiments

In this section we evaluate our approaches on several problems: we verify the theoretical findings of Section 4, we compare our methods to ordinal embedding based approaches on standard datasets, and we illustrate their behaviour on a triplet-based dataset.

5.1 Planted Hierarchical Model

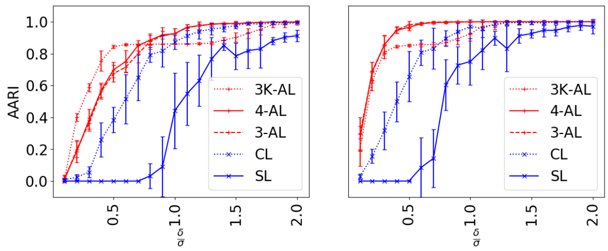
To verify the theoretical findings from Section 4 we use the planted hierarchical model presented in Section 4.1 to generate data and study the performance of the various methods (4–SL, 4–CL, 4–AL, 3–AL, 3K–AL).

Recall that the generative model depends on the mean similarity within each cluster μ, the variance σ^2, the separability constant δ, the depth of the planted partition L and the number of examples in each cluster N_0. From the theoretical analysis it is clear that the hardness of the problem depends on the signal-to-noise ratio $\frac{\delta}{\sigma}$. Hence, to study the behaviour of the different methods when this quantity changes, we set $\mu = 0.8$, $\sigma = 0.1$, $N_0 = 30$, and $L = 3$ and we vary $\delta \in \{0.01, 0.02, \ldots, 0.2\}$. The guarantees of 4–AL also depend on pure initial clusters of sufficient sizes m for exact recovery and thus we considered $m \in \{1, 2, 3, 5, 10\}$. These initial clusters were obtained by uniformly sampling without replacement from the N_0 examples contained in each of the 2^L ground-truth clusters. As a measure of performance we report the Averaged Adjusted Rand Index (AARI) between the ground truth $\mathcal{C}$ and the hierarchies $\mathcal{C}'$ learned by the different methods. Let $\mathcal{C}^\ell$ and $\mathcal{C}'^\ell$ be the partitions of $\mathcal{X}$ at level ℓ of the hierarchies, then:

$$\mathrm{AARI}\left(\mathcal{C}, \mathcal{C}'\right) = \frac{1}{L} \sum_{\ell \in \{1, \ldots, L\}} \mathrm{ARI}\left(\mathcal{C}^\ell, \mathcal{C}'^\ell\right)$$

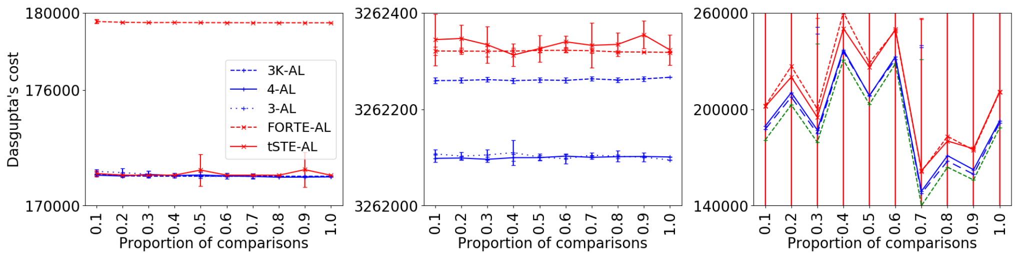
where ARI is the Adjusted Rand Index (Hubert and Arabie, 1985), a widely used measure to compare partitions. We use the average across the different levels $\mathcal{C}^\ell$ and $\mathcal{C}'^\ell$ to take into account the hierarchical structure. The AARI takes values in the interval $[0, 1]$ and the higher the value the more similar the hierarchies are. $\mathrm{AARI}\left(\mathcal{C}, \mathcal{C}'\right) = 1$ implies that the two hierarchies are identical. For all the experiments we report the mean and the standard deviation of 10 repetitions.

In Figure 2 we present the results for $m = 1$ and $m = 3$. We defer the other results to the supplementary material. Firstly, similar to the theory, 4–SL and 4–CL can recover the planted



(a) $m = 1$ (b) $m = 3$

Figure 2: AARI of the proposed methods (higher is better) on data obtained from the planted hierarchical model with $\mu = 0.8$, $\sigma = 0.1$, $L = 3$, $N_0 = 30$. In Figure 2a the methods get no initial clusters ($m = 1$) and in Figure 2b they get initial clusters of size $m = 3$. Best viewed in color.



(a) Zoo (b) Glass (c) 20news

Figure 3: Dasgupta's score (lower is better) of the different methods on the Zoo, Glass and 20news datasets. The embedding dimension for FORTE–AL and tSTE–AL is 2 for all the datasets. Best viewed in color.

hierarchy only for very large values of $\frac{\delta}{\sigma}$. Moreover, 4–AL and 3–AL have similar performance and, given small initial clusters, they are able to exactly recover the true hierarchy even when $\frac{\delta}{\sigma} < 1$. Finally, without initial clusters ($m = 1$ case), 3K–AL performs significantly better than the other methods for very small values of $\frac{\delta}{\sigma}$.

5.2 Standard Clustering Datasets

In this second set of experiments we compare our methods to two baselines that use ordinal embedding as a first step.

Baselines. We consider two ordinal embedding methods, t-STE (van der Maaten and Weinberger, 2012) and FORTE (Jain et al., 2016), followed by a standard average linkage approach using a cosine similarity as the base metric (tSTE-AL and FORTE-AL). These two methods are parametrized by the embedding dimension d. Since it is often difficult to automatically tune parameters in clustering (because of the lack of ground-truth) we considered several embedding dimensions and reported the best results in the main paper. We also reported the results for the other embedding dimensions in Appendix B.

Datasets. We consider 3 different datasets commonly used in hierarchical clustering: Zoo, Glass and 20news (Heller and Ghahramani, 2005; Vikram and Dasgupta, 2016). The Zoo dataset is composed of 100 animals with 16 features. The Glass dataset has 9 features for 214 examples. The 20news dataset is composed of 11314 news articles. Following Vikram and Dasgupta (2016) we pre-processed the 20news dataset using a bag of words approach followed by PCA to retain 100 relevant features. We randomly sampled 200 examples for hierarchical clustering.

To fit the comparison-based setting we generated some triplets and quadruplets using the cosine similarity:

$$w_{ij} = \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\|\mathbf{x}_i\| \, \|\mathbf{x}_i\|}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the representations of objects x_i and x_j and $\langle \cdot, \cdot \rangle$ is the dot product. As mentioned earlier in the paper it is not realistic to assume that all the comparisons are available and, in the experiments, we used a random subset of $p\%$ comparisons, where p varies between 10% and 100%. Since 4–CL and 4–SL are not applicable when we only have a subset of comparisons, we did not consider them.

Evaluation Function. Contrary to the planted hierarchical model we do not have access to a ground-truth hierarchy and thus we cannot use the AARI measure to evaluate the performance of the different methods. Instead we use a recently proposed cost (Dasgupta, 2016) that has been specifically designed to evaluate hierarchical clustering methods (Dasgupta's cost). Given a base similarity measure w, the cost of a hierarchy $\mathcal{C}$ is

$$\mathrm{cost}(\mathcal{C}, w) = \sum_{x, x_j \in \mathcal{X}} w_{ij} \left| \mathcal{C}^{lca}(x_i, x_j) \right|$$

where w_{ij} is the similarity between x_i and x_j and $\mathcal{C}^{lca}(x_i, x_j)$ is the smallest cluster containing both x_i and x_j in the hierarchy. A low cost is achieved if similar objects (high w_{ij}) are merged towards the bottom of the tree (small $\mathcal{C}^{lca}(x_i, x_j)$), and vice-versa. Hence, a lower value of the cost indicates a better hierarchy. For all the experiments we report the mean and the standard deviation of 10 repetitions.

Results. We report the results in Figure 3. First, notice that the proportion of comparisons does not have a large impact on the results which are, on average, the same for all regimes. Our approaches consistently outperform the embedding based methods on Glass and 20news as they do not need to first embed the examples and thus do not impose a rather strong euclidean structure to the data. On Zoo, tSTE–AL is competitive when the embedding dimension is correctly chosen. In the appendix, we report results for higher embedding dimensions that show that a poor choice can worsen the results of tSTE–AL and FORTE–AL.

5.3 Triplet-based Dataset

In the final experiment, we consider a comparison-based dataset. The car dataset (Kleindessner and von Luxburg, 2017b) is composed of 60 different type of cars and 12112 triplet comparisons. Since there is no known underlying ground-truth or similarity it is difficult to quantitatively evaluate the quality of the hierarchies. Nevertheless, to visually demonstrate the capabilities of the different methods, we reported in the supplementary material the dendrograms obtained by 3–AL, 3K–AL, FORTE–AL and tSTE–AL (4–AL is not applicable to this triplet data set).

6 Conclusion

We investigated the problem of hierarchical clustering in a comparison based setting. We showed that single and complete linkage algorithms (4–SL and 4–CL) could be easily used in this setting, and we proposed three new approaches based on average linkage (3–AL, 4–AL, and 3K–AL). We studied the performance of the algorithms under a planted hierarchical model. We showed that 4–SL and 4–CL can exactly recover the planted hierarchy if the signal-to-noise ratio grows with the number of examples, whereas 4–AL and 3K–AL can, under reasonable assumptions, exactly recover the hierarchy even for constant ratios. Empirically we confirmed our theoretical findings and compared our methods to two ordinal embedding based baselines.

References

Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., and Belongie, S. (2007). Generalized non-metric multidimensional scaling. In *International Conference on Artificial Intelligence and Statistics*, pages 11–18.

Amid, E. and Ukkonen, A. (2015). Multiview triplet embedding: Learning attributes in multiple maps. In *International Conference on Machine Learning*, pages 1472–1480.

Amini, A. A. and Levina, E. (2018). On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1):149–179.

Arias-Castro, E. et al. (2017). Some theory for ordinal embedding. *Bernoulli*, 23(3):1663–1693.

Borg, I. and Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications.* Springer.

Chaudhuri, K., Dasgupta, S., Kpotufe, S., and von Luxburg, U. (2014). Consistent procedures for cluster tree estimation and pruning. *IEEE Transactions on Information Theory*, 60(12):7900–7912.

Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., and Mathieu, C. (2018). Hierarchical clustering: Objective functions and algorithms. In *Symposium on Discrete Algorithms*, pages 378–397.

Dasgupta, S. (2016). A cost function for similarity-based hierarchical clustering. In *Symposium on Theory of Computing*, pages 118–127. ACM.

Emamjomeh-Zadeh, E. and Kempe, D. (2018). Adaptive hierarchical clustering using ordinal queries. In *Symposium on Discrete Algorithms*. ACM.

Haghiri, S., Garreau, D., and von Luxburg, U. (2018). Comparison-based random forests. In *International Conference on Machine Learning.*

Haghiri, S., Ghoshdastidar, D., and von Luxburg, U. (2017). Comparison-based nearest neighbor search. In *International Conference on Artificial Intelligence and Statistics*, pages 851–859.

Hartigan, J. A. (1981). Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374):388–394.

Heikinheimo, H. and Ukkonen, A. (2013). The crowd-median algorithm. In *AAAI Conference on Human Computation and Crowdsourcing.*

Heller, K. A. and Ghahramani, Z. (2005). Bayesian hierarchical clustering. In *International conference on Machine learning*, pages 297–304. ACM.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Jain, L., Jamieson, K. G., and Nowak, R. (2016). Finite sample prediction and recovery bounds for ordinal embedding. In *Advances in Neural Information Processing Systems*, pages 2711–2719.

Jamieson, K. G. and Nowak, R. (2011a). Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems*, pages 2240–2248.

Jamieson, K. G. and Nowak, R. D. (2011b). Low-dimensional embedding using adaptively selected ordinal data. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 1077–1084. IEEE.

Janson, S. and Ruciński, A. (2002). The infamous upper tail. *Random Structures & Algorithms*, 20(3):317–342.

Kazemi, E., Chen, L., Dasgupta, S., and Karbasi, A. (2018). Comparison based learning from weak oracles. In *International Conference on Artificial Intelligence and Statistics.*

Kleindessner, M. and Luxburg, U. (2014). Uniqueness of ordinal embedding. In *Conference on Learning Theory*, pages 40–67.

Kleindessner, M. and Luxburg, U. (2015). Dimensionality estimation without distances. In *International Conference on Artificial Intelligence and Statistics*, pages 471–479.

Kleindessner, M. and von Luxburg, U. (2017a). Kernel functions based on triplet similarity comparisons. In *Advances in Neural Information Processing Systems.*

Kleindessner, M. and von Luxburg, U. (2017b). Lens depth function and k-relative neighborhood graph: versatile tools for ordinal data analysis. *The Journal of Machine Learning Research*, 18(1):1889–1940.

Moseley, B. and Wang, J. (2017). Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search. In *Advances in Neural Information Processing Systems 30*, pages 3097–3106.

Roy, A. and Pokutta, S. (2016). Hierarchical clustering via spreading metrics. In *Advances in Neural Information Processing Systems*, pages 2316–2324.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140.

Stewart, N., Brown, G. D., and Chater, N. (2005). Absolute identification by relative judgment. *Psychological review*, 112(4):881.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., and Kalai, A. T. (2011). Adaptively learning the crowd kernel. In *International Conference on Machine Learning*, pages 673–680. Omnipress.

Terada, Y. and von Luxburg, U. (2014). Local ordinal embedding. In *International Conference on Machine Learning*, pages 847–855.

Ukkonen, A. (2017). Crowdsourced correlation clustering with relative distance comparisons. *arXiv preprint arXiv:1709.08459.*

Ukkonen, A., Derakhshan, B., and Heikinheimo, H. (2015). Crowdsourced nonparametric density estimation using relative distances. In *AAAI Conference on Human Computation and Crowdsourcing.*

van der Maaten, L. and Weinberger, K. (2012). Stochastic triplet embedding. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE.

Vikram, S. and Dasgupta, S. (2016). Interactive bayesian hierarchical clustering. In *International Conference on Machine Learning*, pages 2081–2090.

Young, F. W. (1987). *Multidimensional scaling: History, theory, and applications.* Lawrence Erlbaum Associates.

Zhang, L., Maji, S., and Tomioka, R. (2015). Jointly learning multiple measures of similarities from triplet comparisons. *arXiv preprint arXiv:1503.01521.*

A Proofs of the Theoretical Results

We provide detailed proofs of Theorems 1–3. We recall some of the key quantities associated with the planted model, which include (i) N, the number of objects; (ii) L, the number of levels in the hierarchy; (iii) $N_0 = \frac{N}{2^L}$, the size of the pure clusters; (iv) δ, the separation between the expected similarities across consecutive levels; and (v) σ, the standard deviation of the similarities.

We also define $\ell_{ij}^{lca} = \ell^{lca}(x_i, x_j)$ as the level of the ground truth tree in which the least common ancestor (lca) of x_i and x_j resides. We extend this definition to the level of lca of two clusters G, G', denoted by $\ell^{lca}(G, G')$. If G, G' are both subsets of the same pure cluster, we write $\ell^{lca}(G, G') = L$. Hence, the range of ℓ^{lca} is $\{0, 1, \ldots, L\}$. Also, let $\beta_\ell = 2\Phi\left(\frac{\ell\delta}{\sqrt{2}\sigma}\right) - 1$ for an integer $\ell \in [-L, L]$, where Φ denotes the standard normal distribution function. Note that $\beta_0 = 0$ and $\beta_1 = \beta$ mentioned in the paper.

A.1 Exact recovery of planted hierarchy by Single Linkage (4–SL) and Complete Linkage (4–CL)

Theorem 1. *Let $\eta \in (0, 1)$. If $\frac{\delta}{\sigma} \geq 4\sqrt{\ln\left(\frac{N}{\eta}\right)}$, and all quadruplet comparisons are available in $\mathcal{Q}$, then 4–SL and 4–CL exactly recover the planted hierarchy with probability $1 - \eta$.*

Proof. Let $Z \sim \mathcal{N}(0, 1)$. It can be easily verified that $\mathbf{P}(|Z| \geq t) \leq \sqrt{\frac{2}{\pi}}\frac{1}{t}\exp(-0.5t^2)$. For $t \geq 1$, we may simply bound this by $\exp(-0.5t^2)$. Now, observe that for every $i \neq j$, $\frac{w_{ij} - \mu_{ij}}{\sigma} \sim \mathcal{N}(0, 1)$. Using this, we can write

$$\mathbf{P}\left(\bigcup_{i \neq j}\left\{|w_{ij} - \mu_{ij}| \geq \frac{\delta}{2}\right\}\right) \leq \sum_{i \neq j}\mathbf{P}\left(|Z| \geq \frac{\delta}{2\sigma}\right) \leq N^2 \exp\left(-\frac{\delta^2}{8\sigma^2}\right)$$

since $\delta > 2\sigma$ under stated condition. The above probability is smaller than η for $\delta \geq 4\sigma\sqrt{\ln(\frac{N}{\eta})}$. Thus, under the stated condition, $|w_{ij} - \mu_{ij}| < \frac{\delta}{2}$ for all $i \neq j$. We now show that the above scenario leads to exact recovery of the hierarchy by single or complete linkage clustering. Note that

$$\mathbf{E}[w_{ij}] = \mu_{ij} = \mu - (L - \ell_{ij}^{lca})\delta$$

Due to the concentration of the similarity score we have,

$$w_{ij} \in \left(\mu - (L - \ell_{ij}^{lca})\delta - \frac{\delta}{2}, \mu - (L - \ell_{ij}^{lca})\delta + \frac{\delta}{2}\right),$$

for all $i \neq j$ with probability $1 - \eta$. Thus, the similarity scores corresponding to the different levels of the ground truth do not overlap, and this ensures that the agglomerative algorithms merge objects or clusters in same order as prescribed by the ground truth. For instance, at the first stage, where the goal is to extract the pure clusters, we have $w_{ij} > \mu - \frac{\delta}{2}$ if x_i, x_j belong to the same pure cluster, and $w_{ij} < \mu - \frac{\delta}{2}$ otherwise. Hence, both single and complete linkage merge objects in the same cluster first before merging objects from different clusters. The same argument also holds for the subsequent levels as well. Hence, the result. □

A.2 Exact recovery of planted hierarchy by Quadruplets based Average Linkage (4–AL)

Theorem 2. *Let $\eta \in (0, 1)$. There is a constant $C > 0$ such that if:*
1. $N_0 > c\ln N$ for $c = C\max\left\{\ln\left(\frac{1}{\eta}\right), \frac{\sigma^2}{\delta^2}e^{L^2\delta^2/4\sigma^2}\right\}$,
2. an initialization step partitions $\mathcal{X}$ into pure clusters of sizes between $c\ln N$ to $2c\ln N$, and

3. all quadruplet comparisons are available in $\mathcal{Q}$,
then starting from the initial partition, 4–AL exactly recovers the planted hierarchy with probability $1-\eta$.
In particular, if N is sufficiently large, and δ, σ are fixed, 4–AL exactly recovers the planted hierarchy with high probability if $L \ll \sqrt{\ln N}$.

Proof. Let $\varphi = \min\limits_{\ell \in [-L+1, L]} (\beta_\ell - \beta_{\ell-1})$, where β_ℓ is as defined in the beginniing of the appendix. Observe that the minimum occurs at $\ell = L$ with $\varphi > \frac{1}{\sqrt{\pi}} \frac{\delta}{\sigma} e^{-L^2\delta^2/4\sigma^2}$. We first analyze the algorithm under expectation. The subsequent reasoning is a technical version of the proof sketch given in the main paper. Assume that at some stage of the agglomerative iterations, we have a partition $G_1, \ldots, G_K$ of $\mathcal{X}$. Assume that the partition adheres to the ground truth, that is, either each G_p is a subset of a pure cluster or an union of several pure clusters that corresponds to one of the nodes in the top L levels of the true hierarchy. Consider $p, q, r, s \in \{1, \ldots, K\}$ such that $p \neq q$, $r \neq s$, $\ell^{lca}(G_p, G_q) = \ell$ and $\ell^{lca}(G_r, G_s) = \ell'$. From the definition of $\mathbb{W}_{\mathcal{Q}}$, we have

$$\begin{aligned}\mathbf{E}[\mathbb{W}_{\mathcal{Q}}\left(G_p, G_q \| G_r, G_s\right)] &= \sum_{x_i \in G_p} \sum_{x_j \in G_q} \sum_{x_k \in G_r} \sum_{x_l \in G_s} \frac{\mathbf{P}(w_{ij} > w_{kl}) - \mathbf{P}(w_{ij} < w_{kl})}{|G_p|\,|G_q|\,|G_r|\,|G_s|} \\ &= \sum_{x_i \in G_p} \sum_{x_j \in G_q} \sum_{x_k \in G_r} \sum_{x_l \in G_s} \frac{2\mathbf{P}(w_{ij} > w_{kl}) - 1}{|G_p|\,|G_q|\,|G_r|\,|G_s|}\end{aligned}$$

since $\mathbf{P}(w_{ij} = w_{kl}) = 0$ as both scores are normal distributed. Observe that $\mu_{ij} = \mu - (L-\ell)\delta$ for every $i \in G_p, j \in G_q$, and $\mu_{kl} = \mu - (L-\ell')\delta$ for every $k \in G_r, l \in G_s$. Hence, for every i, j, k, l, we can use two independent standard normal random variables Z, Z' to write

$$\begin{aligned}\mathbf{P}(w_{ij} > w_{kl}) &= \mathbf{P}\left(\mu_{ij} + \sigma Z > \mu_{kl} + \sigma Z'\right) \\ &= \mathbf{P}\left(Z' - Z < \frac{(\ell - \ell')\delta}{\sigma}\right) = \Phi\left(\frac{(\ell-\ell')\delta}{\sqrt{2}\sigma}\right)\end{aligned} \tag{10}$$

since $(Z' - Z) \sim \mathcal{N}(0, 2)$. Based on the above relation, we have

$$\mathbf{E}[\mathbb{W}_{\mathcal{Q}}\left(G_p, G_q \| G_r, G_s\right)] = \beta_{\ell - \ell'}\,,$$

which justifies that $\mathbf{E}[\mathbb{W}_{\mathcal{Q}}\left(G_p, G_q \| G_r, G_s\right)] \geq \beta_1$ whenever $\ell^{lca}(G_p, G_q) > \ell^{lca}(G_r, G_s)$, as stated in the main paper. Now, consider $p, q, p', q' \in \{1, \ldots, K\}$ such that $p \neq q$, $p' \neq q'$, $\ell^{lca}(G_p, G_q) = \ell$ and $\ell^{lca}(G_{p'}, G_{q'}) = \ell - 1$ for some $\ell \geq 1$. Thus, according to the planted model, one should merge G_p, G_q before $G_{p'}, G_{q'}$. We verify that this is indeed the case under expectation since

$$\begin{aligned}\mathbf{E}[W(G_p, G_q)] - &\mathbf{E}[W(G_{p'}, G_{q'})] \\ &= \frac{1}{K(K-1)} \sum_{\substack{r,s=1 \\ r \neq s}}^{K} \mathbf{E}[\mathbb{W}_{\mathcal{Q}}\left(G_p, G_q \| G_r, G_s\right)] - \mathbf{E}[\mathbb{W}_{\mathcal{Q}}\left(G_{p'}, G_{q'} \| G_r, G_s\right)] \\ &= \frac{1}{K(K-1)} \sum_{\substack{r,s=1 \\ r \neq s}}^{K} \beta_{\ell - \ell^{lca}(G_r, G_s)} - \beta_{\ell - 1 - \ell^{lca}(G_r, G_s)} \\ &\geq \varphi\end{aligned}$$

since each term in the summation is larger than φ. Chaining of the above argument shows that $\mathbf{E}[W(G_p, G_q)] - \mathbf{E}[W(G_{p'}, G_{q'})] \geq \varphi$ whenever $\ell^{lca}(G_p, G_q) > \ell^{lca}(G_{p'}, G_{q'})$. Under the assumptions stated in Theorem 2, we later prove that with probability $1 - \eta$,

$$|W(G, G') - \mathbf{E}[W(G, G')]| < \frac{\varphi}{2} \tag{11}$$

for every pair of clusters G, G' formed during the agglomerative steps of the algorithm starting from the $\mathcal{O}(\ln N)$ sized pure clusters. Based on (11) and the above argument, it is evident that $W(G_p, G_q) > W(G_{p'}, G_{q'})$ whenever $\ell^{lca}(G_p, G_q) > \ell^{lca}(G_{p'}, G_{q'})$ and, in particular, the cluster pair that achieves the maximum at any stage of iteration must be merged at the earliest according to the planted hierarchy. This guarantees exact recovery of the planting by the algorithm.

We now prove (11). For this, we first state a concentration inequality that we prove later. Let G_1, G_2, G_3, G_4 be four clusters, each of size at least $c \ln N$, such that G_1, G_2 are disjoint and so are G_3, G_4. Here, $c = C \max \left\{ \ln\left(\frac{1}{\eta}\right), \frac{\sigma^2}{\delta^2} e^{L^2\delta^2/4\sigma^2} \right\}$ for some absolute constant $C > 0$, as specified in Theorem 2. Then

$$\mathbf{P}\left(|\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4) - \mathbf{E}[\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4)]| \geq \frac{\varphi}{2} \right) \leq 2N^{2-(\varphi^2 c^2 \ln N)/128}. \tag{12}$$

We wish to use (12) to argue that with probability $1-\eta$, all clusters in the initial partition (obtained from the initialisation in the second condition) satisfy

$$|\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4) - \mathbf{E}[\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4)]| < \frac{\varphi}{2}.$$

Note that we do not know how the initial partition is achieved, but we can ensure that

$$\begin{aligned}
&\mathbf{P}\left(\exists G_1, G_2, G_3, G_4 : c \ln N \leq |G_1|, |G_2|, |G_3|, |G_4| \leq 2c \ln N, \right. \\
&\qquad\qquad\qquad\qquad \left. |\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4) - \mathbf{E}[\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4)]| \geq \frac{\varphi}{2} \right) \\
&\qquad\qquad \leq \sum_{i_1, i_2, i_3, i_4 = c \ln N}^{2c \ln N} \binom{N}{i_1}\binom{N}{i_2}\binom{N}{i_3}\binom{N}{i_4} 2N^{2-(\varphi^2 c^2 \ln N)/128} \\
&\qquad\qquad \leq (c \ln N)^4 \left(\frac{eN}{c \ln N} \right)^{8c \ln N} 2N^{2-(\varphi^2 c^2 \ln N)/128} \leq N^{(15-\varphi^2 c^2/128) c \ln N}
\end{aligned}$$

since $N > c \ln N > e$. Choosing $C > 0$ large enough we can have the above probability bounded by $N^{-(c \ln N)/2}$. The latter is bounded by η for $c \geq 2 \ln(\frac{1}{\eta})$. Thus, with probability $1-\eta$, we know that for every tuple of four clusters, obtained at initialization, $\mathbb{W}_{\mathcal{Q}}$ deviates from its mean by at most $\frac{\varphi}{2}$. In fact, the same deviation also holds when we merge some of these clusters. For instance, let G_1, G_1', G_2, G_3, G_4 be part of a partition at some stage and suppose G_1, G_1' are merged. Then

$$\mathbb{W}_{\mathcal{Q}}(G_1 \cup G_1', G_2 \| G_3, G_4) = \frac{|G_1|}{|G_1| + |G_1'|} \mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4) + \frac{|G_1'|}{|G_1| + |G_1'|} \mathbb{W}_{\mathcal{Q}}(G_1', G_2 \| G_3, G_4),$$

which is a convex combination of $\mathbb{W}_{\mathcal{Q}}$ computed at the previous stage. Hence, if each of them deviates from mean by at most $\frac{\varphi}{2}$, then the convex combination after merging also deviates from mean by at most $\frac{\varphi}{2}$. The same also holds for other instances of merging throughout the hierarchy, which shows that with probability $1-\eta$, at any stage of agglomeration, we have $|\mathbb{W}_{\mathcal{Q}}(G_p, G_q \| G_r, G_s) - \mathbf{E}[\mathbb{W}_{\mathcal{Q}}(G_p, G_q \| G_r, G_s)]| < \frac{\varphi}{2}$ for any tuple of four clusters in the partition. Now, observe that $W(G_p, G_q)$ is an average of several $\mathbb{W}_{\mathcal{Q}}$, and so, (11) holds.

We complete the proof of Theorem 2 by proving the concentration inequality in (12). Since $\{w_{ij} = w_{kl}\}$ occurs with zero probability for any $i, j, k, l (i \neq j, k \neq l)$, we may write

$$\begin{aligned}
&|\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4) - \mathbf{E}[\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4)]| \\
&\quad = \frac{2}{|G_1|\,|G_2|\,|G_3|\,|G_4|} \left| \sum_{x_i \in G_1} \sum_{x_j \in G_2} \sum_{x_k \in G_3} \sum_{x_l \in G_4} \left(\mathbb{I}_{(x_i, x_j, x_k, x_l) \in \mathcal{Q}} - \mathbf{P}(w_{ij} > w_{kl}) \right) \right|.
\end{aligned}$$

Note that each term in the summation is a centred random variable in the range $[-1, 1]$. Let us denote each of them by B_{ijkl}, and observe that they have dependencies among themselves. We use the technique described in Section 2.3.2 of (Janson and Ruciński, 2002). Consider the dependency graph for these variables, which is a graph on $m = |G_1||G_2||G_3||G_4|$ vertices and two vertices are adjacent if they are dependent. Some of the vertices have degree $|G_1||G_2| - 1$ (dependent with other variables with same k, l), while other vertices have degree $|G_3||G_4| - 1$. Let us denote the maximum degree by d. One can find an equitable colouring for such a graph using $(d+1)$ colours, where equitable denotes that all colour classes are of nearly equal sizes $\lfloor \frac{m}{d+1} \rfloor$ or $\lceil \frac{m}{d+1} \rceil$. Denoting the colour classes by $\mathcal{C}_1, \ldots, \mathcal{C}_{d+1}$, we can bound the probability using union bound and Hoeffding's inequality as

$$\begin{aligned}
\mathbf{P}\left(|\mathbb{W}_{\mathcal{Q}}(G_1, G_2\|G_3, G_4) - \mathbf{E}[\mathbb{W}_{\mathcal{Q}}(G_1, G_2\|G_3, G_4)]| > \frac{\varphi}{2}\right) &= \mathbf{P}\left(\left|\sum_{i,j,k,l} B_{ijkl}\right| > \frac{m\varphi}{4}\right) \\
&\leq \sum_{\ell=1}^{d+1} \mathbf{P}\left(\left|\sum_{(i,j,k,l)\in\mathcal{C}_\ell} B_{ijkl}\right| > \frac{m\varphi}{4(d+1)}\right) \\
&\leq \sum_{\ell=1}^{d+1} 2\exp\left(-\frac{\frac{\varphi^2 m^2}{16(d+1)^2}}{4|\mathcal{C}_\ell|}\right).
\end{aligned}$$

The bound in (12) follows by first noting that $|\mathcal{C}_\ell| \leq \frac{2m}{d+1}$, and then using the fact $\frac{m}{d+1} \geq \min\{|G_1||G_2|, |G_3||G_4|\} \geq (c \ln N)^2$. For the outer summation, we simply use $(d+1) \leq N^2$ to obtain the bound in (12).

The final claim in the theorem for large N can be easily derived by comparing the definition on $N_0 = \frac{N}{2^L}$ with the first assumption that $N_0 \gg e^{L^2\delta^2/4\sigma^2} \ln N$. Taking logarithm leads to the stated criterion. $\square$

A.3 Exact recovery of planted hierarchy by Triplet Kernel based Average Linkage (3K–AL)

Theorem 3. *Let $\eta \in (0, 1)$. There is a constant $C > 0$ such that if $N_0 > C\frac{8^L}{\beta_1^4} \ln\left(\frac{N}{\eta}\right)$, and $\mathcal{T}$ contains all valid triplets, then 3K–AL exactly recovers the planted hierarchy with probability $1-\eta$. If N is sufficiently large, and δ, σ are fixed, the above condition on N_0 holds for $L \ll \ln N$.*

Proof. We begin by computing the expected kernel matrix given that all the triplets are observed. Recall that

$$\begin{aligned}
K_{ij} = K(x_i, x_j) &= \sum_{\substack{k,l=1\\k<l}}^{N} \frac{\left(\mathbb{I}_{(x_i,x_k,x_l)\in\mathcal{T}} - \mathbb{I}_{(x_i,x_l,x_k)\in\mathcal{T}}\right)}{\sqrt{|\{(x, y, z \in \mathcal{T} : x = x_i\}|}} \frac{\left(\mathbb{I}_{(x_j,x_k,x_l)\in\mathcal{T}} - \mathbb{I}_{(x_j,x_l,x_k)\in\mathcal{T}}\right)}{\sqrt{|\{(x, y, z) \in \mathcal{T} : x = x_j\}|}} \\
&= \frac{1}{\binom{N}{2}} \sum_{\substack{k,l=1\\k<l}}^{N} \left(\mathbb{I}_{(x_i,x_k,x_l)\in\mathcal{T}} - \mathbb{I}_{(x_i,x_l,x_k)\in\mathcal{T}}\right)\left(\mathbb{I}_{(x_j,x_k,x_l)\in\mathcal{T}} - \mathbb{I}_{(x_j,x_l,x_k)\in\mathcal{T}}\right), \qquad (13)
\end{aligned}$$

where we note that if all the triplets are observed, then for every $x \in \mathcal{X}$, the set of valid comparisons $\mathcal{T}$ contains exactly $\binom{N}{2}$ terms with x as pivot. Hence, for $i \neq j$, the expectation can be computed as

$$\begin{aligned}
\mathbf{E}[K_{ij}] &= \frac{1}{\binom{N}{2}} \sum_{\substack{k,l=1\\k<l}}^{N} \left(\mathbf{P}(w_{ik} > w_{il}) - \mathbf{P}(w_{ik} < w_{il})\right)\left(\mathbf{P}(w_{jk} > w_{jl}) - \mathbf{P}(w_{jk} < w_{il})\right) \\
&= \frac{1}{2\binom{N}{2}} \sum_{\substack{k,l=1\\l\neq k}}^{N} \left(2\mathbf{P}(w_{ik} > w_{il}) - 1\right)\left(2\mathbf{P}(w_{jk} > w_{jl}) - 1\right)
\end{aligned}$$

$$= \frac{1}{2\binom{N}{2}} \Bigg[\sum_{k \neq i,j} 2 \left(2\mathbf{P}(w_{ij} > w_{jk}) + 2\mathbf{P}(w_{ij} > w_{ik}) - 2 \right) \tag{14}$$

$$+ \sum_{k \neq i,j} \sum_{l \neq i,j,k} \left(2\mathbf{P}(w_{ik} > w_{il}) - 1 \right) \left(2\mathbf{P}(w_{jk} > w_{jl}) - 1 \right) \Bigg]$$

$$= \frac{1}{2\binom{N}{2}} \Bigg[\sum_{k \neq i,j} 2 \left(\beta_{\ell^{lca}_{ij} - \ell^{lca}_{jk}} + \beta_{\ell^{lca}_{ij} - \ell^{lca}_{ik}} \right) + \sum_{k \neq i,j} \sum_{l \neq i,j,k} \beta_{\ell^{lca}_{ik} - \ell^{lca}_{il}} \beta_{\ell^{lca}_{jk} - \ell^{lca}_{jl}} \Bigg]. \tag{15}$$

The second line follows by noting that each term is symmetric in k, l due to the product, and hence we may sum over all $k \neq l$. We also use the fact that $\mathbf{P}(w_{ik} = w_{il}) = 0$ for all $k \neq l$. In the third step, we separate the terms where either k or l takes the values i, j. Finally, we use the derivation in (10) to express each term using β_ℓ.

Recall that, under the planted hierarchy, $\mathcal{X}$ is partitioned in pure clusters $\mathcal{G}_1, \ldots, \mathcal{G}_{2^L}$. We abuse notation to write $\mathcal{G}_r$ as the set $\{i : x_i \in \mathcal{G}_r\}$. In (15), observe that each term in the sum depends only on the groups containing i, j, k, l, and hence, we may only compute it for each group and multiply by the number of terms in the group. If $i, j \in \mathcal{G}_1$, then k can take only $(N_0 - 2)$ values in $\mathcal{G}_1$, and N_0 values in other groups (similar for l). We may do the entire computation only at group level, and then use a multiplicative factor of $(1 \pm \epsilon)$ with $\epsilon = \frac{4}{N_0}$ to account for fluctuations in the number of terms from each group. Here, $\mathbf{E}[K_{ij}] = (1 \pm \epsilon)a$ denotes $(1-\epsilon)a \leq \mathbf{E}[K_{ij}] \leq (1+\epsilon)a$.

Thus, for i, j such that $i \neq j$ and $\ell^{lca}_{ij} = \ell$, we have

$$\begin{aligned}
&N(N-1)\mathbf{E}[K_{ij}] \\
&= (1 \pm \epsilon) \Bigg[N_0 \sum_{r=1}^{2^L} 2 \left(\beta_{\ell^{lca}_{ij} - \ell^{lca}(j, \mathcal{G}_r)} + \beta_{\ell^{lca}_{ij} - \ell^{lca}(i, \mathcal{G}_r)} \right) + N_0^2 \sum_{r,s=1}^{2^L} \beta_{\ell^{lca}(i,\mathcal{G}_r) - \ell^{lca}(i,\mathcal{G}_s)} \beta_{\ell^{lca}(j,\mathcal{G}_r) - \ell^{lca}(j,\mathcal{G}_s)} \Bigg] \\
&= (1 \pm \epsilon) \Bigg[2N_0 \sum_{p=0}^{L} \beta_{\ell - p} \#\{r : \ell^{lca}(i, \mathcal{G}_r) = p\} + 2N_0 \sum_{p'=0}^{L} \beta_{\ell - p'} \#\{r : \ell^{lca}(j, \mathcal{G}_r) = p'\} \\
&\quad + N_0^2 \sum_{p,q,p',q'=0}^{L} \beta_{p-q} \beta_{p'-q'} \#\{r : \ell^{lca}(i, \mathcal{G}_r) = p, \ell^{lca}(j, \mathcal{G}_r) = p'\} \#\{s : \ell^{lca}(i, \mathcal{G}_s) = q, \ell^{lca}(j, \mathcal{G}_s) = q'\} \Bigg],
\end{aligned} \tag{16}$$

where the second equality explicitly mentions that we need to count the number of different pure clusters that are merged with i or j under the true hierarchy. We now consider different cases. First, if i, j belong to same group, then $\ell = L$ and for every r, s, $\ell^{lca}(i, \mathcal{G}_r) = \ell^{lca}(j, \mathcal{G}_r)$ and $\ell^{lca}(i, \mathcal{G}_s) = \ell^{lca}(j, \mathcal{G}_s)$. So,

$$\begin{aligned}
&N(N-1)\mathbf{E}[K_{ij}] \\
&= (1 \pm \epsilon) \Bigg[4N_0 \sum_{p=0}^{L} \beta_{L-p} \#\{r : \ell^{lca}(i, \mathcal{G}_r) = p\} + N_0^2 \sum_{p,q=0}^{L} \beta_{p-q}^2 \#\{r : \ell^{lca}(i, \mathcal{G}_r) = p\} \#\{s : \ell^{lca}(i, \mathcal{G}_s) = q\} \Bigg] \\
&= (1 \pm \epsilon) \Bigg[4N_0 \sum_{p=0}^{L} \beta_{L-p} (2^{L-1-p} \vee 1) + N_0^2 \sum_{p,q=0}^{L} \beta_{p-q}^2 (2^{L-1-p} \vee 1)(2^{L-1-q} \vee 1) \Bigg] \\
&=: \kappa_L
\end{aligned} \tag{17}$$

where $\vee$ denotes max. The numbers of clusters are computed based on the fact that there is only one cluster merged at levels L or $L-1$, and otherwise 2^{L-1-p} groups are merged with i at level-p. We call this term κ_L since it does not depend on i, j or the group they belong to.

If i, j are not in same group, that is, $\ell = \ell^{lca}_{ij} < L$, then we observe that for the terms involving one β-term, the computation is similar to κ_L and we get $4N_0 \sum_{p=0}^{L} \beta_{\ell-p}(2^{L-1-p} \vee 1)$. For the terms quadratic in β-s, we observe:

- if $p < \ell$, then for any $\mathcal{G}_r$ such that $\ell^{lca}(i, \mathcal{G}_r) = p$, we also have $\ell^{lca}(j, \mathcal{G}_r) = p$. So we may only consider cases $p = p'$ when $p < \ell$.

- there is no $\mathcal{G}_r$ such that $\ell^{lca}(i,\mathcal{G}_r) = \ell^{lca}(j,\mathcal{G}_r) = \ell$ which happens because the hierarchy is a binary tree and $\mathcal{G}_r$ must either merge first with i or with j (so, we don't need to consider $p = p' = \ell$). Note that this is the main difference from the case $\ell = L$.
- if $p > \ell$, then for any $\mathcal{G}_r$ with $\ell^{lca}(i,\mathcal{G}_r) = p$, we have $\ell^{lca}(j,\mathcal{G}_r) = \ell$. So we may set $p' = \ell$ whenever we have $p > \ell$. Similarly, we should also count the cases $p' > \ell, p = \ell$.

The above facts also hold for q, q'. Thus, we can decompose the summation into 9 parts based on three conditions on p, p' ($p = p' < \ell$; $p > \ell, p' = \ell$; $p = \ell, p' > \ell$) and similar three on q, q'. Out of these 9 cases, two terms ($p = q = \ell$; $p' = q' = \ell$) do not contribute since they involve a term $\beta_{\ell-\ell} = \beta_0 = 0$. For the other 7 cases, we should count $\#\{r : \ell^{lca}(i,\mathcal{G}_r) = p, \ell^{lca}(j,\mathcal{G}_r) = p'\}$ and $\#\{s : \ell^{lca}(i,\mathcal{G}_s) = q, \ell^{lca}(j,\mathcal{G}_s) = q'\}$. To compute these, we note that when $p = p' < \ell$, $\#\{r : \ell^{lca}(i,\mathcal{G}_r) = \ell^{lca}(j,\mathcal{G}_r) = p\} = 2^{L-1-p} \vee 1$ as used above. But when $p = \ell, p' > \ell$, we have $\#\{r : \ell^{lca}(i,\mathcal{G}_r) = \ell, \ell^{lca}(j,\mathcal{G}_r) = p'\} = 2^{L-1-p'} \vee 1$ since this counts only those groups which merge with i at level-p', and p plays no role in the count. The same also holds for counting $\#\{s : \ell^{lca}(i,\mathcal{G}_s) = q, \ell^{lca}(j,\mathcal{G}_s) = q'\}$ for different q, q'. Hence, we compute the second summation in (16) as

$$\begin{aligned}
N_0^2 \Bigg[&\sum_{p,q=0}^{\ell-1} \beta_{p-q}^2 (2^{L-1-p} \vee 1)(2^{L-1-q} \vee 1) + 2 \sum_{p=0}^{\ell-1} \sum_{q=\ell+1}^{L} \beta_{p-q}\beta_{p-\ell} (2^{L-1-p} \vee 1)(2^{L-1-q} \vee 1) \\
&+ 2 \sum_{p=\ell+1}^{L} \sum_{q=0}^{\ell-1} \beta_{p-q}\beta_{\ell-q} (2^{L-1-p} \vee 1)(2^{L-1-q} \vee 1) \\
&+ 2 \sum_{p,q=\ell+1}^{L} \beta_{p-\ell}\beta_{\ell-q} (2^{L-1-p} \vee 1)(2^{L-1-q} \vee 1) \Bigg].
\end{aligned}$$

We changed indices from p', q' to p, q to express the 7 terms as four terms (the last three computed twice). Note that the above expectation depends only on ℓ, and hence we call it as κ_ℓ, which can be simplified further by observing that the second and third terms are identical if p, q are interchanged (we use the negation property of $\beta_{-r} = \beta_r$). Also, $2^{L-1-p} \geq 1$ when $p < \ell$. Hence, for $\ell = \ell_{ij}^{lca} < L$, we write $\mathbf{E}[K_{ij}]$ as

$$\begin{aligned}
\kappa_\ell &:= N(N-1)\mathbf{E}[K_{ij}] \\
&= (1 \pm \epsilon) 4N_0 \sum_{p=0}^{L} \beta_{\ell-p} (2^{L-1-p} \vee 1) + (1 \pm \epsilon) N_0^2 \Bigg[\sum_{p,q=0}^{\ell-1} 2^{2L-2-p-q} \beta_{p-q}^2 \\
&\qquad + \sum_{p=0}^{\ell-1} \sum_{q=\ell+1}^{L} 2^{L+1-p} (2^{L-1-q} \vee 1) \beta_{q-p}\beta_{\ell-p} - \sum_{p,q=\ell+1}^{L} 2(2^{L-1-p} \vee 1)(2^{L-1-q} \vee 1)\beta_{p-\ell}\beta_{q-\ell} \Bigg].
\end{aligned} \tag{18}$$

Observe that, similar to κ_L in (17), the above quantity does not depend on i, j, or the groups they belong to. It simply depends on the fact that $\ell_{ij}^{lca} = \ell$. It will subsequently help to note that for every ℓ

$$\begin{aligned}
\kappa_\ell &\leq 8N_0(L+1)2^L\beta_L + 2N_0^2(L+1)^2 2^{2L}\beta_L^2 \\
&\leq 2^{3(L+2)} N_0^2
\end{aligned}$$

for $N_0 > 1$ and $L > 1$. Here, we use the fact that $\beta_L < 1$.

The above discussion shows that the expected kernel matrix $\mathbf{E}[K]$ has exactly $(L+1)$ distinct values given by $\frac{\kappa_\ell}{N(N-1)}$ for $\ell = 0, \ldots, L$, where κ_ℓ is given in (17) or (18). We next show that these distinct values are considerably separated from each other, that is $\kappa_{\ell+1} - \kappa_\ell$ is sufficiently large for every ℓ. To compute such a lower bound, one may initially ignore the ϵ-factor during subtraction, and then subtract another term of $2\epsilon \left(2^{3(L+2)} N_0^2\right)$, which accounts for the approximation. We begin with a lower bound for $\kappa_L - \kappa_{L-1}$ as

$$\kappa_L - \kappa_{L-1}$$

$$
\begin{aligned}
&> 4N_0 \sum_{p=0}^{L} (\beta_{L-p} - \beta_{L-1-p})(2^{L-1-p} \vee 1) + N_0^2 \left[2\beta_1^2 + \sum_{p=0}^{L-2} 2^{L-p} (\beta_{L-p} - \beta_{L-1-p})^2 \right] - 2\epsilon \left(2^{3(L+2)} N_0^2 \right) \\
&> 2N_0^2 \beta_1^2 - N_0 2^{3(L+3)}
\end{aligned}
$$

The first inequality follows the lower bounding approach mentioned above. We note that the β^2-terms where both $p, q \leq L-2$ cancel out and the remaining ones, that is for p or q larger than $L-2$ simplify as given in the second line. The last step follows by observing that both terms involving summation are positive and may be ignored. We also use $\epsilon = \frac{4}{N_0}$. Finally, we use the assumption on N_0, which implies $8^L < \frac{N_0 \beta_1^4}{C \ln N}$, to claim that $\kappa_L - \kappa_{L-1} > N_0^2 \beta_1^2$. We now bound the difference between $\kappa_{\ell+1} - \kappa_\ell$ for $\ell \leq L-2$, which can be written as

$$
\begin{aligned}
\kappa_{\ell+1} - \kappa_\ell > 4N_0 & \sum_{p=0}^{L} (\beta_{\ell+1-p} - \beta_{\ell-p})(2^{L-1-p} \vee 1) - 2\epsilon \left(2^{3(L+2)} N_0^2 \right) \\
& + N_0^2 \left[2 \sum_{p=0}^{\ell-1} 2^{2L-2-\ell-p} \beta_{\ell-p}^2 + \sum_{q=\ell+2}^{L} 2^{L+1-\ell} (2^{L-1-q} \vee 1) \beta_{q-\ell} \beta_1 \right. \\
& \quad - \sum_{p=0}^{\ell-1} 2^{L+1-p} 2^{L-2-\ell} \beta_{\ell+1-p} \beta_{\ell-p} \\
& \quad + \sum_{p=0}^{\ell-1} \sum_{q=\ell+2}^{L} 2^{L+1-p} (2^{L-1-q} \vee 1) \beta_{q-p} (\beta_{\ell+1-p} - \beta_{\ell-p}) \\
& \quad + \sum_{p,q=\ell+2}^{L} 2(2^{L-1-p} \vee 1)(2^{L-1-q} \vee 1)(\beta_{p-\ell}\beta_{q-\ell} - \beta_{p-\ell-1}\beta_{q-\ell-1}) \\
& \quad \left. + 2(2^{L-2-\ell})^2 \beta_1^2 + 2 \sum_{q=\ell+2}^{L} 2(2^{L-2-\ell})(2^{L-1-q} \vee 1) \beta_1 \beta_{q-\ell} \right]
\end{aligned}
$$

Here, the first term is due the linear β-term in (18) and the second term takes care of the ϵ-factor. Among the β^2-terms, the first one arises from the first β^2-term in (18), where the surviving terms are for $q = \ell, p < \ell$ or vice-versa. The next three terms come from the difference of the second β^2-term for κ_ℓ and $\kappa_{\ell+1}$ — the first among these is because $p = \ell$ appears only for $\kappa_{\ell+1}$, the second is because $q = \ell + 1$ appears only for κ_ℓ, and the third is the difference between terms present in both. The last three β^2-terms arise from the last term in (18) — the first of these accounts for difference when $p, q \geq \ell + 2$, the next takes care of $p = q = \ell + 1$ (which appears only for κ_ℓ) and the last is for the cases $p = \ell + 1, q > \ell + 1$ or vice-versa.

It is easy to see that the fifth β^2-term is positive since $\beta_{p-\ell}\beta_{q-\ell} > \beta_{p-\ell-1}\beta_{q-\ell-1}$. We show that the sum of first, third and fourth β^2-terms is also positive. For this, we write the sum as

$$
\begin{aligned}
& \sum_{p=0}^{\ell-1} \left[2^{2L-1-\ell-p} \beta_{\ell-p}^2 - 2^{2L-1-\ell-p} \beta_{\ell+1-p} \beta_{\ell-p} + \sum_{q=\ell+2}^{L} 2^{L+1-p} (2^{L-1-q} \vee 1) \beta_{q-p} (\beta_{\ell+1-p} - \beta_{\ell-p}) \right] \\
& > \sum_{p=0}^{\ell-1} \beta_{\ell-p} (\beta_{\ell+1-p} - \beta_{\ell-p}) 2^{-p} \left[-2^{2L-1-\ell} + \sum_{q=\ell+2}^{L} 2^{L+1} (2^{L-1-q} \vee 1) \right] \\
& = \sum_{p=0}^{\ell-1} \beta_{\ell-p} (\beta_{\ell+1-p} - \beta_{\ell-p}) 2^{-p} \left[-2^{2L-1-\ell} + 2^{L+1} + \sum_{q=\ell+2}^{L-1} 2^{2L-q} \right]
\end{aligned}
$$

Computing the geometric progression, we get that the term within the bracket is zero. Hence, the overall expression is positive. Note that in the second step, we use $\beta_{q-p} > \beta_{\ell-p}$ since $q \geq \ell + 2$. Thus, only the second, sixth and seventh β^2-terms in the difference remain, which are all positive, and after computing the geometric progression in second and seventh terms, we get a lower bound

$$\kappa_{\ell+1} - \kappa_\ell > -2\epsilon\left(2^{2L+1}N_0\right)^2 + N_0^2 2^{2L-2\ell-2}\beta_1^2 > -N_0 2^{3(L+3)} + 4N_0^2\beta_1^2$$

since $\ell \leq L-2$. Thus, under condition on N_0, we have $\kappa_{\ell+1} - \kappa_\ell > N_0^2\beta_1^2$ for all $\ell = 0, \ldots, L-1$, and hence, the distinct off-diagonal entries of $\mathbf{E}[K]$ are separated at least by $\frac{N_0^2\beta_1^2}{N(N-1)}$.

We now use concentration inequality to show that with probability $1-\eta$, we have $|K_{ij} - \mathbf{E}[K_{ij}]| < \frac{N_0^2\beta_1^2}{2N(N-1)}$ for all $i \neq j$. This implies that all random entries of K corresponding to different levels of hierarchy in the ground truth tree are non-overlapping. Hence, one can simply use the arguments in the proof of Theorem 1 to show that average linkage (or even single/complete linkage) recovers the planted hierarchy. For the concentration, we show that under the condition on N_0, for any $i \neq j$,

$$\mathbf{P}\left(|K_{ij} - \mathbf{E}[K_{ij}]| > \frac{N_0^2\beta_1^2}{2N(N-1)}\right) \leq \frac{\eta}{N^2}. \tag{19}$$

By union bound, this implies that the bound on deviation holds for all $i \neq j$ with probability at least $1-\eta$. To prove (19), we first define

$$\begin{aligned} B_{kl} = &\left(\mathbb{I}_{(x_i,x_k,x_l\in\mathcal{T})} - \mathbb{I}_{(x_i,x_l,x_k\in\mathcal{T})}\right)\left(\mathbb{I}_{(x_j,x_k,x_l\in\mathcal{T})} - \mathbb{I}_{(x_j,x_l,x_k\in\mathcal{T})}\right) \\ &- \left(2\mathbf{P}(w_{ik} > w_{il}) - 1\right)\left(2\mathbf{P}(w_{jk} > w_{jl}) - 1\right) \end{aligned}$$

for every $k, l \in \{1, \ldots, N\}$, $k \neq l$. Note that $K_{ij} - \mathbf{E}[K_{ij}] = \frac{1}{N(N-1)} \sum\limits_{k,l:k\neq l} B_{kl}$, which is a sum of centered random variables, each lying in the range $[-2, 2]$. However, $\{B_{kl} : k \neq l\}$ are dependent, and we again use the equitable colouring based approach for concentration (Janson and Ruciński, 2002). Consider the dependency graph among the $N^2 - N$ random variables in the summation. We have that degree of each vertex is $4N-3$ since B_{kl} has edge with all $B_{k'l'}$ where k or l is in first or second index ($4N$ terms) other than itself and the two diagonal terms. Thus, we can find a $(4N-2)$-equitable colouring with largest colour class being of size at most $\lceil\frac{N^2-N}{4N-2}\rceil$, which gives

$$\begin{aligned} \mathbf{P}\left(|K_{ij} - \mathbf{E}[K_{ij}]| > \frac{N_0^2\beta_1^2}{2N(N-1)}\right) &\leq \sum_{p=1}^{4N-2} \mathbf{P}\left(\left|\sum_{(k,l)\in\mathcal{C}_p} B_{kl}\right| > \frac{N_0^2\beta_1^2}{2(4N-2)}\right) \\ &\leq \sum_{p=1}^{4N-2} 2\exp\left(-\frac{\frac{N_0^4\beta_1^4}{4(4N-2)^2}}{4^2\,\frac{2(N^2-N)}{(4N-2)}}\right) \leq 8N\exp\left(-\frac{N_0^4\beta_1^4}{2^9N^3}\right). \end{aligned}$$

Using $\frac{N_0^3}{N^3} = \frac{1}{8^L}$ and the condition $\frac{N_0\beta_1^4}{8^L} > C\ln(\frac{N}{\eta})$ leads to the bound in (19) if $C > 0$ is chosen large enough. □

B Additional Experiments

In this section we present additional plots and discussions that were not included in the main paper.

B.1 Planted Hierarchical Model

In Figure 4 we present supplementary results for the planted hierarchical model. As suggested by the theory, 4–AL strongly benefits from bigger initial clusters achieving perfect recovery for signal-to-noise ratio $\frac{\delta}{\sigma}$ as low as 0.3 when $m = 10$. 3–AL performs very similarly. On the other hand, the effect of initial clustering has very little effect on the performance of 4–SL, 4–CL or 3K–AL. Without initial clustering, that is $m = 1$, 3K–AL outperforms all other methods for small values of $\frac{\delta}{\sigma}$. This clearly demonstrates the potential of the Kendall's τ kernel function in the regime of small signal-to-noise ratio. However, as $\frac{\delta}{\sigma}$ increases, 4–AL or 3–AL eventually prove to be a better choice. The reason for this effect could be theoretically investigated. It is also evident that the average linkage based methods (4–AL, 3–AL and 3K–AL) clearly perform better than 4–SL and 4–CL, and 4–SL cannot achieve exact recovery even for relatively large $\frac{\delta}{\sigma}$. However, 4–CL has a slightly better performance than 4–SL, which is not evident from the theoretical result. This suggests that a better sufficient condition might be possible for 4–CL.

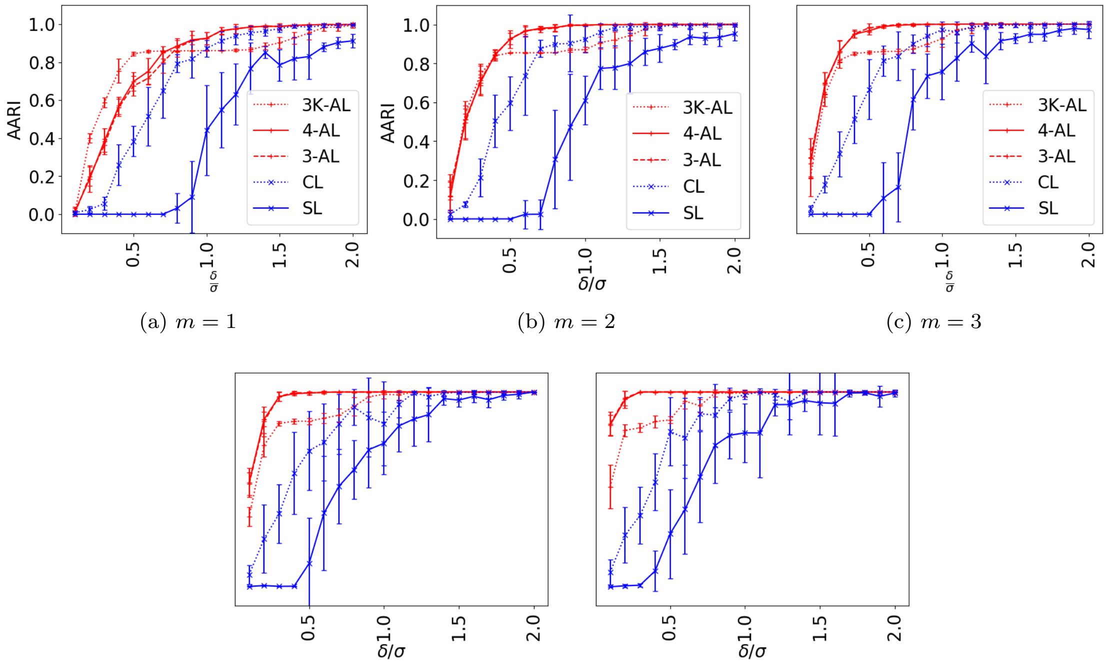
B.2 Standard Clustering Datasets

In Figures 5, 6, and 7 we present supplementary results for the standard clustering dataset. The performance of tSTE-AL and FORTE-AL depends on the dimension that should be carefully chosen. Unfortunately, in clustering, tuning parameters can be difficult as there is no ground-truth. We point out that in Figure 5 the plot for 3K–AL is not clearly visible since its performance is close to that of 4–AL and 3–AL.

B.3 Triplet-based Dataset

The car dataset (Kleindessner and von Luxburg, 2017b) is composed of 60 different type of cars and 6056 ordinal comparisons, collected via crowd-sourcing, of the form *Which car is most central among the three x_i, x_j and x_k?*. These statements translate easily to the triplet setting: if x_i is most central in the set of three then we recover two triplets (x_j, x_i, x_k) and (x_k, x_i, x_j). Overall we obtained 12112 triplet comparisons that we used to learn a hierarchy among the cars.

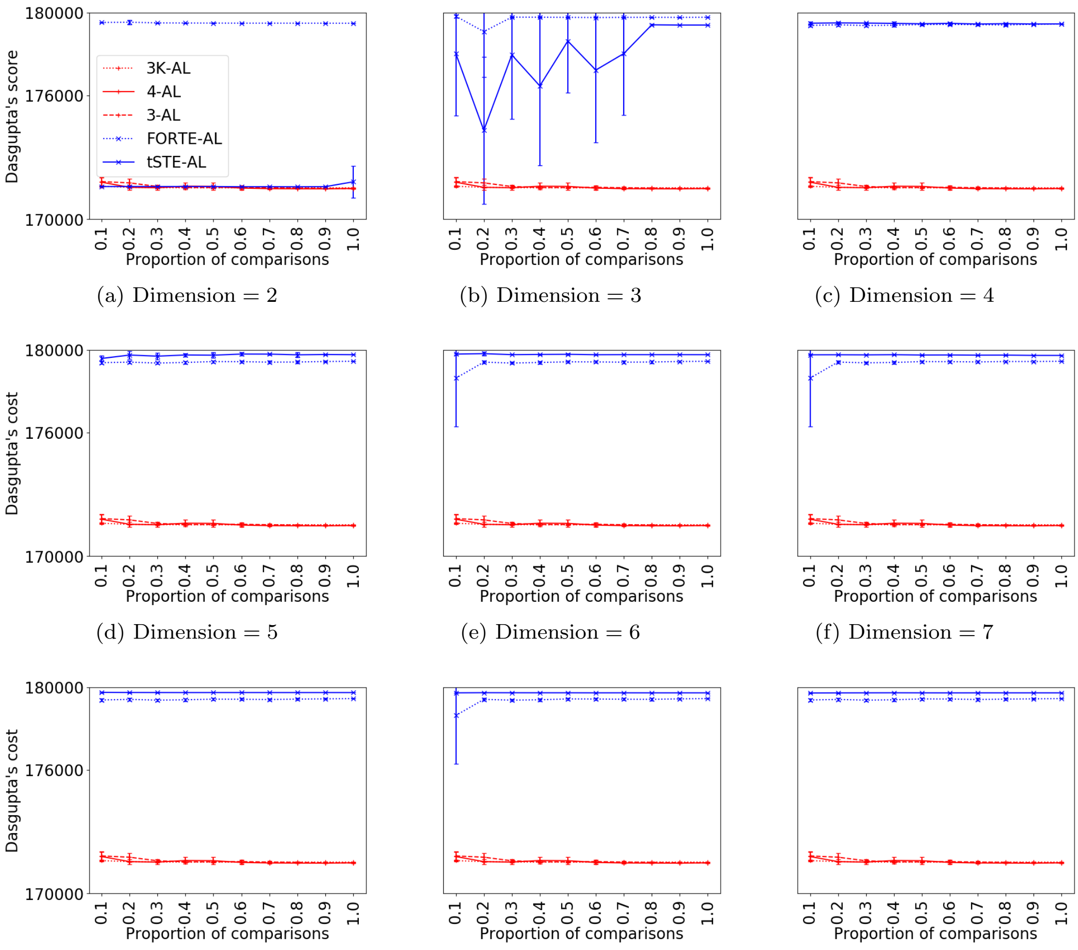
The hierarchies obtained by 3–AL, 3K–AL, FORTE–AL and tSTE–AL are attached to this supplementary as png files. The names of the files are respectively cars.3–AL.png, cars.3K–AL.png, cars.FORTE–AL.embedding_dimension.png and cars.tSTE–AL.embedding_dimension.png.



(a) $m = 1$ (b) $m = 2$ (c) $m = 3$

(d) $m = 5$ (e) $m = 10$

Figure 4: AARI of the proposed methods (higher is better) on data obtained from the planted hierarchical model with $\mu = 0.8$, $\sigma = 0.1$, $L = 3$, $N_0 = 30$ and different values of m, the size of the initial clusters. Best viewed in color.

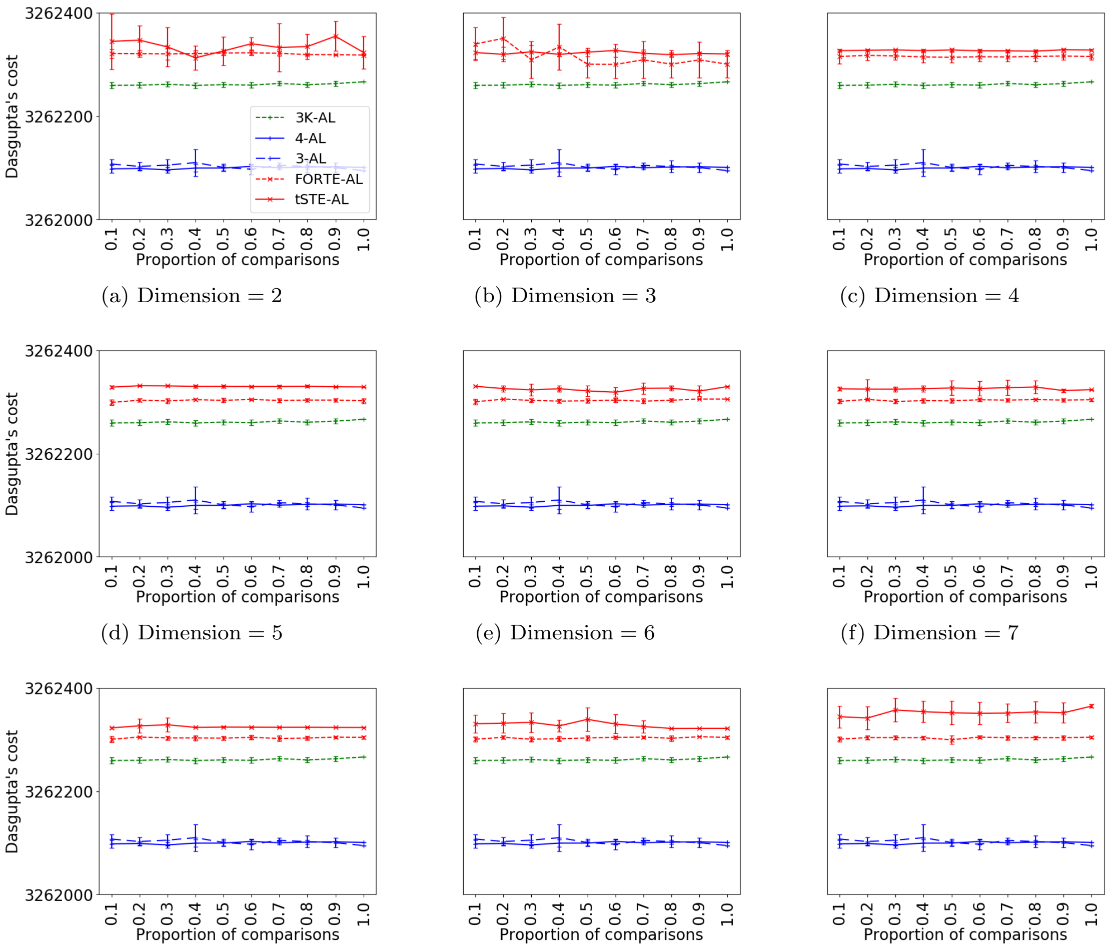


(a) Dimension = 2 (b) Dimension = 3 (c) Dimension = 4

(d) Dimension = 5 (e) Dimension = 6 (f) Dimension = 7

(g) Dimension = 8 (h) Dimension = 9 (i) Dimension = 10

Figure 5: Dasgupta's score of the different methods on the Zoo dataset with increasing embedding dimensions for FORTE–AL and tSTE–AL. Best viewed in color.

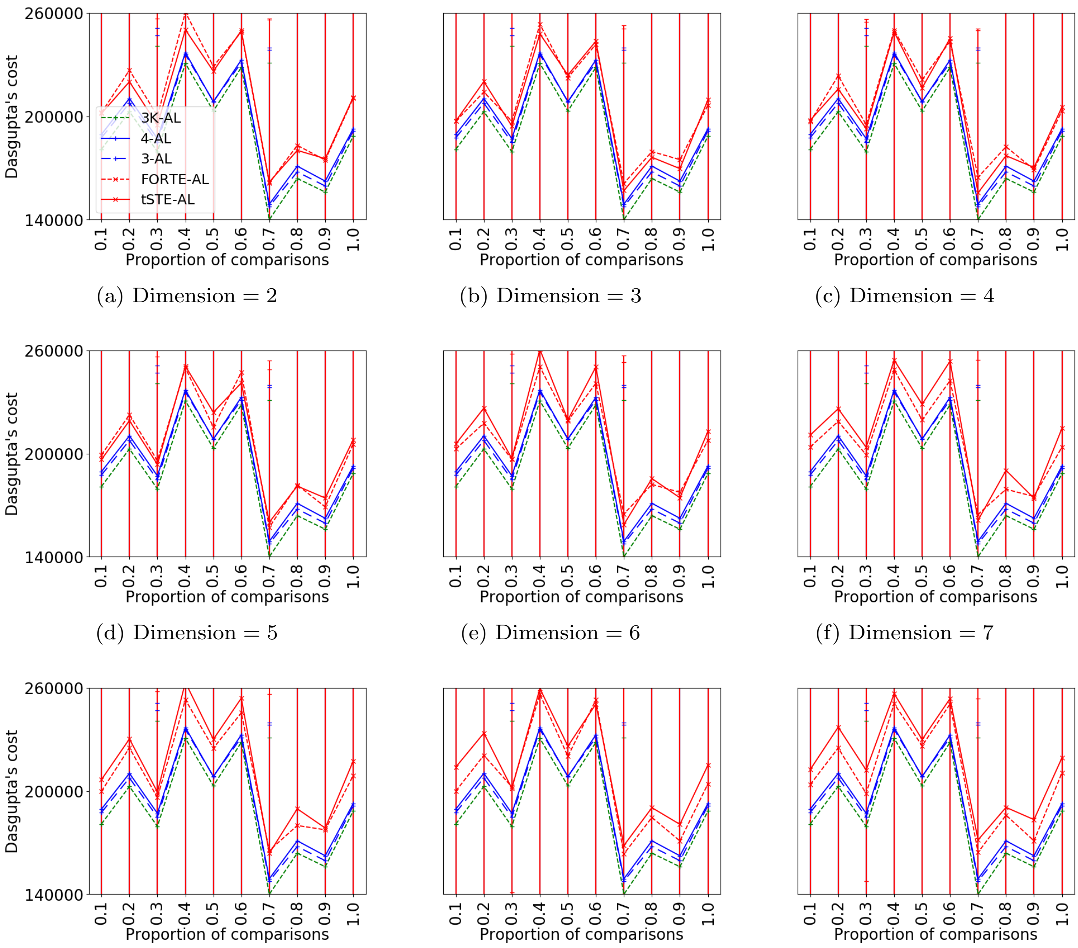


(a) Dimension = 2 (b) Dimension = 3 (c) Dimension = 4

(d) Dimension = 5 (e) Dimension = 6 (f) Dimension = 7

(g) Dimension = 8 (h) Dimension = 9 (i) Dimension = 10

Figure 6: Dasgupta's score of the different methods on the Glass dataset with increasing embedding dimensions for FORTE–AL and tSTE–AL. Best viewed in color.



(a) Dimension = 2 (b) Dimension = 3 (c) Dimension = 4

(d) Dimension = 5 (e) Dimension = 6 (f) Dimension = 7

(g) Dimension = 8 (h) Dimension = 9 (i) Dimension = 10

Figure 7: Dasgupta's score of the different methods on the 20news dataset with increasing embedding dimensions for FORTE–AL and tSTE–AL. Best viewed in color.