Foundations of Comparison-Based Hierarchical Clustering

Debarghya Ghoshdastidar[*,†]
Department of Informatics, TU Munich
ghoshdas@in.tum.de

Michaël Perrot[*]
Max Planck Institute for Intelligent Systems
michael.perrot@tuebingen.mpg.de

Ulrike von Luxburg
Department of Computer Science, University of Tübingen
Max Planck Institute for Intelligent Systems
luxburg@informatik.uni-tuebingen.de

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 is available, that is, 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 show that single and complete linkage are inherently comparison-based and we develop variants of average linkage. We provide statistical guarantees for the different methods under a planted hierarchical partition model. We also empirically demonstrate the performance of the proposed approaches 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, quadruplet comparisons of the form "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 has been widely discussed in the 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; 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 a 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, 2011; Tamuz et al., 2011; Van Der Maaten and Weinberger, 2012; Terada and von Luxburg, 2014; Amid and Ukkonen, 2015). As a consequence, the statistical

[*]Both authors contributed equally to the paper.
[†]This work was done when the author was affiliated to the University of Tübingen.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

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 (Jamieson and Nowak, 2011; Kleindessner and Luxburg, 2014; Jain et al., 2016; Arias-Castro et al., 2017), it is not known if one can design provably consistent learning algorithms using 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 for estimation of data dimension, centroid or density (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 by 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).

Motivation. 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 would be to combine the analysis of a classic 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 either adaptively or non-adaptively chosen sets of comparisons.

Contribution 1: 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). We observe that it is easy to perform agglomerative hierarchical clustering using only comparisons since one can directly reformulate single linkage and complete linkage clustering algorithms in the quadruplet comparisons 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 has stronger theoretical guarantees (Moseley and Wang, 2017; Cohen-Addad et al., 2018), it cannot be used in the comparison-based setup since it relies on an averaging of similarity scores. We propose two variants of average linkage clustering that can be applied to the quadruplet comparisons framework. We numerically compare the merits of these new methods with single and complete linkage and embedding based approaches.

Contribution 2: Guarantees for true hierarchy recovery. 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 independently 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. It cannot account for noise in the underlying objects and their associated similarities. Instead, we consider a theoretical model that generates random pairwise similarities with a planted hierarchical structure (Balakrishnan et al., 2011). This induces considerable dependence among

input : Set of objects $\mathcal{X} = \{x_1, \dots, x_N\}$; Cluster-level similarity $W : 2^{\mathcal{X}} \times 2^{\mathcal{X}} \to \mathbb{R}$.
output : Binary tree, or dendrogram, representing a hierarchical clustering of $\mathcal{X}$.
begin
  Let $\mathcal{B}$ be a collection of N singleton trees $\mathcal{C}_1, \dots, \mathcal{C}_N$ 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 $\mathcal{C}''.root = \{\mathcal{C}.root \cup \mathcal{C}'.root\}$, $\mathcal{C}''.left = \mathcal{C}$, and $\mathcal{C}''.right = \mathcal{C}'$.
    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.

the quadruplets, and makes the analysis challenging. We derive conditions under which different comparison-based agglomerative algorithms can exactly recover the hierarchy with high probability.

2 Background

In this section we introduce standard hierarchical clustering with known similarities, we describe the model used for the theoretical analyses, and we formalize the comparison-based framework.

2.1 Agglomerative hierarchical clustering with known similarity scores

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 $w : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The goal of hierarchical clustering is to group the N objects to form a binary tree such that x_i and x_j are merged in the bottom of the tree if their similarity score $w_{ij} = w(x_i, x_j)$ is high, and vice-versa. Here, we briefly review popular agglomerative clustering algorithms (Cohen-Addad et al., 2018). They rely on the similarity score w between 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 most similar clusters. This is described in Algorithm 1, where different choices of W lead to different algorithms. Given two clusters G and G', popular choices for $W(G, G')$ are

$$W(G, G') = \underbrace{\max_{x_i \in G, x_j \in G'} w_{ij}}_{\textbf{Single Linkage (SL)}}, \quad \text{or} \quad \underbrace{\min_{x_i \in G, x_j \in G'} w_{ij}}_{\textbf{Complete Linkage (CL)}}, \quad \text{or} \quad \underbrace{\sum_{x_i \in G, x_j \in G'} \frac{w_{ij}}{|G||G'|}}_{\textbf{Average Linkage (AL)}}.$$

2.2 Planted hierarchical model

Theoretically, we study the problem of hierarchical clustering under a *noisy hierarchical block matrix* (Balakrishnan et al., 2011) where, given N objects, the matrix of pairwise similarities can be written as $M + R$, where $M = (\mu_{ij})_{1 \le i,j \le N}$ is a symmetric ideal similarity matrix characterizing the planted hierarchy among the examples and $R = (r_{ij})_{1 \le i,j \le N}$ is a symmetric perturbation matrix that accounts for the noise in the observed similarity scores. In this paper, we assume that the entries $\{r_{ij}\}_{1 \le i < j \le N}$ are mutually independent and normally distributed, that is $r_{ij} \sim \mathcal{N}\left(0, \sigma^2\right)$, for some fixed variance σ^2. The ideal similarity matrix M is constructed in the following way. We assume that the planted hierarchy is a balanced binary tree of height L (see Figure 1), where the 2^L leaf nodes $\mathcal{G}_1, \dots, \mathcal{G}_{2^L}$ correspond to "pure clusters", each of size N_0. Thus, the total number of objects in $\mathcal{X}$ is $N = N_0 2^L$. For some constants $\delta > 0$ and μ, the ideal similarities are 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 ideal similarity is set to $\mu_{ij} = \mu - L\delta$ (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-$2, \dots, L-1$**:** The above process is repeated $L-1$ times, and in step ℓ, the ideal similarity across two newly-formed sub-groups is $\mu_{ij} = \mu - (L-\ell)\delta$.

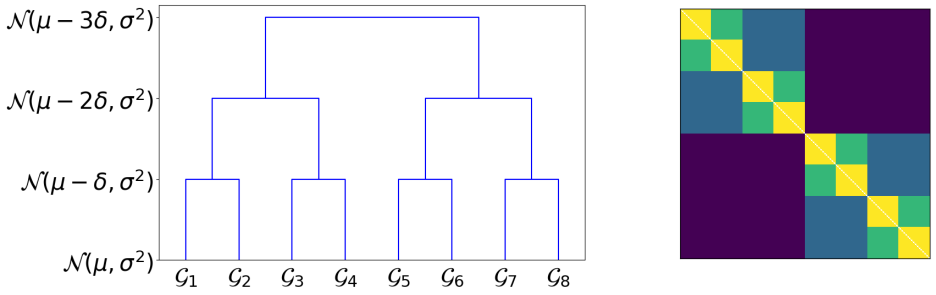


Figure 1: **(Left)** Illustration of the planted hierarchical model for $L = 3$ along with specification of the distributions for similarities at different levels; **(Right)** Hierarchical block structure in the expected pairwise similarity matrix, where darker implies smaller similarity.

Step-L: The above steps form 2^L clusters, $\mathcal{G}_1, \ldots, \mathcal{G}_{2^L}$, each of size N_0. The ideal similarity between two objects x_i, x_j belonging to the same cluster is $\mu_{ij} = \mu$ (yellow blocks in Figure 1).

This gives rise to similarities of the form $w_{ij} = \mu_{ij} + r_{ij}$ for all $i < j$. By symmetry of M and R, $w_{ji} = w_{ij}$. We can equivalently assume that, for all $i < j$, the similarities are independently drawn as $w_{ij} = w_{ji} \sim \mathcal{N}\left(\mu_{ij}, \sigma^2\right)$. Note that the pairwise similarity gets smaller in expectation when two objects are merged higher in the true hierarchy. We consider the problem of exact recovery of the above planted structure, that is correct identification of all the pure clusters $\mathcal{G}_1, \ldots, \mathcal{G}_{2^L}$ and recovery of the entire hierarchy among the clusters.

2.3 The comparison-based framework

In Section 2.1 we assumed that, even without a representation of the objects, we had access to a similarity function w. In the rest of this paper, we consider the ordinal setting, where w is not available, and information about similarities can only be accessed through quadruplet comparisons. We assume that we are given a set $\mathcal{Q} \subseteq \{(i, j, k, l) : x_i, x_j, x_k, x_l \in \mathcal{X}, w_{ij} > w_{kl}\}$, that is, for every ordered tuple $(i, j, k, l) \in \mathcal{Q}$, 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 small subset of all possible quadruplets. Since noise is inherent in the similarities, we do not consider it in the comparisons. We assume $\mathcal{Q}$ is obtained in either of the two following ways:
Active comparisons: In this case, the algorithm can adaptively ask an oracle quadruplet queries of the form $w_{ij} \gtrless w_{kl}$ and the outcome will be either $w_{ij} > w_{kl}$ or $w_{ij} < w_{kl}$.
Passive comparisons: In this case, for every tuple (i, j, k, l), we assume that with some sampling probability $p \in (0, 1]$, there is a comparison $w_{ij} \gtrless w_{kl}$ and based on the outcome either $(i, j, k, l) \in \mathcal{Q}$ or $(k, l, i, j) \in \mathcal{Q}$. We also assume that the observations of the quadruplets are independent.

3 Comparison-based hierarchical clustering

In this section, we discuss that single linkage and complete linkage can be easily implemented in the comparison-based setting, provided that we have access to $\Omega\left(N^2\right)$ adaptively selected quadruplets. However, their statistical guarantees are very weak. It prompts us to study two variants of average linkage. On the one hand, Quadruplets-based Average Linkage (4–AL) uses average linkage-like ideas to directly estimate the cluster level similarities from the quadruplet comparisons. On the other hand, Quadruplets Kernel Average Linkage (4K–AL) uses the quadruplet comparisons to estimate the similarities between the different objects and then uses standard average linkage. We show that both of these variants have good statistical performances in the following senses: (i) they can exactly recover the planted hierarchy under mild assumptions on the signal-to-noise ratio $\frac{\delta}{\sigma}$ and the size of the pure clusters $N_0 = \frac{N}{2^L}$ in the model introduced in Section 2.2, (ii) 4K–AL only needs $\mathcal{O}\left(N \ln N\right)$ active comparisons to achieve exact recovery, and (iii) both 4K–AL and 4–AL can achieve exact recovery using only a small subset of passively obtained quadruplets (sampling probability $p \ll 1$).

3.1 Single linkage (SL) and complete linkage (CL)

The single and complete linkage algorithms inherently fall in the comparison-based framework. To see this, first notice that the $\arg\max$ and $\arg\min$ functions used in these methods only depend on quadruplet comparisons. Although it is not possible to exactly compute the linkage value $W(G, G')$, one can retrieve, in each cluster, the pair of objects that achieve the maximum or minimum similarity. Then, 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'. This discussion emphasizes that CL and SL fall well in the comparison-based framework when the quadruplets can be adaptively chosen—in order to select pairs with minimum or maximum similarities. The next proposition, proved in the appendix, bounds the number of active comparisons necessary and sufficient to use SL and CL.

Proposition 1 (Active query complexity of SL and CL). *The SL and CL algorithms require at least $\Omega\left(N^2\right)$ and at most $\mathcal{O}\left(N^2 \ln N\right)$ number of active quadruplet comparisons.*

We now state a sufficient condition for exact recovery of the planted model for both SL and CL as well as a matching (up to constant) necessary condition for SL. The proof is in the appendix.

Theorem 1 (Exact recovery of planted hierarchy by SL and CL). *Assume that $\eta \in (0, 1)$. If $\frac{\delta}{\sigma} \geq 4\sqrt{\ln\left(\frac{N}{\eta}\right)}$, then SL and CL exactly recover the planted hierarchy with probability $1 - \eta$. Conversely, for $\frac{\delta}{\sigma} \leq \frac{1}{4}\sqrt{\ln\left(\frac{N}{2^L}\right)}$ and large $\frac{N}{2^L}$, SL fails to recover the hierarchy with probability $\frac{1}{2}$.*

Theorem 1 implies that a necessary and sufficient condition for exact recovery by single linkage is that the signal-to-noise ratio grows as $\sqrt{\ln N}$ with the number of examples. This strong requirement raises the question of whether one can achieve exact recovery under weaker assumptions and with less quadruplets. The subsequent sections provide an affirmative answer to this question.

3.2 Quadruplets kernel average linkage (4K–AL)

Average linkage is difficult to cast to the ordinal framework due to the averaging of pairwise similarities, w_{ij}, which cannot be computed using only comparisons. A first way to overcome this issue is to use the quadruplet comparisons to derive some kind of proxies for the similarities w_{ij}. These proxy similarities can then be directly used in the standard formulation of average linkage. To derive them we use ideas that are close in spirit to the triplet comparisons-based kernel developed by Kleindessner and von Luxburg (2017a). Furthermore, we propose two different definitions depending on whether we use active comparisons (Equation 1) or passive comparisons (Equation 3).

Active case. We first consider the active case, where the quadruplet comparisons to be evaluated can be chosen by the algorithm. A pair of distinct items (i_0, j_0) is chosen uniformly at random, and a set of landmark points $\mathcal{S}$ is constructed such that every $k \in \{1, \ldots, N\}$ is independently added to $\mathcal{S}$ with probability q. The proxy similarity between two distinct objects x_i and x_j is then defined as

$$K_{ij} = \sum_{k \in \mathcal{S} \setminus \{i,j\}} \left(\mathbb{I}_{\left(w_{ik} > w_{i_0 j_0}\right)} - \mathbb{I}_{\left(w_{ik} < w_{i_0 j_0}\right)}\right)\left(\mathbb{I}_{\left(w_{jk} > w_{i_0 j_0}\right)} - \mathbb{I}_{\left(w_{jk} < w_{i_0 j_0}\right)}\right). \tag{1}$$

The underlying idea is that two similar objects should behave similarly with respect to any third object, that is if x_i and x_j are similar then we should have $w_{ik} \approx w_{jk}$ for any other object x_k. Since we cannot directly access the similarities, we instead use comparisons to a reference similarity $w_{i_0 j_0}$ to evaluate the closeness between w_{ik} and w_{jk}.

The next theorem presents exact recovery guarantees for 4K–AL with actively obtained comparisons.

Theorem 2 (Exact recovery of planted hierarchy by 4K–AL with active comparisons). *Let $\eta \in (0, 1)$ and $\Delta = \frac{\eta^2}{100}\frac{\delta}{\sigma}e^{-2L^2\delta^2/\sigma^2}$. There exists an absolute constant $C > 0$ such that if $N_0 > \frac{4}{\Delta}\sqrt{N}$ and we set $q > \max\left\{C\frac{2^{2L}}{N\Delta^4}\ln\left(\frac{N}{\eta}\right), \frac{3}{N}\ln\left(\frac{2}{\eta}\right)\right\}$, then with probability at least $1 - \eta$, 4K–AL exactly recovers the planted hierarchy using at most $2qN^2$ number of actively chosen quadruplet comparisons.*

In particular, if $L = \mathcal{O}(1)$, the above statement implies that even with $\frac{\delta}{\sigma}$ constant, 4K–AL exactly recovers the planted hierarchy with probability $1 - \eta$ using only $\mathcal{O}(N \ln N)$ active comparisons.

The above result shows that, in comparison to SL or CL, the proposed 4K–AL method achieves consistency for smaller signal-to-noise ratio $\frac{\delta}{\sigma}$, and can also do so with only $\mathcal{O}(N \ln N)$ active comparisons, which is much smaller than that needed by SL and CL. Our result also aligns with the conclusion of Emamjomeh-Zadeh and Kempe (2018), who showed that $\mathcal{O}(N \ln N)$ active triplet comparisons suffice to recover hierarchy under a different (data-independent) noise model. It is also worth noting that the condition $N_0 = \Omega\left(\sqrt{N}\right)$ is necessary for exact recovery of the planted hierarchy since the condition is necessary even in the case of planted flat clustering (Chen and Xu, 2016, Figure 1).

From a theoretical perspective, it is sufficient to use a single random reference similarity $w_{i_0 j_0}$. However, in practice, we observe better performances when considering a set $\mathcal{R}$ of multiple reference pairs. Hence, in the experiments, we use the following extension of the above kernel function:

$$K_{ij} = \sum_{(i_0,j_0)\in\mathcal{R}} \sum_{k\in\mathcal{S}\setminus\{i,j\}} \left(\mathbb{I}_{\left(w_{ik} > w_{i_0 j_0}\right)} - \mathbb{I}_{\left(w_{ik} < w_{i_0 j_0}\right)}\right)\left(\mathbb{I}_{\left(w_{jk} > w_{i_0 j_0}\right)} - \mathbb{I}_{\left(w_{jk} < w_{i_0 j_0}\right)}\right). \quad (2)$$

Passive case. Theorem 2 shows that 4K–AL can exactly recover the planted hierarchy even for a constant signal-to-noise ratio, provided that it can actively choose the quadruplets. It is natural to ask if the same holds in the passive case, where we do not have the freedom of querying specific comparisons but instead have access to a small pre-computed set of quadruplet comparisons $\mathcal{Q}$. We address this problem using the following variant of the aforementioned quadruplets kernel:

$$K_{ij} = \sum_{k,l=1,k<l}^{N} \sum_{r=1}^{N} \left(\mathbb{I}_{(i,r,k,l)\in\mathcal{Q}} - \mathbb{I}_{(k,l,i,r)\in\mathcal{Q}}\right)\left(\mathbb{I}_{(j,r,k,l)\in\mathcal{Q}} - \mathbb{I}_{(k,l,j,r)\in\mathcal{Q}}\right) \quad (3)$$

for all $i \neq j$. This formulation extends the active kernel in (1) by using all $\binom{N}{2}$ pairs of (k,l) as reference similarities instead of a single pair (i_0, j_0). But each term in the sum contributes only when we simultaneously observe the comparisons between (i,r) and (k,l) and between (j,r) and (k,l). Theorem 3 presents guarantees for 4K–AL with quadruplets obtained from the passive comparisons model in Section 2.3.

Theorem 3 (Exact recovery of planted hierarchy by 4K–AL with passive comparisons). *Let $\eta \in (0,1)$ and $\Delta = \frac{\delta}{2\sigma} e^{-L^2\delta^2/4\sigma^2}$. There exists an absolute constant $C > 0$ such that if $N_0 > \frac{8}{\Delta}\sqrt{N}$ and we set $p > \max\left\{C\frac{2^L}{\Delta^2}\sqrt{\frac{1}{N}\ln\left(\frac{N}{\eta}\right)}, \frac{2}{N^4}\ln\left(\frac{2}{\eta}\right)\right\}$, then with probability at least $1-\eta$, the 4K–AL algorithm exactly recovers the planted hierarchy using at most pN^4 quadruplet comparisons, which are passively obtained based on the model described in Section 2.3.*

In particular, if $L = \mathcal{O}(1)$, the above statement implies that even with $\frac{\delta}{\sigma}$ constant, 4K–AL exactly recovers the planted hierarchy with probability $1-\eta$ using $\mathcal{O}\left(N^{7/2}\ln N\right)$ passive comparisons.

The derived conditions for exact recovery are similar to Theorem 2 in terms of $\frac{\delta}{\sigma}$, but passive 4K–AL requires a much larger number of passive comparisons than active 4K–AL. While this may seem disappointing, $\mathcal{O}\left(N^{7/2}\right)$ passive comparisons might, in fact, be necessary in this case. Indeed, Emamjomeh-Zadeh and Kempe (2018, Theorem 2.3) show that in the case of triplets, $\Omega\left(N^3\right)$ passive triplet comparisons are necessary to exactly recover a hierarchy in the worst case. The proof can be easily adapted to the quadruplet comparison setting to prove a worst-case complexity of $\Omega\left(N^4\right)$ passive quadruplets. The above result shows that under the planted model, which is simpler than the worst-case, the query complexity can be improved at least by a factor of $\sqrt{N}$. Further study is required to identify a precise necessary condition. We also believe that the passive query complexity should reduce considerably if the signal-to-noise ratio $\frac{\delta}{\sigma}$ grows with N.

3.3 Quadruplets-based average linkage (4–AL)

In 4K–AL we derived a proxy for the similarities between objects and then used standard average linkage. In this section we consider a different approach where we use the quadruplet comparisons to define a new cluster-level similarity function. This method is particularly well suited when it is not possible to actively query the comparisons. We assume that we are given a set of passively

obtained quadruplets $\mathcal{Q}$ as in the previous section (4K–AL with passive comparisons). Using this set of comparisons, 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

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

The 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. This formulation suggests that an agglomerative clustering should merge G_1, G_2 before G_3, G_4 if $\mathbb{W}_{\mathcal{Q}}\left(G_1, G_2 \| G_3, G_4\right) > 0$. Also, note that $\mathbb{W}_{\mathcal{Q}}\left(G_1, G_2 \| G_3, G_4\right) = -\mathbb{W}_{\mathcal{Q}}\left(G_3, G_4 \| G_1, G_2\right)$ and $\mathbb{W}_{\mathcal{Q}}\left(G_1, G_2 \| G_1, G_2\right) = 0$, which hints that (4) is a preference relation between pairs of clusters. We use the above preference relation $\mathbb{W}_{\mathcal{Q}}$ to define a new cluster-level similarity function W that can be used in Algorithm 1. Hence, given two clusters G_p, G_q, $p \neq q$, we define their similarity 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 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 measure W provides an average linkage approach based on quadruplets (4–AL), whose statistical guarantees are presented below.

Theorem 4 (Exact recovery of planted hierarchy by 4–AL with passive comparisons). *Let $\eta \in (0,1)$ and $\Delta = \frac{\delta}{2\sigma} e^{-L^2\delta^2/4\sigma^2}$. Assume the following:*
(i) An initial step partitions $\mathcal{X}$ into pure clusters of sizes in the range $[m, 2m]$ for some $m \leq \frac{1}{2}N_0$.
(ii) $\mathcal{Q}$ is a passively obtained set of quadruplet comparisons, where each tuple (i,j,k,l) is observed independently with probability $p > \frac{C}{m\Delta^2} \max\left\{\ln N, \frac{1}{m}\ln\left(\frac{1}{\eta}\right)\right\}$ for some constant $C > 0$.
Then, with probability $1-\eta$, starting from the given initial partition and using $|\mathcal{Q}| \leq pN^4$ number of passive comparisons, 4–AL exactly recovers the planted hierarchy.

In particular, if $L = \mathcal{O}(1)$, the above statement implies that, when $\frac{\delta}{\sigma}$ is a constant, 4–AL exactly recovers the planted hierarchy with probability $1-\eta$ using $\mathcal{O}\left(\frac{N^4 \ln N}{m}\right)$ passive comparisons.

Compared to 4K–AL (Theorem 3), the guarantee for 4-AL in Theorem 4 additionally requires an initial partitioning of $\mathcal{X}$ into small pure clusters of size m. This is reasonable in the context of the hierarchical clustering literature since existing consistency results for average linkage also require similar assumptions (Cohen-Addad et al., 2018, Theorem 5.8). In principle, one may use passive 4K–AL to obtain these initial clusters. Theorem 4 shows that if the size of initial clusters is much larger than $\ln N$, then we do not need to observe all the quadruplets. Moreover, if $L = \mathcal{O}(1)$ and we have $\Omega(N_0)$-sized initial clusters, then the subsequent steps of 4–AL require only $\mathcal{O}\left(N^3 \ln N\right)$ passive comparisons out of the $\mathcal{O}\left(N^4\right)$ total number of available comparisons. This is less quadruplets than 4K–AL, but it is still large for practical purposes. It remains an open question whether better sampling rates can be achieved in the passive case. From a practical perspective, our experiments in Section 4 demonstrate that 4–AL performs better than 4K–AL even when no initial clusters are provided, that is $m = 1$.

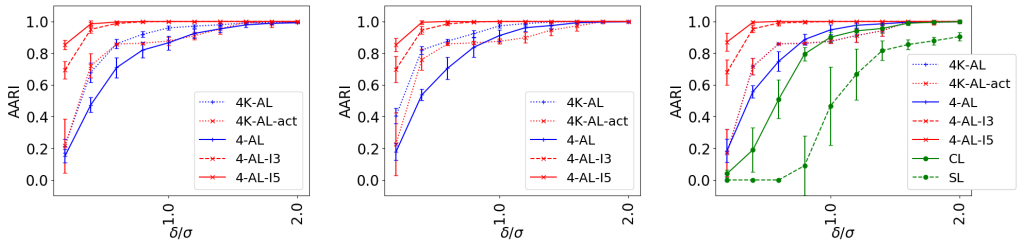
4 Experiments

In this section we evaluate our approaches on several problems: we empirically verify our theoretical findings, we compare our methods[1] to ordinal embedding based approaches on standard datasets, and we illustrate their behaviour on a comparison-based dataset.

4.1 Planted hierarchical model

We first use the planted hierarchical model presented in Section 2.2 to generate data and study the performance of the various methods introduced in Section 3.

[1] The code of our methods is available at `https://github.com/mperrot/ComparisonHC`.



(a) Proportion $p = 0.01$. (b) Proportion $p = 0.1$. (c) Proportion $p = 1$.

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, 2b, and, 2c, the methods get different proportions p of all the quadruplets. Best viewed in color.

Data. Recall that our generative model has several parameters, the within-cluster mean similarity 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 different guarantees presented in Section 3, it is clear that the hardness of the problem depends mainly on the signal-to-noise ratio $\frac{\delta}{\sigma}$, and the probability p of observing samples for the passive methods. Hence, to study the behaviour of the different methods with respect to these two quantities, we set $\mu = 0.8$, $\sigma = 0.1$, $N_0 = 30$, and $L = 3$ and we vary $\delta \in \{0.02, 0.04, \ldots, 0.2\}$ and $p \in \{0.01, 0.02, \ldots, 0.1, 1\}$.

Methods. We study SL, CL, which always use the same number of active comparisons and thus are not impacted by p. We also consider 4K–AL with passive comparisons and its active counterpart, 4K–AL–act, implemented as described in (2) with $q = \frac{\ln N}{N}$ and the number of references in $\mathcal{R}$ chosen so that the number of comparisons observed is the same as for the passive methods. Finally, we study 4–AL with no initial pure clusters and two variants 4–AL–I3 and 4–AL–I5 that have access to initial clusters of sizes 3 and 5 respectively. 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.

Evaluation function. As a measure of performance we use the Averaged Adjusted Rand Index (AARI) between the ground truth hierarchy and the hierarchies learned by the different methods. The main idea behind the AARI is to extend the Adjusted Rand Index (Hubert and Arabie, 1985) to hierarchies by averaging over the different levels (see the appendix for a formal definition). This measure takes values in $[0, 1]$ with higher values for more similar hierarchies—AARI $(\mathcal{C}, \mathcal{C}') = 1$ implies identical hierarchies. We report the mean and the standard deviation of 10 repetitions.

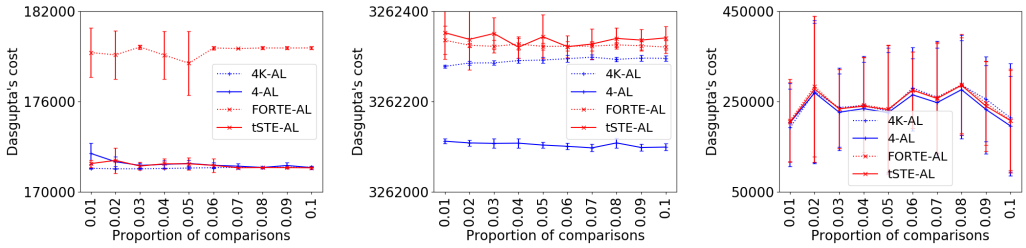
Results. In Figure 2 we present the results for $p = 0.01$, $p = 0.1$ and $p = 1$. We defer the other results to the appendix. Firstly, similar to the theory, SL can hardly recover the planted hierarchy, even for large values of $\frac{\delta}{\sigma}$. CL performs better than SL which implies that it might be possible to derive better guarantees. We observe that 4K–AL, 4K–AL–act, and, 4–AL are able to exactly recover the true hierarchy for smaller signal-to-noise ratio and their performances do not degrade much when the number of sampled comparisons is reduced. Finally, as expected, the best method is 4–AL–I5. It uses large initial clusters but recovers the true hierarchy even for very small values of $\frac{\delta}{\sigma}$.

4.2 Standard clustering datasets

In this second set of experiments we compare our passive methods, 4K–AL with passive comparisons and 4–AL without initial clusters, to two baselines that use ordinal embedding as a first step.

Baselines. We consider 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 consider several embedding dimensions and report the best results in the main paper. In the appendix, we detail the cosine similarity and report results for other embedding dimensions.

Data. We evaluate the different approaches on 3 different datasets commonly used in hierarchical clustering: Zoo, Glass and 20news (Heller and Ghahramani, 2005; Vikram and Dasgupta, 2016). To



(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 set to 2. Best viewed in color.

fit the comparison-based setting we generate the comparisons using the cosine similarity. Since it is not realistic to assume that all the comparisons are available. We use the procedure described in Section 2.3 to passively obtain a proportion $p \in \{0.01, 0.02, \ldots, 0.1\}$ of all the quadruplets. Some statistics on the datasets and details on the comparisons generation are presented in the appendix.

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. Instead, we use the recently proposed Dasgupta's cost (Dasgupta, 2016) that has been specifically designed to evaluate hierarchical clustering methods. The idea of this cost is that similar objects that are merged higher in the hierarchy should be penalized. Hence, a lower cost indicates a better hierarchy. Details are provided in the appendix. For all the experiments we report the mean and the standard deviation of 10 repetitions.

Results. We report the results in Figure 3. We note that the proportion of comparisons does not have a large impact as the results are, on average, stable across all regimes. Our methods are either comparable or better than the embedding-based ones. They do not need to first embed the examples and thus do not impose a strong Euclidean structure on the data. The impact of this structure is more or less pronounced depending on the dataset. Furthermore, as illustrated in the appendix, a poor choice of embedding dimension can drastically worsen the results of the embedding methods.

Comparison-based dataset. In the appendix, we also apply the different methods to a comparison-based dataset called Car (Kleindessner and von Luxburg, 2017b).

5 Conclusion

We investigated the problem of hierarchical clustering in a comparison-based setting. We showed that the single and complete linkage algorithms (SL and CL) could be used in the setting where comparisons are actively queried, but with poor exact recovery guarantees under a planted hierarchical model. We also proposed two new approaches based on average linkage (4K–AL and 4–AL) that can be used in the setting of passively obtained comparisons with good guarantees in terms of exact recovery of the planted hierarchy. An active version of 4K–AL achieves exact recovery using only $\mathcal{O}(N \ln N)$ active comparisons. Empirically, we confirmed our theoretical findings and compared our methods to two ordinal embedding based baselines on standard and comparison-based datasets.

The paper leaves several open problems. From an algorithmic perspective, the key question is whether one can develop similar provable methods in the triplet setting, where one has access to comparisons of the form "x_i is more similar to x_j than to x_k". An equivalent to passive 4K–AL can obtained using the triplet kernel of Kleindessner and von Luxburg (2017a), while triplet-based variants of active 4K–AL and 4–AL require careful designing. We leave the description of such algorithms and their theoretical analysis under planted hierarchy to a follow-up work. From a theoretical perspective, the main question is to derive necessary conditions and query complexities for exact recovery of planted hierarchy, and subsequently, validate whether the proposed algorithms are indeed optimal. Additionally, it would be interesting to analyse the performance of the proposed methods in terms of Dasgupta's score, and in presence of noisy queries, that is when some answers are randomly flipped.

Acknowledgments

This work has been supported by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, DFG, ZUK 63), by the DFG Cluster of Excellence "Machine Learning – New Perspectives for Science", EXC 2064/1, project number 390727645, by the BMBF through the Tuebingen AI Center (FKZ: 01IS18039A), and by the Baden-Württemberg Eliteprogramm for Postdocs.

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.

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

Balakrishnan, S., Xu, M., Krishnamurthy, A., and Singh, A. (2011). Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems*, pages 954–962.

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.

Chen, Y. and Xu, J. (2016). Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *Journal of Machine Learning Research*, 17(27):1–57.

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.

Emamjomeh-Zadeh, E. and Kempe, D. (2018). Adaptive hierarchical clustering using ordinal queries. In *Symposium on Discrete Algorithms*, pages 415–429.

Haghiri, S., Garreau, D., and von Luxburg, U. (2018). Comparison-based random forests. In *International Conference on Machine Learning*, pages 1866–1875.

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 *InternationalConference on Machine Learning*, pages 297–304.

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

Kazemi, E., Chen, L., Dasgupta, S., and Karbasi, A. (2018). Comparison based learning from weak oracles. In *International Conference on Artificial Intelligence and Statistics*, pages 1849–1858.

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*, pages 6807–6817.

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.

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.

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*.

Foundations of Comparison-Based Hierarchical Clustering – Appendix

Debarghya Ghoshdastidar*
Department of Informatics, TU Munich
ghoshdas@in.tum.de

Michaël Perrot*
Max Planck Institute for Intelligent Systems
michael.perrot@tuebingen.mpg.de

Ulrike von Luxburg
Department of Computer Science, University of Tübingen
Max Planck Institute for Intelligent Systems
luxburg@informatik.uni-tuebingen.de

Appendix A contains the proofs of our different theorems, and Appendix B has details on the experiments along with further numerical results.

A The hierarchical model and proofs of the theoretical results

In this section, we illustrate the planted hierarchical model and we provide detailed proofs of Theorems 1–4.

A.1 Notations

We first recall some of the key quantities associated with the planted model, which include:

- N, the number of objects;
- L, the number of levels in the hierarchy;
- $N_0 = \frac{N}{2^L}$, the size of the pure clusters;
- μ, expected similarity between pairs belonging to a pure cluster;
- δ, the separation between the expected similarities across consecutive levels; and
- σ, the standard deviation of the similarities.

Throughout the appendix, we use Z to denote a generic standard normal random variable, that is, $Z \sim \mathcal{N}(0,1)$. 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 $\{0, 1, \ldots, L\}$.

A.2 Analysis of Single Linkage (SL) and Complete Linkage (CL)

Proposition 1 (Active query complexity of SL and CL). *The SL and CL algorithms require at least $\Omega\left(N^2\right)$ and at most $\mathcal{O}\left(N^2 \ln N\right)$ number of active quadruplet comparisons.*

Proof. In the first step of SL or CL, the algorithm merges the pair x_i, x_j if $w_{ij} \geq w_{kl}$ for all $k, l \in \{1, \ldots, N\}$. This requires $\binom{N}{2}$ number of ordinal comparisons to find the minimum, and hence, the active query complexity of SL and CL is at least $\Omega\left(N^2\right)$.

*Both authors contributed equally to the paper.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

To prove an upper bound on the active query complexity, it suffices to observe that single/complete linkage only requires a total ordering of the $\binom{N}{2}$ scalar similarities $\{w_{ij} : i < j\}$. Using a sorting algorithm such as merge-sort, this ordering can be easily obtained from $\mathcal{O}\left(N^2 \ln N\right)$ actively chosen comparisons. □

Theorem 1 (Exact recovery of planted hierarchy by SL and CL). *Assume that $\eta \in (0,1)$. If $\frac{\delta}{\sigma} \geq 4\sqrt{\ln\left(\frac{N}{\eta}\right)}$, then SL and CL exactly recover the planted hierarchy with probability $1-\eta$.*

Conversely, for $\frac{\delta}{\sigma} \leq \frac{1}{4}\sqrt{\ln\left(\frac{N}{2^L}\right)}$ and large $\frac{N}{2^L}$, SL fails to recover the hierarchy with probability $\frac{1}{2}$.

Proof. **We first prove the sufficient condition for exact recovery.** 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 w, we know that w_{ij} lies in the range $\left(\mu-(L-\ell_{ij}^{lca})\delta-\frac{\delta}{2}, \mu-(L-\ell_{ij}^{lca})\delta+\frac{\delta}{2}\right)$ 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 the 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 and hence, the claim.

We now prove the converse statement for SL. We first prove the result for $L = 1$. The argument easily extends to $L > 1$ from the observation that exact recovery of the entire hierarchy involves exact recovery for pairs of clusters at $L-1$ levels. For $L = 1$, there are two pure clusters, $\mathcal{G}_1$ and $\mathcal{G}_2$, that are split at the top level of the true hierarchy.

Recall that single linkage corresponds to a cluster tree on the set of items (Chaudhuri et al., 2014). For any $t \in \mathbb{R}$, we consider the subgraph G_t of the cluster tree with edge set $E_t = \{(i,j) : w_{ij} > t\}$. Observe that G_t is equivalent to a stochastic block model, where

$$\mathbf{P}\big((i,j) \in E_t\big) = \begin{cases} 1-\Phi\left(\dfrac{t-\mu}{\sigma}\right) & \text{for } i,j \text{ in the same cluster, and} \\ 1-\Phi\left(\dfrac{t-\mu+\delta}{\sigma}\right) & \text{when } i,j \text{ belong to different clusters.} \end{cases} \tag{1}$$

Let p, q denote the aforementioned within and inter-cluster edge probabilities in (1), and recall the bounds on the Gaussian tail

$$\frac{1}{\sqrt{2\pi}}\frac{1}{2x}e^{-x^2/2} < 1-\Phi(x) < \frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2}\,, \tag{2}$$

which is valid for all $x \geq 1$. Setting $t = \mu + \sigma\sqrt{2\ln N_0}$, it is easy to verify that

$$p < \frac{1}{\sqrt{2\pi}}\frac{1}{N_0\sqrt{2\ln N_0}} \qquad \text{and} \qquad q > \frac{1}{\sqrt{2\pi}}\frac{1}{2\left(\sqrt{2\ln N_0}+\frac{\delta}{\sigma}\right)}e^{-\left(\sqrt{2\ln N_0}+\frac{\delta}{\sigma}\right)^2/2}.$$

Assuming $\frac{\delta}{\sigma} < \frac{1}{4}\sqrt{\ln N_0}$, the lower bound on q can be simplified as $q > \frac{1}{10N_0\sqrt{N_0\ln N_0}}$. Hence, for large enough N_0, we have $p < \frac{1}{N_0}$ and $q \gg \frac{\ln N_0}{N_0^2}$. Now observe that the two subgraphs of

G_t restricted to $\mathcal{G}_1$ and $\mathcal{G}_2$, $G_{t|_{\mathcal{G}_1}}$ and $G_{t|_{\mathcal{G}_2}}$, are Erdős-Rényi graphs, each with N_0 vertices and edge-probability p. Using a standard result for random graphs (Chapter 8 of Blum et al., 2018), we can conclude that both $G_{t|_{\mathcal{G}_1}}$ and $G_{t|_{\mathcal{G}_2}}$ are disconnected with high probability for $p < \frac{1}{N_0}$. Similarly, since $q \gg \frac{\ln N_0}{N_0^2}$, one can conclude that, with high probability, there exist edges between $G_{t|_{\mathcal{G}_1}}$ and $G_{t|_{\mathcal{G}_2}}$. Based on the cluster tree perspective of single linkage (Chaudhuri et al., 2014), the above conclusions about connectivity of $G_{t|_{\mathcal{G}_1}}$ and $G_{t|_{\mathcal{G}_2}}$ implies that SL merges items from $\mathcal{G}_1$ and $\mathcal{G}_2$ before extracting the pure clusters. For large enough N_0, the probability of this event is greater than $\frac{1}{2}$. □

A.3 Analysis of Active Quadruplets Kernel based Average Linkage (4K–AL)

Recall that the active quadruplet kernel is defined in the following way. A pair of distinct items (i_0, j_0) is chosen uniformly, and a set of landmark points $\mathcal{S}$ is constructed such that every $k \in \{1, \ldots, N\}$ is independently added to $\mathcal{S}$ with probability q. The kernel K is defined as

$$K_{ij} = \sum_{k \in \mathcal{S} \setminus \{i,j\}} \left(\mathbb{I}_{\left(w_{ik} > w_{i_0 j_0}\right)} - \mathbb{I}_{\left(w_{ik} < w_{i_0 j_0}\right)} \right) \left(\mathbb{I}_{\left(w_{jk} > w_{i_0 j_0}\right)} - \mathbb{I}_{\left(w_{jk} < w_{i_0 j_0}\right)} \right) \tag{3}$$

for $i \neq j$. For ease of notation, we introduce the terms $w^* = w_{i_0 j_0}$ and $\xi_k = \mathbb{I}_{(k \in \mathcal{S})}$. It follows that $\xi_1, \ldots, \xi_N \sim_{iid} \text{Bernoulli}(q)$ and, with these notations, we write the kernel function (3) as

$$K_{ij} = \sum_{k \neq i,j} \xi_k \left(2\mathbb{I}_{(w_{ik} > w^*)} - 1 \right) \left(2\mathbb{I}_{(w_{jk} > w^*)} - 1 \right),$$

where the re-arrangement of indicators are under the planted model assumption since any two similarity scores are distinct with probability 1 due to the Gaussian assumption.

We now restate and prove the exact recovery guarantee for 4K–AL with actively obtained comparisons.

Theorem 2 (Exact recovery of planted hierarchy by 4K–AL with active comparisons). *Let $\eta \in (0, 1)$ and $\Delta = \frac{\eta^2}{100} \frac{\delta}{\sigma} e^{-2L^2\delta^2/\sigma^2}$. There exists an absolute constant $C > 0$ such that if $N_0 > \frac{4}{\Delta}\sqrt{N}$ and we set*

$$q > \max\left\{ C \frac{2^{2L}}{N\Delta^4} \ln\left(\frac{N}{\eta}\right), \frac{3}{N} \ln\left(\frac{2}{\eta}\right) \right\},$$

then with probability at least $1 - \eta$, 4K–AL exactly recovers the planted hierarchy using at most $2qN^2$ number of actively chosen quadruplet comparisons.

In particular, if $L = \mathcal{O}(1)$, the above statement implies that even with $\frac{\delta}{\sigma}$ constant, 4K–AL exactly recovers the planted hierarchy with probability $1 - \eta$ using only $\mathcal{O}(N \ln N)$ active comparisons.

Proof. We prove the result by proving the following statements:

- the probability that 4K–AL queries more than $2qN^2$ comparisons is at most $\frac{\eta}{2}$, and
- the probability of not achieving exact recovery is at most $\frac{\eta}{2}$.

To derive the bound on the number of comparisons, we observe that evaluation of the entire kernel matrix requires quadruplet comparisons of the form $\mathbb{I}_{(w_{ik} > w^*)}$ for all $i = 1, \ldots, N$ and $k \in \mathcal{S}$. Hence, the total number of comparisons is $N|\mathcal{S}|$, which can be bounded by showing that the size of $\mathcal{S}$ is at most $2qN$. This follows from Bernstein's inequality since

$$\begin{aligned} \mathbf{P}(|\mathcal{S}| > 2qN) &= \mathbf{P}\left(\sum_{k=1}^{N} \xi_k - qN > qN \right) \\ &\leq \exp\left(-\frac{q^2 N^2}{2Nq(1-q) + \frac{2}{3}qN} \right) \leq \exp\left(-\frac{qN}{3} \right), \end{aligned}$$

which is bounded by $\frac{\eta}{2}$ since $q > \frac{3}{N} \ln\left(\frac{2}{\eta}\right)$.

To derive the exact recovery guarantee, we analyze the kernel matrix K, and also 4K–AL, conditioned on w^*. For this, we need to characterize the behaviour of w^* under the planted model. Since w^* is the similarity of a randomly chosen pair, one can observe that $w^* \sim \sum_{\ell=0}^{L} a_\ell \mathcal{N}\left(\mu - \ell\delta, \sigma^2\right)$ has a mixture of Gaussian distribution, where the weights $a_0 = \frac{2^L\binom{N_0}{2}}{\binom{N}{2}}$ and $a_\ell = \frac{2^{L+\ell-2}N_0^2}{\binom{N}{2}}$ for $\ell = 1, \ldots, L$ are the proportion of similarities corresponding to item pairs merged at level $(L-\ell)$ of the panted hierarchy. We claim that, with probability $1 - \frac{\eta}{4}$,

$$\mu - L\delta - \sigma\sqrt{2\ln\left(\frac{8}{\eta}\right)} < w^* < \mu + \sigma\sqrt{2\ln\left(\frac{8}{\eta}\right)}. \tag{4}$$

The bounds follow from the mixture of Gaussian nature of w^* since

$$\begin{aligned}\mathbf{P}\left(w^* > t\right) &= \sum_{\ell=0}^{L} a_\ell \mathbf{P}\left(\mu - \ell\delta + \sigma Z > t\right) \\ &\leq \mathbf{P}(\mu + \sigma Z > t)\,,\end{aligned}$$

where we use Z to denote a standard normal random variable. Setting $t = \mu + \sigma\sqrt{2\ln\left(\frac{8}{\eta}\right)}$ and using the upper bound on Gaussian tail probability (2), we can bound the above probability by $\frac{\eta}{8}$. A similar argument holds for the lower bound on w^*, where the probability of violating the bound is also at most $\frac{\eta}{8}$. Hence, the bounds in (4) hold with probability $1 - \frac{\eta}{4}$.

We next compute the expected kernel matrix (3) conditioned on the knowledge of w^*. For this, we first define the quantities

$$\begin{aligned}\beta_{\ell,w^*} &= 2\mathbf{P}_{Z\sim\mathcal{N}(0,1)}\left(\mu - (L-\ell)\delta + \sigma Z > w^* \middle| w^*\right) - 1, \text{ and} \\ \beta_\ell &= 2\mathbf{P}_{Z,Z'\sim\mathcal{N}(0,1)}\left(\mu + \sigma Z > \mu - \ell\delta + \sigma Z'\right) - 1 = 2\Phi\left(\frac{\ell\delta}{\sqrt{2}\sigma}\right) - 1\end{aligned} \tag{5}$$

for any $\ell \in \mathbb{R}$ and $w^* \in \mathbb{R}$. Observe that $\beta_{\ell,w^*} = \mathbf{E}\left[2\mathbb{I}_{(w_{ij}>w^*)} - 1 \middle| w^*\right]$ when $\ell_{ij}^{lca} = \ell$, whereas $\beta_\ell = \mathbf{E}\left[2\mathbb{I}_{(w_{ij}>w_{kl})} - 1\right]$ when $\ell_{ij}^{lca} - \ell_{kl}^{lca} = \ell$. In particular, $\beta_0 = 0$. Based on (5) and the observation that the product terms in (3) are independent conditioned on w^*, we write for any $i \neq j$,

$$\mathbf{E}\left[K_{ij} \middle| w^*\right] = \sum_{k\neq i,j} q\beta_{\ell_{ik}^{lca},w^*}\beta_{\ell_{jk}^{lca},w^*}.$$

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 (3), observe that each term in the sum depends only on the groups containing i, j, k, 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. We may perform 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, $\mathbf{E}[K_{ij}|w^*] = (1 \pm \epsilon)a$ denotes $(1-\epsilon)a \leq \mathbf{E}[K_{ij}|w^*] \leq (1+\epsilon)a$. Allowing a fluctuation of $(1 \pm \epsilon)$ also helps to ignore the small effect of the case where (i, k) or (j, k) corresponds to (i_0, j_0), that is, the reference pair for which $w^* = w_{i_0 j_0}$. Thus, for i, j such that $i \neq j$ and $\ell_{ij}^{lca} = \ell$, we have

$$\begin{aligned}\mathbf{E}[K_{ij}|w^*] &= (1 \pm \epsilon)qN_0 \sum_{r=1}^{2^L} \beta_{\ell^{lca}(i,\mathcal{G}_r),w^*}\beta_{\ell^{lca}(j,\mathcal{G}_r),w^*} \\ &= (1 \pm \epsilon)qN_0 \sum_{t,t'=0}^{L} \beta_{t,w^*}\beta_{t',w^*}\#\{r : \ell^{lca}(i,\mathcal{G}_r) = t, \ell^{lca}(j,\mathcal{G}_r) = t'\},\end{aligned} \tag{6}$$

where the second equality explicitly mentions that we need to count the number of different pure clusters that are merged with i or j at different levels of the true hierarchy. We now consider different

cases. First, if i, j belong to same group, then $\ell = L$ and $\ell^{lca}(i, \mathcal{G}_r) = \ell^{lca}(j, \mathcal{G}_r)$ for every r. So,

$$\kappa_L := \mathbf{E}[K_{ij}|w^*] = (1 \pm \epsilon) q N_0 \sum_{t=0}^{L} (2^{L-1-t} \vee 1) \beta_{t,w^*}^2,$$

which we denote by a quantity κ_L noting that it only depends on the level L and not on i, j. Here, $\vee$ denotes the maximum of two values. The numbers of clusters are computed based on the fact that there is only one cluster at levels L or $L-1$, and otherwise 2^{L-1-t} groups are merged with i at level-t. If i, j are not in the same group, that is, $\ell = \ell_{ij}^{lca} < L$, then we observe:

- if $t < \ell$, then for any $\mathcal{G}_r$ such that $\ell^{lca}(i, \mathcal{G}_r) = t$, we also have $\ell^{lca}(j, \mathcal{G}_r) = t$. So we may only consider cases $t = t'$ when $t < \ell$.
- 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 do not need to consider $t = t' = \ell$, which is the main difference from the case $\ell = L$.
- if $t > \ell$, then for any $\mathcal{G}_r$ with $\ell^{lca}(i, \mathcal{G}_r) = t$, we have $\ell^{lca}(j, \mathcal{G}_r) = \ell$. So we may set $t' = \ell$ whenever we have $t > \ell$. Similarly, we should also count the cases $t' > \ell, t = \ell$.

Thus, we can decompose the summation into three parts based on the conditions on t, t' ($t = t' < \ell$; $t > \ell, t' = \ell$; $t = \ell, t' > \ell$). For each case, we should count $\#\{r : \ell^{lca}(i, \mathcal{G}_r) = t, \ell^{lca}(j, \mathcal{G}_r) = t'\}$. To compute these, we note that $\#\{r : \ell^{lca}(i, \mathcal{G}_r) = \ell^{lca}(j, \mathcal{G}_r) = t\} = 2^{L-1-t}$ when $t = t' < \ell$ as used above. But when $t > \ell, t' = \ell$, we have $\#\{r : \ell^{lca}(i, \mathcal{G}_r) = t, \ell^{lca}(j, \mathcal{G}_r) = \ell\} = 2^{L-1-t} \vee 1$ since this counts only those groups which merge with i at level-t, and t' plays no role in the count. A similar argument holds for the case $t = \ell, t' > \ell$. Based on this, we compute $\mathbf{E}[K_{ij}|w^*]$ for the case $\ell_{ij}^{lca} = \ell < L$ and denote the expected value by κ_ℓ, noting that it does not depend on i, j. We have

$$\kappa_\ell := \mathbf{E}[K_{ij}|w^*] = (1 \pm \epsilon) q N_0 \left[\sum_{t=0}^{\ell-1} 2^{L-1-t} \beta_{t,w^*}^2 + 2 \sum_{t=\ell+1}^{L} (2^{L-1-t} \vee 1) \beta_{t,w^*} \beta_{\ell,w^*} \right],$$

where the second term, counted twice, corresponds to both the cases of $t > \ell$ or $t' > \ell$, which behave similarly. Since $\beta_{t,w^*} \in [-1, 1]$, one can easily verify that $|\kappa_\ell| \leq qN$ for all ℓ.

The above discussion leads to the conclusion that $\mathbf{E}[K|w^*]$ has a block diagonal structure with exactly $L+1$ distinct off-diagonal entries, $\kappa_0, \ldots, \kappa_L$, and the block structure corresponds to the planted hierarchy shown in Figure 1 in the main paper (right). We now show that these distinct terms are sufficiently separated, that is, $\kappa_{\ell+1} - \kappa_\ell$ is large for every $\ell = 0, 1, \ldots, L-1$. To derive this, we require a lower bound on

$$\begin{aligned}
\beta_{t+1,w^*} - \beta_{t,w^*} &= 2\mathbf{P}\left(\mu - (L-t-1)\delta + \sigma Z > w^* \middle| w^*\right) - 2\mathbf{P}\left(\mu - (L-t)\delta + \sigma Z > w^* \middle| w^*\right) \\
&= \sqrt{\frac{2}{\pi}} \int_{(w^* - \mu + (L-t-1)\delta)/\sigma}^{(w^* - \mu + (L-t)\delta)/\sigma} e^{-z^2/2} \mathrm{d}z \\
&\geq \sqrt{\frac{2}{\pi}} \frac{\delta}{\sigma} e^{-\left(a^2 \vee (a-\delta)^2\right)/2\sigma^2},
\end{aligned}$$

where $a = w^* - \mu + (L-t)\delta$. Conditioned on the bounds w^* stated in (4), one can see that

$$a^2 \vee (a-\delta)^2 < 2(L+1)^2\delta^2 + 4\sigma^2 \ln\left(\frac{8}{\eta}\right),$$

where we use the fact that $t \in [0, L]$ and the inequality $(x+y)^2 \leq 2(x^2+y^2)$. Plugging this into the above derivation shows that $\beta_{t+1,w^*} - \beta_{t,w^*} > \Delta$ for any $t \in [0, L]$, where Δ is defined in the statement of theorem. We use the above bound to show that

$$\begin{aligned}
\kappa_L - \kappa_{L-1} &> qN_0 \left(\beta_{L,w^*}^2 - \beta_{L-1,w^*}^2\right)^2 - 2\epsilon qN \\
&> qN_0\Delta^2 - q2^{L+3},
\end{aligned}$$

where the second term, involving ϵ, takes care of the fluctuation due to our approximate computations of κ_ℓ and is simply bounded by the upper bound on κ_ℓ. Similarly, for any $\ell < L-1$,

$$\begin{aligned}\kappa_{\ell+1} - \kappa_\ell &> qN_0 \Bigg[2^{L-1-\ell}\beta_{\ell,w^*}^2 - 2^{L-1-\ell}\beta_{\ell,w^*}\beta_{\ell+1,w^*} \\ &\qquad + 2\sum_{t=\ell+2}^{L} (2^{L-1-t} \vee 1)\beta_{t,w^*}(\beta_{\ell+1,w^*} - \beta_{\ell,w^*}) \Bigg] - 2\epsilon qN \\ &= qN_0 2 \sum_{t=\ell+2}^{L} (2^{L-1-t} \vee 1)(\beta_{t,w^*} - \beta_{\ell,w^*})(\beta_{\ell+1,w^*} - \beta_{\ell,w^*}) - 2\epsilon qN \\ &> 2^{L-\ell-1} qN_0 \Delta^2 - q2^{L+3},\end{aligned}$$

where the equality follows since $2^{L-1-\ell} = 2\sum_{t=\ell+2}^{L}(2^{L-1-t} \vee 1)$, and subsequently, we note that $\beta_{t,w^*} - \beta_{\ell,w^*} > \beta_{\ell+1,w^*} - \beta_{\ell,w^*} > \Delta$ for all $t \geq \ell+2$. Hence, we can conclude that for $N_0 > \frac{4}{\Delta}\sqrt{N}$, or equivalently, $N_0 > \frac{2^{L+4}}{\Delta^2}$,

$$\kappa_{\ell+1} - \kappa_\ell > \frac{qN_0\Delta^2}{2} \tag{7}$$

for all $\ell = 0, 1, \ldots, L-1$. We subsequently show that under the condition on q assumed in the theorem, with probability $1 - \frac{\eta}{4}$,

$$K_{ij} - \mathbf{E}[K_{ij}|w^*] < \frac{qN_0\Delta^2}{4} \tag{8}$$

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. We complete the proof by deriving the concentration result of (8). From (3), we observe that, conditioned on w^*, the entry K_{ij} is a sum of $N-2$ independent random variables each lying in the range $[-1, 1]$. Hence, a direct application of Bernstein's inequality implies that

$$\mathbf{P}\left(\left| K_{ij} - \mathbf{E}[K_{ij}|w^*] \right| > \sqrt{3qN \ln\left(\frac{4N^2}{\eta}\right)} \bigvee 3\ln\left(\frac{4N^2}{\eta}\right) \middle| w^* \right) \leq \frac{\eta}{2N^2}\,.$$

Using the symmetry of K and the union bound, it follows that the above entry-wise concentration holds for all $i \neq j$ with probability at least $1 - \frac{\eta}{4}$. Finally, for $q > C\frac{2^{2L}}{N\Delta^4}\ln\left(\frac{N}{\eta}\right)$ with $C > 0$ large enough, it is easy to verify that $\frac{1}{4}qN_0\Delta^2$ is larger than the deviation obtained using Bernstein's inequality. The above argument leads to the claim of Theorem 2.

To verify the claim for fixed L and $\frac{\delta}{\sigma}$, we note that, in this case, Δ is constant and $N_0 = \Omega(N)$. Hence, using $q = \frac{c\ln N}{N}$ for a large enough constant c immediately leads to the exact recovery guarantee and number of comparisons. □

A.4 Analysis of Passive Quadruplets Kernel based Average Linkage (4K–AL)

In the passive setting, we do not have the freedom of querying specific comparisons but have access to only a pre-computed set of quadruplet comparisons $\mathcal{Q} \subset \{(i,j,k,l) : w_{ij} > w_{kl}\}$. Hence, we use a variant of the kernel in (3), which relies only on passively obtained comparisons.

$$K_{ij} = \sum_{\substack{k,l=1 \\ k<l}}^{N} \sum_{r=1}^{N} \left(\mathbb{I}_{(i,r,k,l)\in\mathcal{Q}} - \mathbb{I}_{(k,l,i,r)\in\mathcal{Q}} \right) \left(\mathbb{I}_{(j,r,k,l)\in\mathcal{Q}} - \mathbb{I}_{(k,l,j,r)\in\mathcal{Q}} \right). \tag{9}$$

In principle, the above kernel extends the actively computed kernel (3) by using all $\binom{N}{2}$ pairs of (k,l) as references in comparison to only one used in (3). However, each term in the sum only contributes

when we simultaneously observe the comparisons between (i, r) and (k, l) and between (j, r) and (k, l).

In the following, we assume that the model for obtaining passive comparisons is the one described in Section 2.3 of the main paper. For every tuple (i, r, k, l), we assume that with probability $p \in (0, 1]$, there is a comparison $w_{ir} \gtrless w_{kl}$ and based on the comparison either $(i, r, k, l) \in \mathcal{Q}$ or $(k, l, i, r) \in \mathcal{Q}$. We also assume that the observation of the quadruplet comparisons are independent. Based on this model, we define a set of i.i.d. Bernoullis $\{\xi_{irkl} \sim \text{Bernoulli}(p) : i, r, k, l \text{ such that } i < r, k < l, (i, r) < (k, l)\}$, where we order the indices/ index pairs to avoid repeated counting of the same tuple. It follows that $|\mathcal{Q}| = \sum_{i,r,k,l} \xi_{irkl}$, and from Bernstein's inequality, it follows that $|\mathcal{Q}| = \mathcal{O}\left(pN^4\right)$ with high probability. Using this notation, we may re-write the kernel function in (9) as

$$K_{ij} = \sum_{k<l} \sum_{r \neq i,j} \xi_{irkl}\xi_{jrkl} \left(\mathbb{I}_{(w_{ir}>w_{kl})} - \mathbb{I}_{(w_{ir}<w_{kl})}\right)\left(\mathbb{I}_{(w_{jr}>w_{kl})} - \mathbb{I}_{(w_{jr}<w_{kl})}\right).$$

We now restate and prove the exact recovery guarantee for average linkage with the aforementioned kernel.

Theorem 3 (Exact recovery of planted hierarchy by 4K–AL with passive comparisons). *Let $\eta \in (0, 1)$ and $\Delta = \frac{\delta}{2\sigma} e^{-L^2\delta^2/4\sigma^2}$. There exists an absolute constant $C > 0$ such that if $N_0 > \frac{8}{\Delta}\sqrt{N}$ and we set*

$$p > \max\left\{C\frac{2^L}{\Delta^2}\sqrt{\frac{1}{N}\ln\left(\frac{N}{\eta}\right)}, \frac{2}{N^4}\ln\left(\frac{2}{\eta}\right)\right\},$$

then with probability at least $1 - \eta$, the 4K–AL algorithm exactly recovers the planted hierarchy using at most pN^4 quadruplet comparisons, which are passively obtained based on the model described in Section 2.3 (of the main paper).

In particular, if $L = \mathcal{O}(1)$, the above statement implies that even with $\frac{\delta}{\sigma}$ constant, 4K–AL exactly recovers the planted hierarchy with probability $1 - \eta$ using $\mathcal{O}\left(N^{7/2}\ln N\right)$ passive comparisons.

Proof. **The upper bound on the number of comparisons** follow by noting that $|\mathcal{Q}|$ is a sum of $\binom{\binom{N}{2}}{2}$ i.i.d. Bernoullis, and hence, the bound of pN^4 holds with probability $1 - \frac{\eta}{2}$ for $p > \frac{2}{N^4}\ln(\frac{2}{\eta})$.

The proof for exact recovery has a similar structure as that of Theorem 2, the only difference being that the analysis does not depend on a fixed reference pair. In particular, we can write the expected entries of the kernel matrix in (9) as

$$\begin{aligned}\mathbf{E}[K_{ij}] &= \sum_{k<l} \sum_{r \neq i,j} p^2\big(2\mathbf{P}(w_{ir} > w_{kl}) - 1\big)\big(2\mathbf{P}(w_{jr} > w_{kl}) - 1\big) \\ &= \frac{1}{2}\sum_{k \neq l} \sum_{r \neq i,j} p^2 \beta_{\ell_{ir}^{lca} - \ell_{kl}^{lca}} \beta_{\ell_{jr}^{lca} - \ell_{kl}^{lca}},\end{aligned}$$

where β is defined in (5). As in the proof of Theorem 2, we show that $\mathbf{E}[K_{ij}]$ can take at most $L + 1$ distinct values depending on the level ℓ_{ij}^{lca}. As before, we decompose the above summation depending on ℓ_{ir}^{lca}, ℓ_{jr}^{lca} and ℓ_{kl}^{lca}, and also allow a fluctuation of $(1 \pm \epsilon)$ with $\epsilon = \frac{8}{N_0}$ to take care of minor effects of ignoring cases such as $k = l$ or $r = i, j$. We write the expectation in terms of the clusters as

$$\begin{aligned}\mathbf{E}[K_{ij}] &= \frac{(1 \pm \epsilon)}{2} p^2 N_0^3 \sum_{r,k,l=1}^{2^L} \beta_{\ell^{lca}(i,\mathcal{G}_r) - \ell^{lca}(\mathcal{G}_k,\mathcal{G}_l)} \beta_{\ell^{lca}(j,\mathcal{G}_r) - \ell^{lca}(\mathcal{G}_k,\mathcal{G}_l)} \\ &= \frac{(1 \pm \epsilon)}{2} p^2 N_0^3 \sum_{s,t,t'=0}^{L} \beta_{t-s}\beta_{t'-s} \times \\ &\qquad \#\{r : \ell^{lca}(i, \mathcal{G}_r) = t, \ell^{lca}(j, \mathcal{G}_r) = t'\} \#\{k, l : \ell^{lca}(\mathcal{G}_k, \mathcal{G}_l) = s\} \\ &= (1 \pm \epsilon) p^2 N_0^3 2^{L-1} \sum_{s,t,t'=0}^{L} (2^{L-1-s} \vee 1)\beta_{t-s}\beta_{t'-s} \#\{r : \ell^{lca}(i, \mathcal{G}_r) = t, \ell^{lca}(j, \mathcal{G}_r) = t'\}\end{aligned}$$

The last step holds since every cluster $\mathcal{G}_l$ is merged with $(2^{L-1-s} \vee 1)$ clusters at level-s, and hence, $\#\{k,l : \ell^{lca}(\mathcal{G}_k, \mathcal{G}_l) = s\} = 2^L(2^{L-1-s} \vee 1)$.

We now compute $\kappa_\ell = \mathbf{E}[K_{ij}]$ where $\ell = \ell_{ij}^{lca}$. For, $\ell = L$, that is, when i, j belong to the same cluster, $\ell^{lca}(i, \mathcal{G}_r) = \ell^{lca}(j, \mathcal{G}_r)$ for every cluster. Hence,

$$\kappa_L = (1 \pm \epsilon) p^2 N_0^3 2^{L-1} \sum_{s,t=0}^{L} (2^{L-1-s} \vee 1)(2^{L-1-t} \vee 1)\beta_{t-s}^2 \,.$$

For $\ell_{ij}^{lca} = \ell < L$, we have three possible cases as mentioned in the proof of Theorem 2: $(t = t' < \ell)$; $(t > \ell, t' = \ell)$; and $(t = \ell, t' > \ell)$. Decomposing the summation based on these cases and noting that $(t > \ell, t' = \ell)$ and $(t = \ell, t' > \ell)$ lead to similar terms, we have

$$\kappa_\ell = (1 \pm \epsilon) p^2 N_0^3 2^{L-1} \sum_{s=0}^{L} (2^{L-1-s} \vee 1) \left[\sum_{t=0}^{\ell-1} 2^{L-1-t} \beta_{t-s}^2 + 2 \sum_{t=\ell+1}^{L} (2^{L-1-t} \vee 1) \beta_{t-s} \beta_{\ell-s} \right]$$

for every $\ell = 0, 1, \ldots, L-1$. We now derive a lower bound on the separation $\kappa_{\ell+1} - \kappa_\ell$, which depends on the observation that $|\kappa_\ell| \le \frac{1}{2} p^2 N^3$ for every ℓ, and a lower bound on

$$\begin{aligned}
\beta_{t+1-s} - \beta_{t-s} &\ge \min_{r \in [-L, L-1]} \beta_{r+1} - \beta_r \\
&= \min_{r \in [-L, L-1]} \sqrt{\frac{2}{\pi}} \int_{r\delta/\sqrt{2}\sigma}^{(r+1)\delta/\sqrt{2}\sigma} e^{-z^2/2} \mathrm{d}z \\
&> \frac{1}{\sqrt{\pi}} \frac{\delta}{\sigma} e^{-L^2\delta^2/4\sigma^2}.
\end{aligned}$$

The lower bound is larger than Δ stated in the theorem. Based on this bound and noting that $2^L = \sum_{s=0}^{L} (2^{L-1-s} \vee 1)$, we obtain

$$\begin{aligned}
\kappa_L - \kappa_{L-1} &> p^2 N_0^3 2^{L-1} \sum_{s=0}^{L} (2^{L-s-1} \vee 1)(\beta_{L-s} - \beta_{L-1-s})^2 - \epsilon p^2 N^3 \\
&> \frac{1}{2^{L+1}} p^2 N^3 \Delta^2 - p^2 N^2 2^{L+3},
\end{aligned}$$

which is at least $\frac{1}{2^{L+2}} p^2 N^3 \Delta^2$ if $N > \frac{2^{2L+5}}{\Delta^2}$, or equivalently, $N_0 > \frac{4\sqrt{2}}{\Delta}\sqrt{N}$. Similarly, for $\ell < L - 1$, we have

$$\begin{aligned}
\kappa_{\ell+1} - \kappa_\ell &> p^2 N_0^3 2^{L-1} \sum_{s=0}^{L} (2^{L-1-s} \vee 1) \Big[2^{L-1-\ell} \beta_{\ell-s}^2 - 2^{L-1-\ell} \beta_{\ell+1-s} \beta_{\ell-s} \\
&\qquad + 2 \sum_{t=\ell+2}^{L} (2^{L-1-t} \vee 1) \beta_{t-s} (\beta_{\ell+1-s} - \beta_{\ell-s}) \Big] - \epsilon p^2 N^3 \\
&> p^2 N_0^3 2^L \sum_{s=0}^{L} \sum_{t=\ell+2}^{L} (2^{L-1-s} \vee 1)(2^{L-1-t} \vee 1) \Delta^2 - p^2 N^2 2^{L+3} \\
&= p^2 N_0^3 2^{3L-\ell-2} \Delta^2 - p^2 N^2 2^{L+3} \\
&> \frac{p^2 N^3 \Delta^2}{2^{\ell+2}} \,.
\end{aligned}$$

The second step follows by using $2 \sum_{t=\ell+2}^{L} (2^{L-1-t} \vee 1) = 2^{L-1-\ell}$ and $\beta_{\ell+1-s} - \beta_{\ell-s} > \Delta$. The third step computes the summation, and the fourth holds when $N_0 > \frac{8}{\Delta}\sqrt{N}$. Thus for every ℓ, we obtain a minimum separation

$$\kappa_{\ell+1} - \kappa_\ell > \frac{1}{2^{L+2}} p^2 N^3 \Delta^2.$$

Following the proof idea of Theorem 2, it only remains to show that the fluctuation of $|K_{ij} - \mathbf{E}[K_{ij}]|$ is less than half of this minimum separation for all $i < j$ since, under this scenario, one can argue that entries of K corresponding to different levels of the planted hierarchy are well-separated, and hence, the planted hierarchy is exactly recovered by average linkage. Thus, to complete the proof, we derive the following concentration inequality

$$\mathbf{P}\left(\left|K_{ij} - \mathbf{E}[K_{ij}]\right| > \sqrt{2p^2N^5 \ln\left(\frac{2N^4}{\eta}\right)} \bigvee 2N^2 \ln\left(\frac{2N^4}{\eta}\right)\right) \leq \frac{\eta}{N^2}\,. \tag{10}$$

By union bound, it follows that with probability $1 - \frac{\eta}{2}$, the above bound holds for all $i < j$, whereas setting $p > C\frac{2^L}{\Delta^2}\sqrt{\frac{1}{N}\ln\left(\frac{N}{\eta}\right)}$ for $C > 0$ large enough ensures that the deviation is smaller than $\frac{1}{2^{L+3}}p^2N^3\Delta^2$. To derive (10), we note that $K_{ij} - \mathbf{E}[K_{ij}] = \sum_{k<l}\sum_{r\neq i,j} B_{rkl}$ is a sum of $\binom{N}{2}(N-2)$ random variables, where we use $B_{rkl} \in [-1, 1]$ to denote each term in the summation. One can verify that B_{rkl} has zero mean and its variance is smaller that p^2. Moreover, each B_{rkl} is dependant on all $(N-3)$ random variables, $\{B_{r'kl} : r' \neq r\}$, and all $\binom{N}{2} - 1$ random variables, $\{B_{rk'l'} : (k', l') \neq (k, l)\}$. Hence, if we draw a dependency graph among these random variable, we obtain a regular graph with the vertex degree of each node being $(N + \binom{N}{2} - 4) < N^2$. We use the concentration technique described in Section 2.3.2 of Janson and Ruciński (2002), where the key observation is that for any graph with maximum degree d, one can find an equitable colouring with $d + 1$ colours, that is a colouring where all colour classes (independent sets) differ in size by at most one. In the present context, it implies that one can split the set of random variables into at most N^2 subsets, $\mathcal{C}_1, \ldots, \mathcal{C}_{N^2}$ such that each subset contains at most $\frac{\binom{N}{2}(N-3)}{N^2} < \frac{N}{2}$ variables that are mutually independent. Hence, we can apply union bound followed by Bernstein's inequality to write

$$\begin{aligned}
\mathbf{P}\left(\left|K_{ij} - \mathbf{E}[K_{ij}]\right| > \tau\right) &\leq \mathbf{P}\left(\bigcup_{s=1}^{N^2}\left|\sum_{(r,k,l)\in\mathcal{C}_s} B_{rkl}\right| > \frac{\tau}{N^2}\right) \\
&\leq \sum_{s=1}^{N^2}\mathbf{P}\left(\left|\sum_{(r,k,l)\in\mathcal{C}_s} B_{rkl}\right| > \frac{\tau}{N^2}\right) \\
&\leq 2N^2\exp\left(-\frac{\frac{\tau^2}{N^4}}{p^2N + \frac{2}{3}\frac{\tau}{N^2}}\right) \leq 2N^2\exp\left(-\frac{\tau^2}{2p^2N^5} \bigvee \frac{\tau}{2N^2}\right).
\end{aligned}$$

For $\tau = \sqrt{2p^2N^5\ln\left(\frac{2N^4}{\eta}\right)} \bigvee 2N^2\ln\left(\frac{2N^4}{\eta}\right)$, the probability is smaller than $\frac{\eta}{N^2}$, which results in the conclusion of (10).

To verify the claim for fixed L and $\frac{\delta}{\sigma}$, we note that in this case, Δ is constant and $N_0 = \Omega(N)$. Hence, using $p = c\sqrt{\frac{\ln N}{N}}$ for a large enough constant c immediately leads to the exact recovery guarantee and the number of passive comparisons. □

A.5 Analysis of Quadruplets based Average Linkage (4–AL)

The proposed 4–AL algorithms estimates 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, we define

$$\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4) = \sum_{x_i\in G_1}\sum_{x_j\in G_2}\sum_{x_k\in G_3}\sum_{x_l\in G_4}\frac{\mathbb{I}_{(i,j,k,l)\in\mathcal{Q}} - \mathbb{I}_{(k,l,i,j)\in\mathcal{Q}}}{|G_1|\,|G_2|\,|G_3|\,|G_4|}\,. \tag{11}$$

Based on our model for passive comparisons, where $\xi_{ijkl} \sim$ Bernoulli(p) is the indicator for observing tuple (i, j, k, l), we may re-write the preference relation in (11) as

$$\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4) = \sum_{x_i\in G_1}\sum_{x_j\in G_2}\sum_{x_k\in G_3}\sum_{x_l\in G_4}\frac{\xi_{ijkl}\left(\mathbb{I}_{(w_{ij}>w_{kl})} - \mathbb{I}_{(w_{ij}<w_{kl})}\right)}{|G_1|\,|G_2|\,|G_3|\,|G_4|}\,.$$

Subsequently, we use the above preference relation $\mathbb{W}_{\mathcal{Q}}$ to define a similarity function W in the following way. 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

$$W(G_p, G_q) = \sum_{r,s=1, r\neq s}^{K} \frac{\mathbb{W}_{\mathcal{Q}}(G_p, G_q \| G_r, G_s)}{K(K-1)}. \tag{12}$$

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 the hierarchical clustering principle (Algorithm 1 in the main paper), results in the proposed 4–AL algorithm. Below, we restate and prove the exact recovery guarantee for 4–AL using passively obtained quadruplet comparisons.

Theorem 4 (Exact recovery of planted hierarchy by 4–AL with passive comparisons). *Let $\eta \in (0,1)$ and $\Delta = \frac{\delta}{2\sigma} e^{-L^2\delta^2/4\sigma^2}$. Assume the following:*
(i) An initial step partitions $\mathcal{X}$ into pure clusters of sizes in the range $[m, 2m]$ for some $m \leq \frac{1}{2}N_0$.
(ii) $\mathcal{Q}$ is a passively obtained set of quadruplet comparisons, where each tuple (i,j,k,l) is observed independently with probability $p > \frac{C}{m\Delta^2} \max\left\{\ln N, \frac{1}{m}\ln\left(\frac{1}{\eta}\right)\right\}$ for some constant $C > 0$.
Then, with probability $1-\eta$, starting from the given initial partition and using $|\mathcal{Q}| \leq pN^4$ number of passive comparisons, 4–AL exactly recovers the planted hierarchy.
In particular, if $L = \mathcal{O}(1)$, the above statement implies that, when $\frac{\delta}{\sigma}$ is a constant, 4–AL exactly recovers the planted hierarchy with probability $1-\eta$ using $\mathcal{O}\left(\frac{N^4 \ln N}{m}\right)$ passive comparisons.

Proof. The bound $|\mathcal{Q}| < pN^4$ with probability $1 - \frac{\eta}{2}$ is derived similarly to the bound on $|\mathcal{Q}|$ in Theorem 3. Hence, we only prove the exact recovery guarantee.

We first analyze the algorithm under expectation. 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

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

Now, consider $p, q, p', q' \in \{1, \ldots, K\}$ such that $p \neq q$, $p' \neq q'$, $\ell^{lca}(G_p, G_q) = \ell + 1$ and $\ell^{lca}(G_{p'}, G_{q'}) = \ell$ for some $\ell \in \{0, 1, \ldots, L-1\}$. 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

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

where the last step follows from arguments used in the proof of Theorem 3, which show that $\min_{\ell \in [-L, L-1]} \beta_{\ell+1} - \beta_\ell > \Delta$, where β_ℓ is defined in (5) and Δ is in the statement of the theorem.

Chaining of the above argument shows that $\mathbf{E}[W(G_p, G_q)] - \mathbf{E}[W(G_{p'}, G_{q'})] > p\Delta$ whenever $\ell^{lca}(G_p, G_q) > \ell^{lca}(G_{p'}, G_{q'})$. Under the assumptions stated in Theorem 4, we later prove that with probability $1 - \frac{\eta}{2}$,

$$\left|W(G, G') - \mathbf{E}[W(G, G')]\right| \leq \frac{p\Delta}{2} \tag{13}$$

for every pair of clusters G, G' formed during the agglomerative steps of the algorithm starting from the given pure clusters of size in the range $[m, 2m]$. Based on (13) 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 planted hierarchy by the algorithm.

We now prove (13). 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 in the range $[m, 2m]$, such that G_1, G_2 are disjoint and so are G_3, G_4. Then

$$\mathbf{P}\left(|\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4) - \mathbf{E}[\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4)]| > \frac{p\Delta}{2}\right) \leq 2\exp\left(2\ln N - \frac{p\Delta^2 m^2}{C'}\right) \tag{14}$$

for some absolute constant $C' > 0$. We wish to use (14) to argue that with probability $1 - \frac{\eta}{2}$, all clusters in the initial partition (assumed in the theorem) satisfy the condition $|\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4) - \mathbf{E}[\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4)]| \leq \frac{p\Delta}{2}$. Note that we do not know how the initial partition is achieved, but we can ensure that

$$\begin{aligned}
&\mathbf{P}\Bigg(\exists G_1, G_2, G_3, G_4 : m \leq |G_1|, |G_2|, |G_3|, |G_4| \leq 2m, \\
&\qquad\qquad |\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4) - \mathbf{E}[\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4)]| > \frac{p\Delta}{2}\Bigg) \\
&\leq \sum_{i_1, i_2, i_3, i_4 = m}^{2m} \binom{N}{i_1}\binom{N}{i_2}\binom{N}{i_3}\binom{N}{i_4} 2\exp\left(2\ln N - \frac{p\Delta^2 m^2}{C'}\right) \\
&\leq 2m^4 \left(\frac{eN}{m}\right)^{8m} \exp\left(2\ln N - \frac{p\Delta^2 m^2}{C'}\right). \\
&\leq C'' \exp\left(9m \ln N - \frac{p\Delta^2 m^2}{C'}\right),
\end{aligned}$$

where $C'' > 0$ is an absolute constant such that $\sup_{m \geq 1} 2m^4 (\frac{e}{m})^{2m} < C''$. The above probability is bounded by $\frac{\eta}{2}$ for $p > \frac{C}{m\Delta^2}\left(\ln N \vee \frac{1}{m}\ln\left(\frac{1}{\eta}\right)\right)$ for some constant $C > 0$. Thus, with probability $1 - \frac{\eta}{2}$, we know that for every tuple of four clusters, obtained at initialization, $\mathbb{W}_{\mathcal{Q}}$ deviates from its mean by at most $\frac{p\Delta}{2}$. 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

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

which is a convex combination of $\mathbb{W}_{\mathcal{Q}}$ computed at the previous stage. Hence, if each of them deviates from its mean by at most $\frac{p\Delta}{2}$, then the convex combination after merging also deviates from its mean by at most $\frac{p\Delta}{2}$. The same also holds for other instances of merging throughout the hierarchy, which shows that with probability $1 - \frac{\eta}{2}$, at any stage of agglomeration, we have $|\mathbb{W}_{\mathcal{Q}}(G_p, G_q \| G_r, G_s) - \mathbf{E}[\mathbb{W}_{\mathcal{Q}}(G_p, G_q \| G_r, G_s)]| < \frac{p\Delta}{2}$ 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, (13) holds.

We complete the proof of Theorem 4 by proving the concentration inequality in (14). Since $w_{ij} = w_{kl}$ occurs with zero probability for any $i, j, k, l (i \neq j, k \neq l)$, we may write

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

where ξ_{ijkl} is the indicator of observing the comparison between (i, j) and (k, l). Note that each term in the summation is a centred random variable in the range $[-1, 1]$, and has variance bounded by p. Let us denote each of them by B_{ijkl}, and observe that they have dependencies among themselves. We use the concentration technique of Janson and Ruciński (2002). Consider the dependency graph for these variables, which is a graph on $s = |G_1||G_2||G_3||G_4|$ 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 $\lfloor \frac{s}{d+1} \rfloor$ or $\lceil \frac{s}{d+1} \rceil$. Denoting the colour classes by $\mathcal{C}_1, \ldots, \mathcal{C}_{d+1}$, we can bound the probability using the union bound and Bernstein's inequality as

$$\begin{aligned}
&\mathbf{P}\left(|\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4) - \mathbf{E}[\mathbb{W}_{\mathcal{Q}}(G_1, G_2 \| G_3, G_4)]| > \frac{p\Delta}{2} \right) \\
&= \mathbf{P}\left(\left| \sum_{i,j,k,l} B_{ijkl} \right| > \frac{sp\Delta}{4} \right) \\
&\leq \sum_{\ell=1}^{d+1} \mathbf{P}\left(\left| \sum_{(i,j,k,l) \in \mathcal{C}_\ell} B_{ijkl} \right| > \frac{sp\Delta}{4(d+1)} \right) \\
&\leq \sum_{\ell=1}^{d+1} 2 \exp\left(- \frac{\frac{s^2 p^2 \Delta^4}{16(d+1)^2}}{2p|\mathcal{C}_\ell| + \frac{2}{3} \frac{sp\Delta}{4(d+1)}} \right).
\end{aligned}$$

The bound in (14) follows by first noting that $|\mathcal{C}_\ell| \leq \frac{2s}{d+1}$, and then using the fact $\frac{s}{d+1} \geq \min\{|G_1||G_2|, |G_3||G_4|\} \geq m^2$. For the outer summation, we simply use $(d + 1) \leq N^2$ to obtain the bound in (14).

To verify the claim for fixed L and $\frac{\delta}{\sigma}$, we note that, in this case, Δ is constant and $N_0 = \Omega(N)$. Hence, using $p = \frac{c \ln N}{m}$ for a large enough constant c immediately leads to the exact recovery guarantee and the number of passive comparisons. □

B Details on the experiments

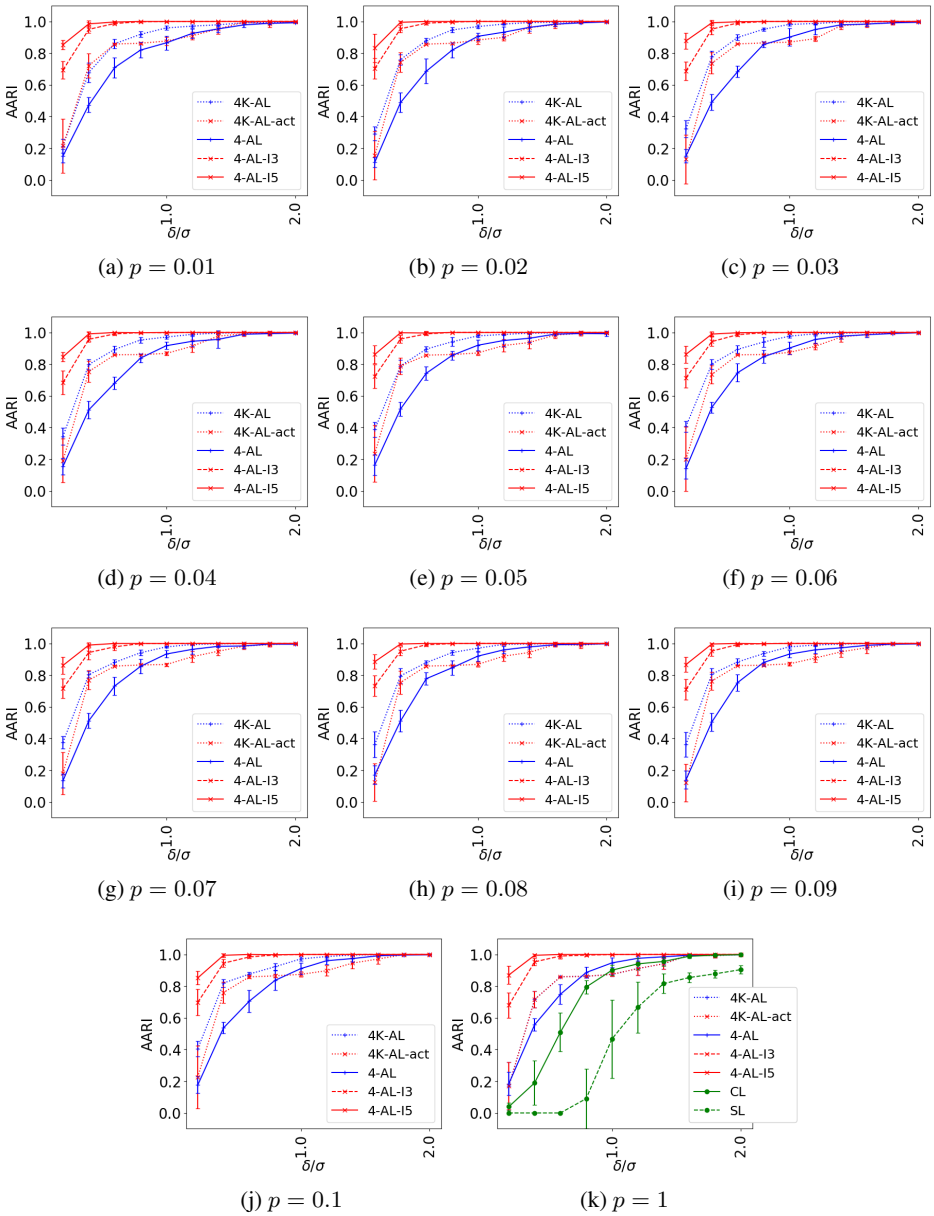
In this section we present some details on the experiments that are not included in the main paper along with some additional plots and discussions.

B.1 Planted Hierarchical Model

Evaluation function. As a measure of performance we report the Averaged Adjusted Rand Index (AARI) between the ground truth hierarchy $\mathcal{C}$ 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:

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

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. $\text{AARI}(\mathcal{C}, \mathcal{C}') = 1$ implies that the two hierarchies are identical. For all the experiments we report the mean and the standard deviation of 10 repetitions.



(a) $p = 0.01$ (b) $p = 0.02$ (c) $p = 0.03$

(d) $p = 0.04$ (e) $p = 0.05$ (f) $p = 0.06$

(g) $p = 0.07$ (h) $p = 0.08$ (i) $p = 0.09$

(j) $p = 0.1$ (k) $p = 1$

Figure S.1: 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 sampling proportions p. Best viewed in color.

Results. In Figure S.1 we present supplementary results for the planted hierarchical model, that is with $p \in \{0.01, 0.02, \ldots, 0.1, 1\}$. Firstly, similar to the theory, SL can hardly recover the planted hierarchy, even for large values of $\frac{\delta}{\sigma}$. CL performs better than SL, which is not evident from the theory. This suggests that a better sufficient condition might be possible for CL. We observe that 4K–AL, 4K–AL–act, and, 4–AL are able to exactly recover the true hierarchy for smaller signal-to-noise ratio and their performances do not degrade much when the number of sampled comparisons is reduced. Finally, as expected, the best methods are 4–AL–I3 and 4–AL–I5. They use large initial clusters but recover the true hierarchy even for very small values of $\frac{\delta}{\sigma}$.

B.2 Standard Clustering Datasets

Data. We provide some details on the datasets used in the paper. We evaluate the different approaches on 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 (it originally contains 101 animals but we chose to remove the 'girl' entry since we feel that it does not fit in a Zoo dataset). 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 generate the quadruplet comparisons 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. Since it is not realistic to assume that all the comparisons are available, we use the procedure described in Section 2.3 in the main paper to passively obtain a proportion $p \in \{0.01, 0.02, \dots, 0.1\}$ of all the quadruplets. Note that tSTE-AL and FORTE-AL are based on ordinal embedding methods that use triplet comparisons of the form "object i is more similar to object j than to object k", that is $w_{ij} > w_{ik}$, rather than quadruplet comparisons. Nevertheless, we can use the same procedure than for the quadruplets to generate the same proportion of triplets that we can use in tSTE and FORTE. To the best of our knowledge, there does not exist ordinal embedding methods based only on quadruplet comparisons.

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 methods. Instead we use the recently proposed Dasgupta's cost (Dasgupta, 2016) that has been specifically designed to evaluate hierarchical clustering methods. Given a base similarity measure w, the cost of a hierarchy $\mathcal{C}$ is

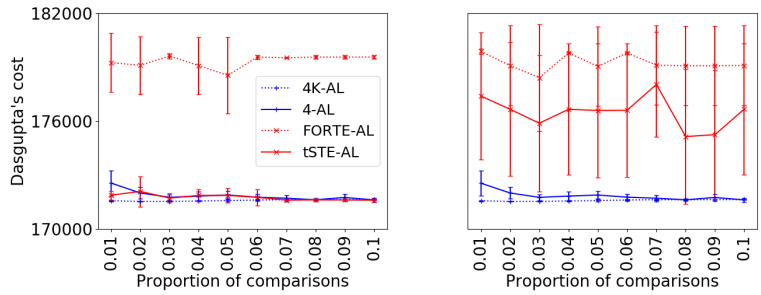
$$\text{cost}(\mathcal{C}, w) = \sum_{x, x_j \in \mathcal{X}} w_{ij} \left| \mathcal{C}^{lca}(x_i, x_j) \right|$$

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. The idea of this cost is that similar objects that are merged higher in the hierarchy should be penalized. Hence, a lower cost indicates a better 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.

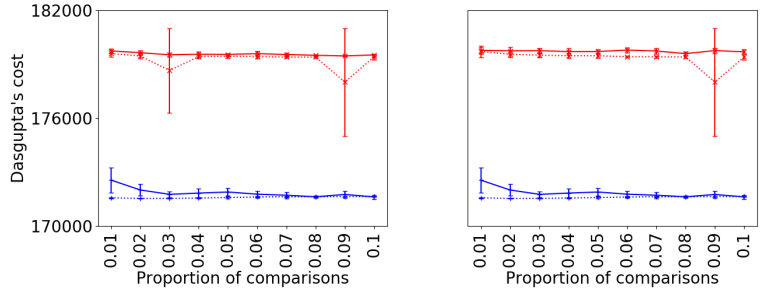
Results. In Figures S.2, S.3, and S.4 we present supplementary results for the standard clustering datasets. We note that the proportion of comparisons does not have a large impact as the results are, on average, stable across all regimes. Our methods are either comparable or better than the embedding-based ones. Our methods do not need to first embed the examples and thus do not impose a strong Euclidean structure on the data. The impact of this structure is more or less pronounced depending on the dataset. Furthermore, the performance of tSTE-AL and FORTE-AL depends on the embedding dimension that should be carefully chosen. For example, on Zoo, the performance of tSTE drops with increasing dimension. Similarly, on Glass, FORTE seems to perform slightly better for larger dimensions. Unfortunately, in clustering, tuning parameters can be difficult as there is no ground-truth.

B.3 Comparison-based datasets

The Car dataset (Kleindessner and von Luxburg, 2017) 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 (j, i, k) and (k, i, j). Then triplet comparisons further translate into quadruplet comparisons by noticing that the triplet (i, j, k) corresponds to the quadruplet (i, j, i, k). Overall we obtained 12112 comparisons that we used to learn a hierarchy among the cars.

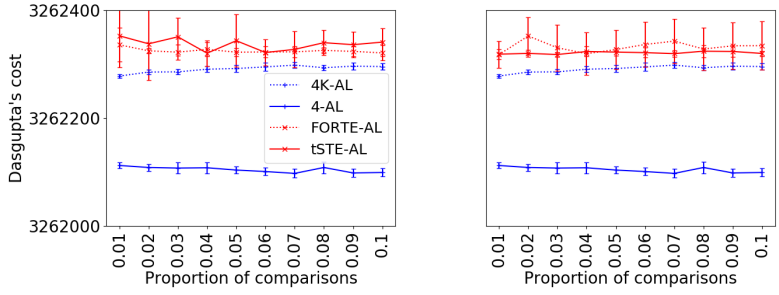


(a) Dimension $= 2$ (b) Dimension $= 3$

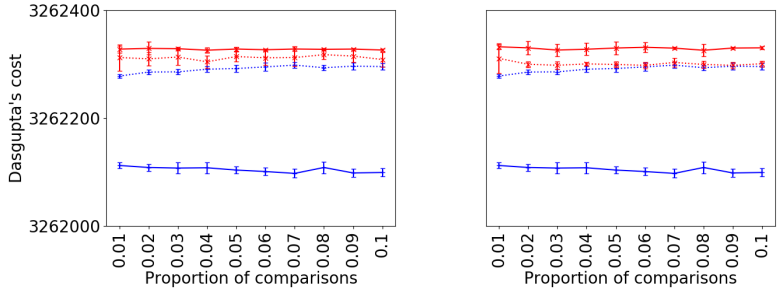


(c) Dimension $= 4$ (d) Dimension $= 5$

Figure S.2: 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.

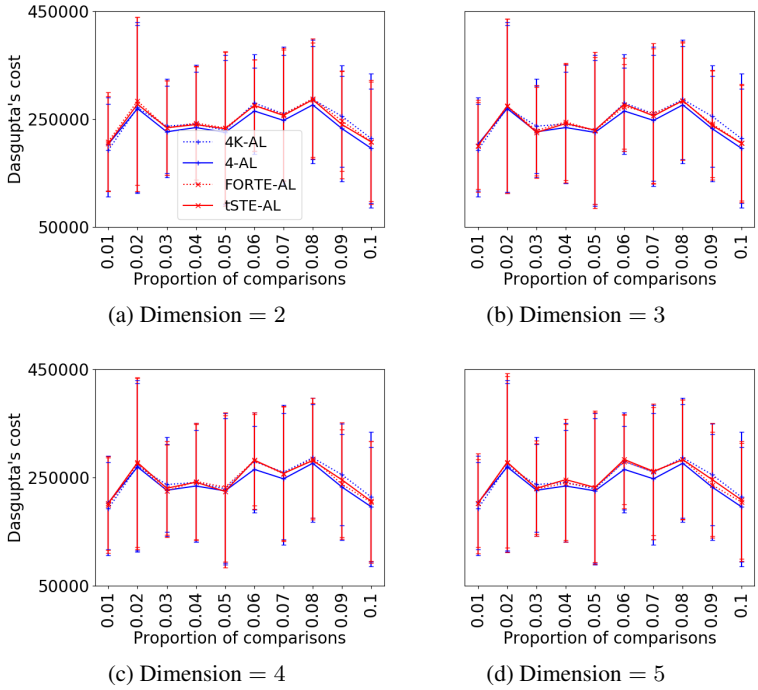


(a) Dimension $= 2$ (b) Dimension $= 3$



(c) Dimension $= 4$ (d) Dimension $= 5$

Figure S.3: 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 $= 5$

Figure S.4: 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.

The hierarchies obtained by 4K–AL, 4–AL, FORTE–AL and tSTE–AL are attached to this supplementary as png files. The names of the files are respectively cars.4K–AL.png, cars.4–AL.png, cars.FORTE–AL.embedding_dimension.png and cars.tSTE–AL.embedding_dimension.png.

References

Blum, A., Hopcroft, J., and Kannan, R. (2018). *Foundations of data science*. Available online at https://www.cs.cornell.edu/jeh/book.pdf.

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.

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

Heller, K. A. and Ghahramani, Z. (2005). Bayesian hierarchical clustering. In *InternationalConference on Machine Learning*, pages 297–304.

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

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

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

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