A Revenue Function for Comparison-Based Hierarchical Clustering

Anonymous authors
Paper under double-blind review

Abstract

Comparison-based learning addresses the problem of learning when, instead of explicit features or pairwise similarities, one only has access to comparisons of the form: *Object A is more similar to B than to C.* Recently, it has been shown that, in Hierarchical Clustering, single and complete linkage can be directly implemented using only such comparisons while several algorithms have been proposed to emulate the behaviour of average linkage. Hence, finding hierarchies (or dendrograms) using only comparisons is a well understood problem. However, evaluating their meaningfulness when no ground-truth nor explicit similarities are available remains an open question.

In this paper, we bridge this gap by proposing a new revenue function that allows one to measure the goodness of dendrograms using only comparisons. We show that this function is closely related to Dasgupta's cost for hierarchical clustering that uses pairwise similarities. On the theoretical side, we use the proposed revenue function to resolve the open problem of whether one can approximately recover a latent hierarchy using few triplet comparisons. On the practical side, we present principled algorithms for comparison-based hierarchical clustering based on the maximisation of the revenue and we empirically compare them with existing methods.

1 Introduction

In the past decade, there has been an exponential growth in the scope of data science and machine learning in domains such as psycho-physics (Shepard, 1962; Stewart et al., 2005; Haghiri et al., 2020) or cultural psychology (Berenhaut et al., 2022), evolutionary biology (Foulds et al., 1979; Semple and Steel, 2003; Catanzaro, 2009), or crowd-sourcing (Heikinheimo and Ukkonen, 2013; Ukkonen, 2017) among others. A type of data that has recently gained some traction in these contexts is *comparisons* (Stewart et al., 2005; Agarwal et al., 2007), particularly in the form of:

Triplet comparison: Binary response to the query—is object i more similar to object j than to object k?

Quadruplet comparison: Binary response to the query —are objects i and j more similar to each other than objects k and l?

Comparisons have been used in the psycho-physics literature for more than 50 years since it is known that humans can provide relative measurements better than absolute ones (Shepard, 1962; Stewart et al., 2005). It led to a surge of popularity of comparisons in the context of crowdsourced data about objects that cannot be represented by Euclidean features, such as food (Wilber et al., 2014) or musical artists (Ellis et al., 2002), or objects for which humans cannot robustly estimate a pairwise similarity, for instance cars (Kleindessner and von Luxburg, 2017) or natural scenes (Heikinheimo and Ukkonen, 2013). The purpose of collecting comparisons is often to learn patterns in the objects, such as latent clusters, or use them for prediction, as in classification. Hence, there has been significant development of algorithms for comparison-based learning (Agarwal et al., 2007; Heikinheimo and Ukkonen, 2013; Haghiri et al., 2017; Kazemi et al., 2018; Perrot and von Luxburg, 2019).

The present paper focuses on comparison-based hierarchical clustering. Clustering refers to partitioning a dataset into groups of similar objects, while hierarchical clustering is the problem of finding partitions of the data at different levels of granularity. It is natural to wonder how one can group similar objects or find a hierarchy of groups when neither features nor pairwise similarities are available, and one only has access to triplet or quadruplet comparisons. For instance, the objects in the food dataset (Wilber et al., 2014) can be broadly categorised into 'sweets or desserts' and 'main or savoury dishes', but the latter can be further sub-divided into meat dishes, soups and others. Surprisingly, interest in comparison-based (hierarchical) clustering stemmed in 1970s when single linkage clustering gained popularity, and researchers realised that the method uses only ordinal informal instead of absolute values of pairwise similarities (Janowitz, 1971; Sibson, 1972; Janowitz, 1979). Around the same time, works also started on the *consensus tree problem*, that is, constructing trees (hierarchies) from given sub-trees or ordinal relations (Adams III, 1972; Aho et al., 1981). The problem has since evolved as an important topic in both computational biology and computer science, most notably addressing the question of phylogenetic tree reconstruction under triplet or other ordinal constraints Semple and Steel (2003); Wu (2004); Snir and Yuster (2011). More recently, Ghoshdastidar et al. (2019) re-discovered that single and complete linkage can be computed using only few *actively chosen* quadruplet comparisons. There are, however, limited practical settings where the learning algorithm can actively decide which comparisons should be queried, and the most relevant case is that of learning from a set of *passively collected* comparisons. This is certainly true in the known practical applications of comparison-based hierarchical clustering, such as (hierarchical) clustering of objects from crowd-sourced comparisons (Ukkonen, 2017; Kleindessner and von Luxburg, 2017), finding communities in languages or in cultural psychology (Berenhaut et al., 2022), and constructing relational database queries Aho et al. (1981) or phylogenetic trees (Semple and Steel, 2003).

One of the fundamental problems in hierarchical clustering is to evaluate the *goodness* of a hierarchy. This issue is obviously inherent to identifying *better* hierarchical clustering algorithms. In the phylogenetics literature, the optimal hierarchy problem typically corresponds to the *minimum evolution problem*, where, given a set of species and a pairwise distance matrix (representing evolutionary distance between species), the goal is to find a weighted tree with minimal total edge weights that preserve the evolutionary distances (Foulds et al., 1979; Catanzaro, 2009). A similar philosophy exists in the early works on hierarchical clustering, where an algorithm is judged to better if the ultrametric induced by the output tree is closer to the specified pairwise dissimilarities among the given objects (Janowitz, 1979). More recently, there has been efforts to mathematically quantify the goodness of a hierarchy in terms of certain *cost or revenue functions* (Dasgupta, 2016; Moseley and Wang, 2017; Wang and Wang, 2020). Such formulations have led to a plethora of new methods for hierarchical clustering that also come with worst-case approximation guarantees (Cohen-Addad et al., 2019; Charikar et al., 2019; Chatziafratis et al., 2021).

Motivation for this work and our contributions. The main motivation for this work stems from the lack of cost or revenue functions that can be used in the comparison-based framework. Available goodness measures for trees can only be defined using pairwise (dis)similarities (Dasgupta, 2016; Moseley and Wang, 2017; Wang and Wang, 2020). Hence, existing works on comparison-based hierarchical clustering either demonstrate the meaningfulness of the computed hierarchies visually or in artificial settings, where comparisons are derived from pairwise similarities (Kleindessner and von Luxburg, 2017; Ghoshdastidar et al., 2019). Neither solution is useful in practice, where one only has access to comparisons. In this paper, we propose new revenue functions for dendrograms that are only based on triplet or quadruplet comparisons (Section 4). We show that the proposed comparison-based revenues are equivalent to Dasgupta's cost or revenue (Dasgupta, 2016; Moseley and Wang, 2017) applied to particular pairwise similarities that can be computed from comparisons. Interestingly, the pairwise similarities that arise from this equivalence are known in the comparison-based clustering literature (Perrot et al., 2020).

Section 5 demonstrates that the proposed revenue function meaningfully captures the goodness of a hierarchical tree. For this purpose, we consider the problem of reconstructing a latent hierarchy (for example, a phylogeny tree) from ordinal constraints (Emamjomeh-Zadeh and Kempe, 2018). In particular, we show that, when all possible triplets among the objects are available, the dendrogram corresponding to the latent hierarchy maximises the proposed revenue function. This, in turn, implies that one can mathematically formulate the triplet-based hierarchical clustering problem as a *maximum triplet comparison revenue prob-*

lem. We further address the question of whether one can approximately recover the latent hierarchy using fewer than $\Omega(n^3)$ triplets. This problem has not been directly addressed in previous works (see Section 2). We show that only $O(n^2 \log n/\epsilon^2)$ passive triplets suffice to obtain a $(1-\epsilon)$-approximation of the optimal revenue.

Finally, Sections 6–7 use the connection of the proposed revenue functions to the *additive similarities* in Perrot et al. (2020) to present two variants of average linkage hierarchical clustering based on passive triplet or quadruplet comparisons. The performance of these approaches is empirically compared with state of the art baselines using synthetic and real datasets.

2 Related Work

In this section, we briefly review the algorithmic developments of comparison-based hierarchical clustering, as well as existing theoretical results related to this problem. As noted earlier, interest in comparison-based hierarchical clustering stemmed from different applications. The current literature consists of two lines of research—works related to reconstruction of phylogenetic trees (Wu, 2004; Snir and Yuster, 2011; Chatziafratis et al., 2021) and those focusing on ordinal data analysis from crowd-sourced data (Kleindessner and von Luxburg, 2017; Ghoshdastidar et al., 2019).

In ordinal data analysis literature, the most widely used principle is that of ordinal embedding, where the underlying idea is to retrieve Euclidean representations of the objects that respect the available comparisons as well as possible (see the review in Vankadara et al. (2019) for more details). The embedded data can be subsequently used for (hierarchical) clustering. While this principle provides flexibility in the choice of clustering methods, the Euclidean restriction of the underlying data often leads to inaccurate representations, and hence, poor performance in the context of hierarchical clustering (Ghoshdastidar et al., 2019). The restrictive assumption of Euclidean embedding is avoided by computing pairwise similarities from available comparisons (Kleindessner and von Luxburg, 2017; Ghoshdastidar et al., 2019). Standard hierarchical clustering algorithms, such as average linkage, can then be applied using the pairwise similarities.

An alternative approach for comparison-based (hierarchical) clustering is to define an appropriate cost or objective based on comparison and directly optimise it. Ukkonen (2017) employs such a technique for clustering using crowd-sourced data, while this principle underlies most techniques in consensus tree problems or phylogenetic tree reconstruction. In the latter context, two well-studied optimisation problems are *maximum rooted triplet consistency* (Wu, 2004; Byrka et al., 2010)—finding a hierarchy that satisfies most, if not all, given triplets—and *maximum quartet consistency* (Snir and Yuster, 2011; Jiang et al., 2000)—where one has access to quartets (sub-trees with four leaves indicating which pairs should be merged first) and the problem is to find a tree that satisfies most given quartets.[1] Other related optimisation problems as well as various constraints other than triplets or quartets have been studied (Snir and Rao, 2010; Chatziafratis et al., 2021). Since the focus of the present paper is to define a revenue for trees (see Section 4), our work naturally belongs to this broad class of hierarchical clustering algorithms based on revenue maximisation. However, in Theorem 1, we relate the proposed revenues to pairwise similarities computed from comparisons. Hence, the present paper connects the optimisation principle to the aforementioned approach of defining pairwise similarities from comparisons.

Prior works on comparison-based hierarchical clustering provide a range of computational and statistical results. On the computational side, it is known that both the problems of maximum rooted triplet consistency and maximum quartet consistency are NP-hard (Byrka et al., 2010; Snir and Yuster, 2011). However, polynomial-time constant factor approximation algorithms are known in both cases, assuming that a uniformly random subset of triplets/quartets is available. For triplets, Wu (2004) provides a $\frac{1}{3}$-approximation algorithm—a fraction at least $\frac{1}{3}$ of the given triplets are satisfied—which is slightly improved in Byrka et al. (2010). For quartets, polynomial-time algorithms that satisfy at least a $(1-\epsilon)$-fraction of the given quartets are known (Jiang et al., 2000; Snir and Yuster, 2011). While the above results focus on finding hierarchies

---

[1] Note that quartets are different from quadruplets though both are defined on four objects. More precisely, a quartet on i, j, k, l corresponds to information that i, j and k, l should be merged in the tree before all four are merged. Using the notation from Section 3, a quadruplet (i, j, k, l) only implies $s_{ij} > s_{kl}$ whereas a quartet on i, j, k, l implies $\min\{s_{ij}, s_{kl}\} > \max\{s_{ik}, s_{il}, s_{jk}, s_{jl}\}$.

that only match the given triplets/quartets, Emamjomeh-Zadeh and Kempe (2018) show that the true (latent) hierarchy can be recovered only if $\Omega(n^3)$ passive (uniformly sampled) triplets are available. In contrast, only $O(n \log n)$ triplets suffice if they are actively queried. In Section 5, we show that only $O(n^2 \log n/\epsilon^2)$ uniformly sampled triplets suffice to obtain a $(1-\epsilon)$-approximation of the optimal triplet revenue.

A different latent model is considered in Ghoshdastidar et al. (2019) and Perrot et al. (2020), where the objects have latent (noisy) pairwise similarities that have a (hierarchical) cluster structure. Noisy triplets/quadruplets are uniformly sampled following the noisy latent similarities. While Ghoshdastidar et al. (2019) focus on quadruplet-based hierarchical clustering and show that $O(n^{3.5} \log n)$ suffice to recover the latent hierarchy, Perrot et al. (2020) show that flat latent clusters can be exactly recovered using only $O(n^2 \log n)$ uniformly sampled triplets/quadruplets. Although our model is different from Perrot et al. (2020), we obtain a similar $O(n^2 \log n)$ upper bound on sample complexity—even when the triplets are noisy.

3 Preliminaries

We consider the problem of hierarchical clustering of a set of n objects, denoted by $[n] = \{1, 2, \ldots, n\}$. In this paper, we assume that a hierarchy or dendrogram on $[n]$ is a binary tree H whose root node is the set $[n]$, each leaf node is a singleton containing one of the n objects, and each internal node represents a set $C \subseteq [n]$ with its two children, C_1 and C_2, denoting a partition of C, that is $\min(|C_1|, |C_2|) > 0$, $C = C_1 \cup C_2$, and $C_1 \cap C_2 = \emptyset$. In the following, we use binary tree or tree to designate a hierarchy. For node C, we use $H(C)$ to denote the subtree rooted at C and $|H(C)|$ represents the number of leaves in the subtree, or equivalently, the number of objects in the set C. For objects $i, j \in [n]$, let $i \vee j$ denote the smallest node in the tree containing both i and j, and $H(i \vee j)$ denote the smallest subtree containing both i and j. The goal of hierarchical clustering is to find a dendrogram H that is *optimal*, or at least *good*, in some sense. In the next subsections, we recall Dasgupta's cost for hierarchical clustering that allows one to measure the goodness of a dendrogram given full access to pairwise similarities, and then describe the comparison-based learning framework, where only triplet or quadruplet comparisons are available. In the paper, we use the standard Landau notations $O(\cdot)$, $\Omega(\cdot)$, $o(\cdot)$, where the asymptotics are defined with respect to n.

3.1 Dasgupta's Cost for Hierarchical Clustering

Suppose one has access to a function $s : [n] \times [n] \to \mathbb{R}$, symmetric, such that $s_{ij} = s(i,j)$ denotes the pairwise similarity between objects $i, j \in [n]$. Dasgupta's cost function (Dasgupta, 2016) for a dendrogram H on $[n]$, with respect to the pairwise similarity s, is defined as

$$Dcost(H, s) = \sum_{i,j \in [n],\ i<j} s_{ij} \cdot |H(i \vee j)| \,. \tag{1}$$

An equivalent definition of the above cost can be found in Wang and Wang (2020), where the cost is expressed in terms of triplets of objects instead of pairs. To find a good dendrogram, Dasgupta (2016) proposed to minimize this cost over all trees. While it is NP-hard to find the optimal solution, several relaxations are known to have constant factor approximation guarantees for Dasgupta's cost or related quantities. In particular, Moseley and Wang (2017) defined Dasgupta's revenue function,

$$Drev(H, s) = n \sum_{i,j \in [n],\ i<j} s_{ij} - Dcost(H, s). \tag{2}$$

Note that since $\sum s_{ij}$ is fixed, maximising $Drev(H, s)$ over all binary trees is equivalent to minimising $Dcost(H, s)$. It can then be shown that the revenue of the tree H obtained from average linkage achieves a revenue $Drev(H, s)$ that is at least $\frac{1}{3}$ of the revenue of the optimal tree that would achieve the best revenue, provided that the similarity function s is non-negative.

3.2 Comparison-based Learning

In the present paper, we assume that the pairwise similarities $\{s_{ij}\}_{i,j\in[n]}$ are not available. Instead the algorithm has access to either a set of triplets $\mathcal{T}$, which is a subset of

$$\mathcal{T}_{all} = \{(i,j,k) \in [n]^3 : s_{ij} > s_{ik},\ i,j,k \text{ distinct}\},$$

or a set of quadruplets $\mathcal{Q} \subseteq \mathcal{Q}_{all}$, where

$$\mathcal{Q}_{all} = \{(i,j,k,l) \in [n]^4 :\ s_{ij} > s_{kl}, i < j,\ k < l,\ (i,j) \neq (k,l)\}.$$

Since the pairwise similarities are assumed symmetric, we set $i < j$ and $k < l$ to avoid considering the same comparison multiple times. We note that the number of possible comparisons is high—$|\mathcal{Q}_{all}| = O(n^4)$ and $|\mathcal{T}_{all}| = O(n^3)$—but, in practice, the observed comparisons, $\mathcal{T}$ or $\mathcal{Q}$ may be fewer than that, about $O(n^2)$ comparisons, as can be seen from Table 3. Note that whenever a triple (i,j,k) is considered, $\mathcal{T}$ contains either (i,j,k) or (i,k,j), depending on whether $s_{ij} \lessgtr s_{ik}$. The same holds for quadruplets. We further assume that the observed set of comparisons $\mathcal{T}$, or $\mathcal{Q}$, is passively collected, that is the algorithm cannot decide which comparisons should be present in it—as opposed to the active setting where the algorithm can choose which comparisons should be observed (Ghoshdastidar et al., 2019).

4 Comparison-based Revenue

We present two comparison-based revenue functions for hierarchical clustering, one in the triplets framework and the other for quadruplet comparisons.

Triplet comparisons. We first consider the case of triplets and assume that the algorithm has access to a passively collected set of triplets $\mathcal{T}$. We define the triplet comparison revenue of a binary tree (dendrogram) H on $[n]$, using triplets $\mathcal{T}$, as

$$Trev(H,\mathcal{T}) = \sum_{(i,j,k)\in\mathcal{T}} \Big(|H(i \vee k)| - |H(i \vee j)|\Big). \tag{3}$$

For every $(i,j,k) \in \mathcal{T}$, we know that i is more similar to j than to k, and hence, we prefer to merge i,j before merging i and k. It means the ideal tree should have $|H(i \vee k)| > |H(i \vee j)|$ for every $(i,j,k) \in \mathcal{T}$. Hence, it is desirable to maximise $|H(i \vee k)| - |H(i \vee j)|$ for every observed triplet $(i,j,k) \in \mathcal{T}$. We then propose to formulate triplets comparison-based hierarchical clustering as the problem of maximizing $Trev(H,\mathcal{T})$ over all binary trees.

Remark 1. *We note that the proposed triplet revenue is significantly different from the triplet based cost presented in Wang and Wang (2020). The most important distinction is that the triplet cost in Wang and Wang (2020) is a reformulation of Dasgupta's cost, and requires knowledge of pairwise similarities. In contrast, the revenue in equation 3 is computed only from triplet comparisons without access to pairwise similarities.*

Quadruplet comparisons. The above formulation can be similarly stated in the quadruplets setting. Assuming that the algorithm has access to a passively collected set of quadruplets $\mathcal{Q}$, we define the quadruplet comparison revenue of a binary tree H on $[n]$ as

$$Qrev(H,\mathcal{Q}) = \sum_{(i,j,k,l)\in\mathcal{Q}} \Big(|H(k \vee l)| - |H(i \vee j)|\Big). \tag{4}$$

Similar to the triplet setting, every $(i,j,k,l) \in \mathcal{Q}$ indicates that i,j should be merged earlier than k,l in H, and we prefer trees such that $|H(k \vee l)| \geq |H(i \vee j)|$. We propose to achieve this by finding a tree that maximises $Qrev(H,\mathcal{Q})$.

Connection with Dasgupta's cost. While one may try to directly maximise the above comparison-based revenue functions, the following equivalence to Dasgupta's cost and revenue allows us to employ existing methods for hierarchical clustering that require pairwise similarities. In the following, let $\mathbb{I}_E$ denote the indicator of event E, that is, $\mathbb{I}_E = 1$ if E happens, and 0 otherwise.

Theorem 1. *For any given set of triplets $\mathcal{T}$ and any dendrogram H on $[n]$,*

$$Trev(H, \mathcal{T}) = -Dcost(H, s^{AddS3}) = Drev(H, s^{AddS3}),$$

where s^{AddS3} refers to the additive similarity from triplets (AddS3) defined by Perrot et al. (2020)

$$s_{ij}^{AddS3} = \sum_{k \neq i,j} \left(\mathbb{I}_{(i,j,k)\in\mathcal{T}} - \mathbb{I}_{(i,k,j)\in\mathcal{T}} + \mathbb{I}_{(j,i,k)\in\mathcal{T}} - \mathbb{I}_{(j,k,i)\in\mathcal{T}} \right)$$

Similarly, for any set of quadruplets $\mathcal{Q}$ and dendrogram H,

$$Qrev(H, \mathcal{Q}) = -Dcost(H, s^{AddS4}) = Drev(H, s^{AddS4}),$$

where s^{AddS4} is the additive similarity from quadruplets (AddS4) defined by Perrot et al. (2020)

$$s_{ij}^{AddS4} = \sum_{k \neq l,\ (k,l) \neq (i,j)} \left(\mathbb{I}_{(i,j,k,l)\in\mathcal{Q}} - \mathbb{I}_{(k,l,i,j)\in\mathcal{Q}} \right).$$

Proof sketch (details in appendix). Proving $Trev(H, \mathcal{T}) = -Dcost(H, s^{AddS3})$ involves a rearrangement of terms, with the observation that, for every i, j, the term $|H(i \vee j)|$ appears in the summation in equation 3 with coefficient -1 when $(i, j, k) \in \mathcal{T}$ or $(j, i, k) \in \mathcal{T}$ and with coefficient $+1$ when $(i, k, j) \in \mathcal{T}$ or $(j, k, i) \in \mathcal{T}$. Adding these coefficients for all $k \neq i, j$ gives us $-s_{ij}^{AddS3}$, and proves the equality. The second equality $-Dcost(H, s^{AddS3}) = Drev(H, s^{AddS3})$ simply follows from the observation that $\sum_{i<j} s_{ij}^{AddS3} = 0$. The proof for quadruplets is similar. ☐

5 Recovering a Latent Hierarchy by Triplet Revenue Maximisation

In this section, we consider the problem of recovering a latent hierarchy from triplet comparisons, earlier studied in Emamjomeh-Zadeh and Kempe (2018). Let H_0 be a hierarchy on $[n]$, from which we derive a set of triplets[2]

$$\mathcal{T}_0 = \{(i, j, k), (j, i, k) \ : \ |H_0(i \vee j)| < \min(|H_0(i \vee k)|, |H_0(j \vee k)|)\}.$$

One can show that any rooted tree H' that satisfies all triplets in $\mathcal{T}_0$ is equivalent to H_0, up to isomorphic transformations, and hence, one can exactly recover H_0 given $\mathcal{T}_0$. Emamjomeh-Zadeh and Kempe (2018) further show one cannot exactly recover H_0 from any passively obtained $\mathcal{T} \subseteq \mathcal{T}_0$ if $|\mathcal{T}| = o(n^3)$. This raises the question—can one approximately recover H_0 from a smaller set of triplets $\mathcal{T}$?

We use the proposed triplets-based revenue to answer this question in the affirmative. Before providing an approximation guarantee, we first show the significance of our formulation in this context by proving that one can recover H_0 from $\mathcal{T}_0$ by maximising $Trev$.

Proposition 2. *Consider the aforementioned setting, where H_0 is a hierarchy on $[n]$ objects, and $\mathcal{T}_0$ is the corresponding set of triplets as defined above. Then*

$$H_0 = \arg\max_H \ Trev(H, \mathcal{T}_0),$$

where the maximisation is over all binary trees H on $[n]$.

---

[2] Emamjomeh-Zadeh and Kempe (2018) consider triples of the form $\{i, j, k\}$ that imply i, j are closer to each other than k, with respect to H_0. Each such triple $\{i, j, k\}$ correspond to two triplets (i, j, k) and (j, i, k) in our setting.

Proof sketch (details in appendix). The proof is in two steps. First, we show that $\mathcal{T}_0$ is the unique optimal set of triplets to maximize the revenue for the tree H_0, that is $\mathcal{T}_0 = \arg\max_{\mathcal{T}} Trev(H_0, \mathcal{T})$. Then, we show that, given two trees H_0 and H_1 such that $\mathcal{T}_0$ and $\mathcal{T}_1$ are the corresponding sets of triplets then we have that $Trev(H_0, \mathcal{T}_1) = Trev(H_1, \mathcal{T}_0)$. Combining these two results, we obtain that $Trev(H_0, \mathcal{T}_0) > Trev(H_0, \mathcal{T}_1) = Trev(H_1, \mathcal{T}_0)$ for any tree H_1 and corresponding set of triplets $\mathcal{T}_1$. This directly implies Proposition 2. ☐

Emamjomeh-Zadeh and Kempe (2018) prove the uniqueness of the hierarchy that satisfies all the triplets in $\mathcal{T}_0$, that is, H_0 maximises the function

$$f(H, \mathcal{T}_0) = \sum_{(i,j,k)\in\mathcal{T}_0} \mathbb{I}_{|H(i\vee k)|>|H(i\vee j)|}.$$

While maximising $Trev(H, \mathcal{T}_0)$ seems to be a relaxation of maximising $f(H, \mathcal{T}_0)$ in this context, Proposition 2 shows that both problems have the same optimal solution H_0.

5.1 Approximate Recovery of H_0 Using Passive Triplets

We consider the setting, where $\mathcal{T}_0$ is not completely available but one has access to a uniformly sampled subset $\mathcal{T} \subseteq \mathcal{T}_0$. We show that $|\mathcal{T}| = O(n^2 \log n/\epsilon^2)$ triplets suffice to obtain a tree $\widehat{H}$ such that $Trev(\widehat{H}, \mathcal{T}_0) \geq (1-\epsilon)\cdot Trev(H_0, \mathcal{T}_0)$, that is, we get a good approximation of H_0 with much fewer than n^3 samples, although we may not exactly recover H_0.

We consider the following uniform sampling to obtain $\mathcal{T}$. Let $p_n \in (0,1]$ denote a sampling probability, depending on n. For every pair of triplets $(i,j,k),(j,i,k) \in \mathcal{T}_0$, we add the pair to $\mathcal{T}$ with probability p_n. One can use standard concentration inequalities to show that when p_n is sufficiently large then $|\mathcal{T}| = O(p_n n^3)$ with high probability. We state the following approximation guarantee for trees derived using $\mathcal{T}$.

Theorem 3. *Let $p_n \in (0,1]$ that depends on n, and $\mathcal{T}$ be as defined above. Consider the hierarchy*

$$\widehat{H} = \arg\max_{H} \ Trev(H, \mathcal{T}).$$

For any $\epsilon \in (0,1)$, there exist constants $n_0, c, c' > 0$ (depending on ϵ) such that for $n > n_0$ and $p_n > c\log n/n\epsilon^2$,

$$Trev(\widehat{H}, \mathcal{T}_0) \geq (1-\epsilon)\cdot Trev(H_0, \mathcal{T}_0) \tag{5}$$

with probability at least $1 - n^{-c'}$. The condition on p_n implies that $|\mathcal{T}| = \Theta(n^2 \log n/\epsilon^2)$ triplets suffice to achieve equation 5.

Proof sketch (details in appendix). Using concentration results, we show that with probability $1 - n^{-c'}$, $\left|Trev(H, \mathcal{T}) - p_n Trev(H, \mathcal{T}_0)\right| = O\left(n^3\sqrt{p_n n \log n}\right)$ for every tree H. Using this concentration for both H_0 and $\widehat{H}$, and noting that $Trev(\widehat{H}, \mathcal{T}) \geq Trev(H_0, \mathcal{T})$, we have

$$Trev(\widehat{H}, \mathcal{T}_0) \geq Trev(H_0, \mathcal{T}_0) - O\left(\sqrt{\frac{n^7 \log n}{p_n}}\right).$$

Finally, we show that $Trev(H_0, \mathcal{T}_0) = \Omega(n^4)$ to arrive at the $(1-\epsilon)$-approximation for $p_n = \Omega(\log n/n\epsilon^2)$. ☐

The result in Emamjomeh-Zadeh and Kempe (2018, Proposition 2.2)—that $\Omega(n^3)$ triplets are necessary to exactly recover H_0—hinges on the fact that it is impossible to correctly guess the hierarchy at the lowest level of the tree H_0 using fewer comparisons. Since errors in the lowest level do not significantly affect $Trev$, we can achieve the $(1-\epsilon)$-approximation in Theorem 3. However, note that $\widehat{H}$ may not be efficiently computable as it requires exhaustive search over all trees. We discuss practical algorithms in the next section.

Theorem 3 holds only in the noiseless setting, where it is assumed that every observed triplet in $\mathcal{T}$ is correct. It is natural to ask if Theorem 3 still holds under a noise setting, where some triplets may be flipped with some probability. To formalise this, let $\mathcal{T} \subseteq \mathcal{T}_0$ be a set of triplets obtained from the sampling procedure in Theorem 3. Let $\mathcal{T}'$ be constructed such that, for every $(i,j,k) \in \mathcal{T}$, $\mathcal{T}'$ contains (i,j,k) with probability $1-\alpha$, or (i,k,j) with probability α. The random flipping of labels is independent for all $(i,j,k) \in \mathcal{T}$. We obtain the following corollary from a minor modification of the above proof (details in appendix).

Corollary 4. *For a fixed flipping probability* $\alpha \in (0, \frac{1}{2})$, *the statement of Theorem 3 holds for the maximizer of* $Trev(H, \mathcal{T}')$, *that is, with probability* $1 - n^{-c'}$, $\max_H \; Trev(H, \mathcal{T}') \geq (1-\epsilon) \cdot Trev(H_0, \mathcal{T}_0)$.

6 Comparison-based Algorithms for Hierarchical Clustering

The equivalence between comparison-based revenues and Dasgupta's revenue, stated in Theorem 1, implies that one may simply employ standard hierarchical clustering algorithms using the pairwise similarities AddS3 or AddS4, depending on whether one has access to triplets or quadruplets. This makes it possible to use the well-established literature on hierarchical clustering with pairwise similarities. In fact, as mentioned before, previous works on passive comparison-based hierarchical clustering also follow this philosophy using other kind of pairwise similarities obtained from the comparisons (Kleindessner and von Luxburg, 2017; Ghoshdastidar et al., 2019). Unlike previous works, our use of AddS3 or AddS4 stems from a revenue maximisation formulation that allows us to consider an approach based on the average linkage (AL) clustering algorithm, that is, the following procedure:

AddS3-AL (or AddS4-AL)	
Given.	A set of triplets $\mathcal{T}$ (or quadruplets $\mathcal{Q}$) on $[n]$
Step 1.	Compute the pairwise similarity function s^{AddS3} (or s^{AddS4}) for every pair of objects
Step 2.	Run average linkage algorithm with s^{AddS3} (or s^{AddS4})
Output.	The tree or dendrogram H on the n objects

Remark on approximation guarantee. Average linkage enjoys strong theoretical guarantees under the assumption that the similarities are always positive. In particular, Moseley and Wang (2017) show that average linkage achieves a worst-case $\frac{1}{3}$-approximation for revenue maximisation. Unfortunately, this result does not readily extend to AddS3-AL and AddS4-AL, as these similarities may be negative in some cases. A possible approach could be to add a positive constant to all the similarities to ensure that they are positive. Although this does not change the optimal tree or the one obtained from average linkage, a $\frac{1}{3}$-approximation for the modified revenues (considering revised similarities) does not imply a $\frac{1}{3}$-approximation for the original revenues.

Based on the proof of Moseley and Wang (2017), one can show that AddS3-AL (or AddS4-AL) returns a tree with non-negative triplet (or quadruplet) comparison revenue. Whether approximation guarantees may also be derived for AddS3-AL and AddS4-AL remains open.

7 Experiments

In this section, we propose two sets of experiments to demonstrate the practical relevance of our new revenue function and the corresponding algorithm. In our first set of experiments, our goal is to show the usefulness of revenue maximisation as a solution to find hierarchies that are closer to the ground truth. To this end, we consider a planted model and demonstrate the alignment between AARI scores, a supervised metric of goodness for clustering, and our proposed revenue function. In our second set of experiments, we aim to show that the heuristic proposed in Section 6 to maximize the revenue performs well in practice. Thus, on several real datasets, we compare our approach to two different state of the art approaches in comparison-based hierarchical clustering.

7.1 Planted Model

In this first set of experiments, we study the behaviour of the proposed revenues in a controlled setting. Hence, we generate data using a planted model for comparison-based hierarchical clustering (Ghoshdastidar et al., 2019) and we use 3 triplets-based and 2 quadruplets-based methods to learn dendrograms.

Data. To generate the data in this first set of experiments, we use a standard planted model in comparison-based hierarchical clustering (Balakrishnan et al., 2011; Ghoshdastidar et al., 2019). More precisely, given n objects, we create a real similarity matrix $S = (s_{ij})_{1 \leq i,j \leq n}$ such that $s_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2)$ is normally distributed and corresponds to the similarity between objects i and j. The choice of μ_{ij} defines a planted hierarchy, a complete binary tree of height L, built on top of 2^L ground clusters, that is sets of n_0 objects denoted as $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_{2^L}$. The total number of points (leaves) in the complete hierarchy is thus $n = n_0 2^L$. For every pair of objects i and j that belong to the same ground cluster, $\mu_{ij} = \mu$—a constant. On the other hand, for objects i, j from two distinct clusters, if $H(i \vee j)$ is rooted at level ℓ, we define $\mu_{ij} = \mu - (L - \ell)\delta$. We observe that the tree is rooted at level-0 and the constants—separation δ and noise level σ—control the hardness of the problem. In particular, smaller values of δ make the similarities between examples that belong to the same cluster more difficult to distinguish from similarities between examples that belong to different clusters. The signal-to-noise ratio is thus $\frac{\delta}{\sigma}$. In all the experiments, we set $\mu = 0.8$, $\sigma = 0.1$, $n_0 = 30$, $L = 3$ and we vary $\delta \in \{0.02, 0.04, ..., 0.2\}$. Since we are in a comparison-based setting, we do not directly use the similarities of the planted model to learn dendrograms but instead generate comparisons. Given $\mathcal{T}_{all}$ and $\mathcal{Q}_{all}$ the sets containing all possible triplets and quadruplets (see preliminaries), we obtain $\mathcal{T} \subseteq \mathcal{T}_{all}$ and $\mathcal{Q} \subseteq \mathcal{Q}_{all}$ by uniformly sampling kn^2 comparisons with $k > 0$.

Evaluation Function. To measure the closeness between the dendrograms obtained by the different approaches and the ground truth trees, we use the Averaged Adjusted Rand Index (Ghoshdastidar et al., 2019). The AARI is an extension to hierarchies of a well-known measure in standard clustering called Adjusted Rand Index (ARI; see Hubert and Arabie, 1985). The underlying idea is to average the ARI obtained over the top L levels of the tree. This measure takes values in $[0, 1]$ with higher values for more similar hierarchies, an AARI of 1 implying identical trees. Our goal is to empirically verify that the hierarchies with higher revenues are the ones closest to the ground truths as indicated by a higher AARI. Indeed, this would show that our revenue function is appropriate to evaluate the goodness of a dendrogram and that maximizing the revenue is indeed a good unsupervised way to select hierarchies. The results reported are averaged over 10 independent trials.[3] We defer the standard deviations to the appendix for the sake of readability.

Methods. We compare AddS3-AL and AddS4-AL, the two methods proposed in this work, to various comparison-based algorithms for learning dendrograms, such as 4K-AL (Ghoshdastidar et al., 2019), a quadruplets-based method, along with two triplets-based approaches MulK3-AL (Kleindessner and von Luxburg, 2017; Perrot et al., 2020) and tSTE-AL (Van Der Maaten and Weinberger, 2012). The former two are similarity-based approaches where the idea is use the comparisons to learn a similarity. The latter is an ordinal embedding approach where the idea is to recover a representation of the data that respects the comparisons as well as possible, and then use the cosine similarity $s_{ij} = \dfrac{\langle x_i, x_j \rangle}{||x_i||_2 ||x_j||_2}$ to compare the examples. To learn the dendrograms we then apply standard average linkage to the various similarities.

Results. In Figure 1, we present the AARI and Revenue of different triplet-based methods for several signal to noise ratios using n^2 comparisons.[4] We observe that, given a set signal to noise ratio, the ordering between the methods remains the same for the revenue and the AARI, that is the method with the highest revenue is also the one with the highest AARI. In other words, a higher revenue indicates that the corresponding dendrogram is better. In Table 1, we verify that this remains true for a constant signal to noise ratio of 1.5 and various number of observed comparisons. In particular, we notice that when the revenue of AddS3-AL

---

[3] The randomness stems from three sources: the noise in the similarities (we use the time + 42 as the seed), the triplets selection (we use the time as the seed), and the optimization procedure (initialization, batch selection) in tSTE (we use the time as the seed).

[4] Note that we also considered other amounts of comparisons. However, the trends were similar to the ones observed here and thus we chose to defer these results to the appendix.

Table 1: Revenue and AARI of various methods for a fixed signal to noise ratio $\frac{\delta}{\sigma} = 1.5$ and varying number of triplets (decreased by factor of 2). The planted setting consists of a total of $n = 240$ objects. In each line the highest revenue and AARI are underlined, taking into account standard deviation (see appendix). This shows that the two measures are well aligned.

Number of triplets	AddS3-AL		tSTE-AL		MulK3-AL	
	Revenue	AARI	Revenue	AARI	Revenue	AARI
$16n^2$	$\underline{7.336 \times 10^7}$	$\underline{0.937}$	7.286×10^7	0.849	7.304×10^7	0.863
$8n^2$	$\underline{3.667 \times 10^7}$	$\underline{0.901}$	3.648×10^7	0.867	3.631×10^6	0.854
$4n^2$	$\underline{1.820 \times 10^7}$	$\underline{0.855}$	$\underline{1.817 \times 10^7}$	$\underline{0.852}$	1.784×10^7	0.821
$2n^2$	8.927×10^6	0.778	$\underline{9.094 \times 10^6}$	$\underline{0.856}$	8.357×10^6	0.660
n^2	4.333×10^6	0.699	$\underline{4.559 \times 10^6}$	$\underline{0.856}$	3.754×10^6	0.539
$n^2/2$	2.097×10^6	0.614	$\underline{2.289 \times 10^6}$	$\underline{0.849}$	1.255×10^6	0.353
$n^2/4$	9.363×10^5	0.501	$\underline{1.138 \times 10^6}$	$\underline{0.847}$	1.680×10^5	0.086
$n^2/8$	4.310×10^5	0.404	$\underline{5.710 \times 10^5}$	$\underline{0.841}$	7.938×10^4	0.010
$n^2/16$	2.128×10^4	0.323	$\underline{2.853 \times 10^5}$	$\underline{0.667}$	5.156×10^3	0.003
$n^2/32$	1.048×10^4	0.151	$\underline{1.438 \times 10^4}$	$\underline{0.543}$	1.314×10^3	0.000

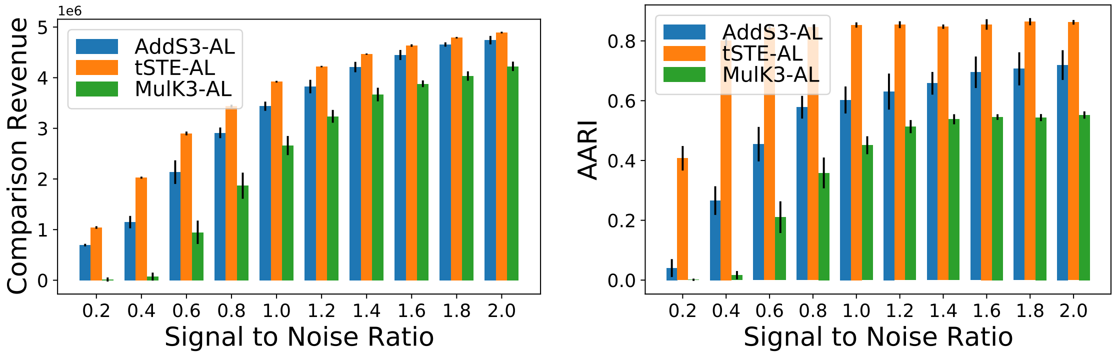


Figure 1: Revenue and AARI (higher is better) of several triplets-based methods using n^2 comparisons. Given various signal to noise ratios, a higher revenue implies higher AARI values (better dendrograms).

becomes higher than the revenue of tSTE-AL, that is using more than $4n^2$ triplets, the AARI also follows the same trend, thus confirming that selecting the dendrogram with the highest revenue is indeed a good way to select meaningful hierarchies. In the appendix, we show that the same behaviour can be observed for various signal to noise ratios as well as in the quadruplet case.

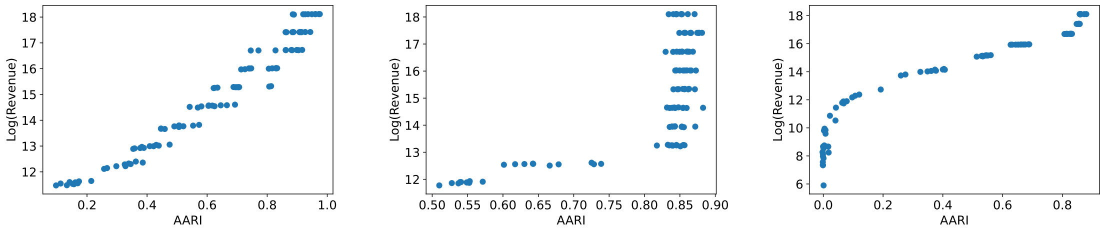
We further investigate the dependence between AARI and the triplet revenue in Figure 2, where we plot the AARI and the corresponding triplet revenue (in log scale) for different runs and number of triplets, considered in Table 1. Although the variation of the revenue for increasing AARI seems to depend on the method under consideration, all plots show a monotonic trend. To validate this we use two measures of rank correlation—Kendall's-τ and Spearman's-ρ—that capture how well the dependence is captured as monotonic function. For both AddS3-AL and MulK3-AL, the rank correlations are greater than 0.9, indicating a highly monotonic trend. Although the rank correlations are smaller for tSTE-AL, but still large enough and corresponding p-value is about 10^{-14}, indicating rank correlation.

7.2 Real Data

The previous experiments establish that our revenue functions are good at identifying meaningful dendrograms in an unsupervised way. In the following experiments, we investigate the behaviour of the proposed

Table 2: Experiments on real datasets. For the triplets-based methods, AddS3-AL tends to obtain the dendrograms with the best revenues. For the quadruplets-based approaches, AddS4-AL and 4K-AL obtain comparable results. Using the original Cosine similarities only yields slightly better hierarchies than the comparison-based methods. For first 3 datasets, multiple runs are used and standard deviation (see appendix) is considered for highlighting the best method(s).

Dataset	Triplet				Quadruplet		
	AddS3-AL	tSTE-AL	MulK3-AL	Cosine-AL	AddS4-AL	4K-AL	Cosine-AL
Zoo	$2.759{\times}10^5$	$2.180{\times}10^5$	$2.029{\times}10^5$	$2.815{\times}10^5$	$2.859{\times}10^5$	$2.865{\times}10^5$	$2.962{\times}10^5$
Glass	$2.182{\times}10^6$	$1.993{\times}10^6$	$1.479{\times}10^6$	$2.115{\times}10^6$	$2.427{\times}10^6$	$2.447{\times}10^6$	$2.493{\times}10^6$
MNIST	$1.876{\times}10^9$	$2.028{\times}10^9$	$1.749{\times}10^9$	$2.027{\times}10^9$	$1.935{\times}10^9$	$1.910{\times}10^9$	$2.062{\times}10^9$
Car	$1.521{\times}10^5$	$1.562{\times}10^5$	$1.264{\times}10^5$	-	$1.521{\times}10^5$	$1.125{\times}10^5$	-
Food	$6.137{\times}10^6$	$5.993{\times}10^6$	$6.096{\times}10^6$	-	$6.137{\times}10^6$	$6.137{\times}10^6$	-
Vogue	$2.722{\times}10^4$	$2.104{\times}10^4$	$3.022{\times}10^3$	-	$2.722{\times}10^4$	$2.549{\times}10^4$	-
Nature	$2.650{\times}10^5$	$2.056{\times}10^5$	$1.231{\times}10^5$	-	$2.650{\times}10^5$	$2.228{\times}10^5$	-
Imagenet	$7.179{\times}10^7$	$6.571{\times}10^7$	$3.440{\times}10^7$	-	$7.179{\times}10^7$	$6.994{\times}10^7$	-



(a) **AddS3-AL** ($\tau = 0.934,\ \rho = 0.992$)

(b) **tSTE-AL** ($\tau = 0.515,\ \rho = 0.673$)

(c) **MulK3-AL** ($\tau = 0.907,\ \rho = 0.982$)

Figure 2: Scatter plots for the AARI and triplet revenue from different runs and varying number of triplets in the setting of Table 4. The three methods (AddS3, tSTE and MulK3) show different trends. However, in each case, the triplet revenue and AARI have high Kendall-τ and Spearman-ρ correlation.

approaches on real data. In particular, we show that they are competitive with standard comparison-based hierarchical clustering approaches on various datasets.

Data. We consider 8 different datasets. On the one hand, we consider 3 standard clustering datasets: Zoo, Glass, and MNIST (Heller and Ghahramani, 2005; LeCun et al., 2010; Vikram and Dasgupta, 2016). The Zoo dataset originally consisted of 101 animals each with 16 features. But we choose to remove the entry with class "girl" since we do not feel it belongs to the zoo dataset. The Glass dataset has 9 features for 214 examples. For the MNIST dataset we consider two subsets of the MNIST test dataset that originally contains 10000 examples distributed among the ten digits. A 2-dimensional embedding of the entire MNIST test data was constructed with t-SNE (van der Maaten, 2014). From this we randomly sampled 200 examples for each digit to form a dataset of 2000 entries and normalized the embeddings so that each example

Table 3: Description of datasets used in the experiments.

Dataset	Query	#Objects	#Triplets
Zoo	Cosine Similarity	100	100000
Glass	Cosine Similarity	214	45796
MNIST	Cosine Similarity	2000	4000000
Car	Most Central Triplet	60	14194
Food	Standard Triplet	100	190376
Vogue	Odd-out Triplet	60	2214
Nature	Odd-out Triplet	120	6710
Imagenet	Rank 2 from 8	1000	328549

lies in $[-1, 1]$. Since we are in a comparison-based setting, we generate n^2 comparisons using the cosine similarity. To model mistakes from human annotators, we randomly and uniformly flip 5% of the comparisons (Emamjomeh-Zadeh and Kempe, 2018), where by flipping (i, j, k) we mean replacing it with (i, k, j). On the other hand, we consider 5 comparison-based datasets, Car, Food, Vogue Cover, Nature Scene and ImageNet Images v0.1, from the cblearn repository.[5] The number of objects and the kind of query used to obtain comparisons are summarized in Table 3. The comparisons are transformed into triplets (final number of triplets noted in Table 3), which are also used in the quadruplet setting. In Table 3, a central triplet is a query of the form—which of the three objects (i,j,k) is most central. Provided that the answer is i, this implies object i is more similar to both j and k than they are to each other. Two standard triplets (j,i,k) and (k,i,j) are thus obtained. An Odd-out triplet is a query of the form—which of the three objects (i,j,k) is the odd one out. If i is picked as the odd one, it gives two standard triplets of the form (j,k,i) and (k,j,i). A rank 2 from 8 query is of the form—among 8 objects $(i_0, \dots, i_7)$, rank the 2 that appear to be most similar to the reference object i_0. If i_1 and i_2 are ranked as the most similar to i_0 in this order, then 11 standard triplets of the form $(i_0, i_1, i_k)_{k=2}^{7}$ and $(i_0, i_2, i_k)_{k=3}^{7}$ are obtained.

Evaluation Function. Since the datasets considered here do not come with a ground truth hierarchy, we cannot compute the AARI. Hence, we only report the revenue. The results reported are averaged over 10 independent trials[6] and defer the standard deviations to the appendix.

Methods. Besides the methods already used in the planted setting, we also consider the Cosine baseline where it is assumed that the pairwise cosine similarities are available, and we apply average linkage directly on the similarities used to generate the comparisons. This baseline is not applicable to the comparison-based datasets where we only have access to the comparisons and not to the similarities.

Results. The results are reported in Table 2. We can notice that AddS3-AL tends to be better than tSTE-AL and MulK3-AL while AddS4-AL and 4K-AL are comparable. As is expected, the Cosine baseline based on the original similarities obtains the best performances in most cases, but it only seems to yield slightly better hierarchies than the comparison-based methods. This would tend to confirm that hierarchical clustering with average linkage is indeed a problem that can be solved using only a limited number of comparisons, instead of using all similarities.

8 Conclusion

In this paper, we proposed novel revenue functions that allow us to measure the goodness of a dendrogram in an unsupervised way using only triplet or quadruplet comparisons. This suggest natural algorithms for hierarchical clustering based on the maximization of such revenues. Drawing theoretical connections with existing work on cost and revenue functions in standard hierarchical clustering, we propose two algorithms based on average linkage for hierarchical clustering using only comparisons. We empirically show that our revenue functions successfully identify the dendrograms that are closest to the ground truth. We also show that the proposed approaches to learn hierarchies perform well on real datasets and are competitive with state of the art methods.

We further used the proposed revenue function to resolve an open theoretical problem of recovering a latent hierarchy using fewer than $\Omega(n^3)$ passive triplets. We showed that $O(n^2 \log n/\epsilon^2)$ passive triplets suffice to obtain a $(1-\epsilon)$-approximation of the optimal triplet revenue. We conclude with the following open questions:
(i) Are $\Omega(n^2 \log n)$ passive triplets necessary for a $(1-\epsilon)$-approximation?
(ii) At this point, we are unable to obtain a polynomial-time approximation scheme (PTAS) for the revenue maximisation problem, but we believe this should be possible. It may also be possible to obtain more efficient algorithms that are linear in time with respect to the number of triplets. A linear-time method exists for maximum quartet consistency (Snir and Yuster, 2011).

---

[5] `https://github.com/dekuenstle/cblearn`

[6] The randomness stems from two main sources: the triplets generation in the Zoo, Glass, and MNIST dataset (we use the time as the seed for the triplets selection and the time + 42 as the seed for the random flips) and the optimization procedure (initialization, batch selection) in tSTE (we use the time as the seed). For all the other datasets and methods, every step is deterministic and, thus, we only need to report the results of a single run.

References

E. N. Adams III. Consensus techniques and the comparison of taxonomic trees. *Systematic Biology*, 21(4): 390–397, 1972.

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

A. V. Aho, Y. Sagiv, T. G. Szymanski, and J. D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. Comput.*, 10(3):405–421, 1981. doi: 10.1137/0210030.

S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh. Noise thresholds for spectral clustering. *Advances in Neural Information Processing Systems*, 24:954–962, 2011.

K. S. Berenhaut, K. E. Moore, and R. L. Melvin. A social perspective on perceived distances reveals deep community structure. *Proceedings of the National Academy of Sciences*, 119(4):e2003634119, 2022.

J. Byrka, S. Guillemot, and J. Jansson. New results on optimizing rooted triplets consistency. *Discret. Appl. Math.*, 158(11):1136–1147, 2010. doi: 10.1016/j.dam.2010.03.004.

D. Catanzaro. The minimum evolution problem: Overview and classification. *Networks*, 53(2):112–125, 2009. doi: 10.1002/net.20280.

M. Charikar, V. Chatziafratis, and R. Niazadeh. Hierarchical clustering better than average-linkage. In T. M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 2291–2304. SIAM, 2019. doi: 10.1137/1.9781611975482.139.

V. Chatziafratis, M. Mahdian, and S. Ahmadian. Maximizing agreements for ranking, clustering and hierarchical clustering via MAX-CUT. In A. Banerjee and K. Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1657–1665. PMLR, 2021.

V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn, and C. Mathieu. Hierarchical clustering: Objective functions and algorithms. *Journal of the ACM (JACM)*, 66(4):1–42, 2019.

S. Dasgupta. A cost function for similarity-based hierarchical clustering. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 118–127, New York, NY, USA, 2016. Association for Computing Machinery. doi: 10.1145/2897518.2897527. URL `https://doi.org/10.1145/2897518.2897527`.

D. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence. The quest for ground truth in musical artist similarity. In *Proceedings of ISMIR*, 2002.

E. Emamjomeh-Zadeh and D. Kempe. Adaptive hierarchical clustering using ordinal queries. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 415–429. SIAM, 2018.

L. Foulds, M. Hendy, and D. Penny. A graph theoretic approach to the development of minimal phylogenetic trees. *J. Mol. Evol*, 13:149, 1979.

D. Ghoshdastidar, M. Perrot, and U. von Luxburg. Foundations of comparison-based hierarchical clustering. *Advances in Neural Information Processing Systems*, 32:7456–7466, 2019.

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

S. Haghiri, F. A. Wichmann, and U. von Luxburg. Estimation of perceptual scales using ordinal embedding. *Journal of vision*, 20(9):14–14, 2020.

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

K. A. Heller and Z. Ghahramani. Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning*, pages 297–304, 2005.

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

M. Janowitz. Monotone equivariant cluster methods. *SIAM Journal on Applied Mathematics*, 37(1):148–165, 1979.

M. F. Janowitz. *Mathematical Taxonomy.* Wiley Series in Probability and Mathematical Statistics. Wiley, 1971. ISBN 978-0471440505.

T. Jiang, P. E. Kearney, and M. Li. A polynomial time approximation scheme for inferring evolutionary trees from quartet topologies and its application. *SIAM J. Comput.*, 30(6):1942–1961, 2000. doi: 10.1137/S0097539799361683.

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

M. Kleindessner and U. von Luxburg. Kernel functions based on triplet similarity comparisons. In *Advances in Neural Information Processing Systems*, pages 6807–6817, 2017.

Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database, 2010.

B. Moseley and J. Wang. Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

M. Perrot and U. von Luxburg. Boosting for comparison-based learning. In *IJCAI*, 2019.

M. Perrot, P. Esser, and D. Ghoshdastidar. Near-optimal comparison based clustering. In *Advances in Neural Information Processing Systems*, volume 33, pages 19388–19399. Curran Associates, Inc., 2020.

C. Semple and M. A. Steel. *Phylogenetics*, volume 24. Oxford University Press on Demand, 2003.

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

R. Sibson. Order invariant methods for data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(3):311–338, 1972.

S. Snir and S. Rao. Quartets maxcut: A divide and conquer quartets algorithm. *IEEE ACM Trans. Comput. Biol. Bioinform.*, 7(4):704–718, 2010. doi: 10.1109/TCBB.2008.133.

S. Snir and R. Yuster. A linear time approximation scheme for maximum quartet consistency on sparse sampled inputs. *SIAM J. Discret. Math.*, 25(4):1722–1736, 2011. doi: 10.1137/110820555. URL `https://doi.org/10.1137/110820555`.

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

A. Ukkonen. Crowdsourced correlation clustering with relative distance comparisons. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1117–1122. IEEE, 2017.

L. van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(93):3221–3245, 2014. URL `http://jmlr.org/papers/v15/vandermaaten14a.html`.

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

L. C. Vankadara, S. Haghiri, F. U. Wahab, and U. von Luxburg. Large scale ordinal embedding: training neural networks with structure-free inputs. *arXiv preprint*, arXiv:1912.01666, 2019.

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

D. Wang and Y. Wang. An improved cost function for hierarchical cluster trees. *Journal of Computational Geometry*, 11(1):283–331, 2020. doi: 10.20382/jocg.v11i1a11.

M. Wilber, S. Kwak, and S. Belongie. Cost-effective hits for relative similarity comparisons. In *Human Computation and Crowdsourcing (HCOMP)*, Pittsburgh, 2014.

B. Y. Wu. Constructing the maximum consensus tree from rooted triples. *J. Comb. Optim.*, 8(1):29–39, 2004. doi: 10.1023/B:JOCO.0000021936.04215.68.

A Proof of Theorem 1

Recall the formulation of $Trev$ for a tree H and a set of triplets $\mathcal{T}$:

$$\begin{aligned}
&Trev(H, \mathcal{T}) \\
&\quad = \sum_{i,j,k} \mathbb{I}_{(i,j,k)\in\mathcal{T}} \Big(|H(i \vee k)| - |H(i \vee j)| \Big) \\
&\qquad\qquad \text{(sum over all distinct } i, j, k.\text{)} \\
&\quad = \sum_{i,j,k} \mathbb{I}_{(i,j,k)\in\mathcal{T}} |H(i \vee k)| - \sum_{i,j,k} \mathbb{I}_{(i,j,k)\in\mathcal{T}} |H(i \vee j)| \\
&\quad = \sum_{i,j,k} \mathbb{I}_{(i,k,j)\in\mathcal{T}} |H(i \vee j)| - \sum_{i,j,k} \mathbb{I}_{(i,j,k)\in\mathcal{T}} |H(i \vee j)| \\
&\qquad\qquad \text{(by change of variable between } j \text{ and } k \text{ in the first term.)} \\
&\quad = \sum_{i \neq j} \Big(\sum_{k \neq i,j} \mathbb{I}_{(i,k,j)\in\mathcal{T}} - \mathbb{I}_{(i,j,k)\in\mathcal{T}} \Big) |H(i \vee j)| \\
&\quad = \sum_{i<j} |H(i \vee j)| \times \\
&\qquad \Big(\sum_{k \neq i,j} \mathbb{I}_{(i,k,j)\in\mathcal{T}} - \mathbb{I}_{(i,j,k)\in\mathcal{T}} + \mathbb{I}_{(j,k,i)\in\mathcal{T}} - \mathbb{I}_{(j,i,k)\in\mathcal{T}} \Big) \\
&\qquad\qquad \text{(by change of variable between } i \text{ and } j \text{ when } i > j.\text{)} \\
&\quad = \sum_{i<j} -s_{ij}^{AddS3} |H(i \vee j)|
\end{aligned}$$

by definition of s^{AddS3}. This concludes the proof for the triplets-based revenue.

Using a similar approach, recall the formulation of $Qrev$ for a tree H and quadruplet set $\mathcal{Q}$:

$$
\begin{aligned}
&Qrev(H, \mathcal{Q}) \\
&\quad = \sum_{i,j,k,l} \mathbb{I}_{(i,j,k,l)\in\mathcal{Q}} \Big(|H(k \vee l)| - |H(i \vee j)| \Big) \\
&\quad = \sum_{i,j,k,l} \mathbb{I}_{(i,j,k,l)\in\mathcal{Q}} |H(k \vee l)| - \sum_{i,j,k,l} \mathbb{I}_{(i,j,k,l)\in\mathcal{Q}} |H(i \vee j)| \qquad \text{(sum over all } i<j, k<l, (i,j) \neq (k,l).) \\
&\quad = \sum_{i,j,k,l} \mathbb{I}_{(k,l,i,j)\in\mathcal{Q}} |H(i \vee j)| - \sum_{i,j,k,l} \mathbb{I}_{(i,j,k,l)\in\mathcal{Q}} |H(i \vee j)| \\
&\qquad\qquad \text{(by swapping the role of } i,j \text{ and } k,l \text{ in the first term.)} \\
&\quad = \sum_{i<j} \Big(\sum_{\substack{k<l \\ (k,l)\neq(i,j)}} \mathbb{I}_{(k,l,i,j)\in\mathcal{Q}} - \mathbb{I}_{(i,j,k,l)\in\mathcal{Q}} \Big) |H(i \vee j)| \\
&\quad = \sum_{i<j} -s_{ij}^{AddS4} |H(i \vee j)|
\end{aligned}
$$

by definition of s^{AddS4}. This concludes the proof for the quadruplets-based revenue.

B Proof of Proposition 2

The proposition follows immediately from the following two lemmas.

Lemma 5. *Let H_0 be a binary tree on $[n]$ and $\mathcal{T}_0$ be the set of triplets induced by H_0. Then*

$$\mathcal{T}_0 = \arg\max_{\mathcal{T}} Trev(H_0, \mathcal{T}), \tag{6}$$

where the maximum is unique over all triplet sets that are induced by some binary tree on $[n]$.

Lemma 6. *Let H_0, H_1 be two binary trees on $[n]$ and $\mathcal{T}_0, \mathcal{T}_1$ be the set of triplets induced by H_0 and H_1, respectively. Then $Trev(H_0, \mathcal{T}_1) = Trev(H_1, \mathcal{T}_0)$.*

Combining the above two lemmas, we obtain that

$$Trev(H_0, \mathcal{T}_0) > Trev(H_0, \mathcal{T}_1) = Trev(H_1, \mathcal{T}_0)$$

for any tree H_1 and corresponding set of triplets $\mathcal{T}_1$. This directly implies the Proposition 2. We now complete the proof by proving Lemmas 5 and 6.

Proof of Lemma 5. Recall that

$$Trev(H_0, \mathcal{T}_0) = \sum_{(i,j,k)\in\mathcal{T}_0} \underbrace{\Big(|H_0(i \vee k)| - |H_0(i \vee j)| \Big)}_{=:D_0(i,j,k)}$$

is a sum of positive terms. For convenience, we denote each difference by $D_0(i,j,k) > 0$.

Let $\mathcal{T}_1$ be the set of triplets generated by another tree H_1. Note that at least one pair of triplets in $\mathcal{T}_1$ has to be different from $\mathcal{T}_0$, otherwise H_1 and H_0 would be isomorphic transformations of one another. Without loss of generality, assume that the pair $(i,j,k), (j,i,k)$ has been replaced by the pair $(i,k,j), (k,i,j)$, that is, i,k are merged in H_1 before they are merged to j. Observe that we can write

$$
\begin{aligned}
&Trev(H_0, \mathcal{T}_0) - Trev(H_0, \mathcal{T}_1) \\
&= \sum_{\substack{(i,j,k), \\ (i,k,j) \\ \in\mathcal{T}_0\setminus\mathcal{T}_1}} D_0(i,j,k) + D_0(j,i,k) - D_0(i,k,j) - D_0(k,i,j)
\end{aligned}
$$

where each term in the summation can be computed as

$$\begin{aligned}&D_0(i,j,k)+D_0(j,i,k)-D_0(i,k,j)-D_0(k,i,j)\\&=|H_0(i\vee k)|+|H_0(j\vee k)|-2|H_0(i\vee j)|\\&\qquad-\Big(|H_0(i\vee j)|+|H_0(k\vee j)|-2|H_0(i\vee k)|\Big)\\&=3D_0(i,j,k),\end{aligned}$$

which is strictly positive for every $(i,j,k)\in\mathcal{T}_0$. Summing over all $(i,j,k),(j,i,k)\in\mathcal{T}_0\backslash\mathcal{T}_1$, we have that $Trev(H_0,\mathcal{T}_0)>Trev(H_0,\mathcal{T}_1)$ for any $\mathcal{T}_1$ generated by another tree H_1. ☐

Proof of Lemma 6. Let $\{s_{0ij}\}_{i,j}$ be the pairwise AddS3 similarity induced by $\mathcal{T}_0$, and $\{s_{1ij}\}_{i,j}$ be the AddS3 similarity from $\mathcal{T}_1$. Due to the definition of $\mathcal{T}_0$, we note that, for any $k\neq i,j$, the term $(\mathbb{I}_{(i,j,k)\in\mathcal{T}_0}-\mathbb{I}_{(i,k,j)\in\mathcal{T}_0}+\mathbb{I}_{(j,i,k)\in\mathcal{T}_0}-\mathbb{I}_{(j,k,i)\in\mathcal{T}_0})$ either takes the value 2 if $k\notin(i\vee j)$— that is, i,j is merged in H_0 before k—or the value -1 if k is merged to either i or j before $(i\vee j)$. Summing over all $k\neq i,j$ gives

$$\begin{aligned}s_{0ij}&=2(n-|H_0(i\vee j)|)-(|H_0(i\vee j)|-2)\\&=2n+2-3|H_0(i\vee j)|\end{aligned}$$

for every i,j. Using the same arguments s_{1ij} can be expressed as $s_{1ij}=2n+2-3|H_1(i\vee j)|$. We can now use the equivalence in Theorem 1 to write

$$\begin{aligned}&Trev(H_0,\mathcal{T}_1)-Trev(H_1,\mathcal{T}_0)\\&=-\sum_{i<j}(s_{1ij}|H_0(i\vee j)|-s_{0ij}|H_1(i\vee j)|)\\&=-\sum_{i<j}\Big((2n+2-3|H_1(i\vee j)|)|H_0(i\vee j)|\\&\qquad-(2n+2-3|H_0(i\vee j)|)|H_1(i\vee j)|\Big)\\&=-(2n+2)\sum_{i<j}\Big(|H_0(i\vee j)|-|H_1(i\vee j)|\Big)\\&=0\end{aligned}$$

since $\sum\limits_{i<j}|H_0(i\vee j)|=\sum\limits_{i<j}|H_1(i\vee j)|=\frac{1}{3}(n^3-n)$ is Dasgupta's cost for any tree on $[n]$ when all pairwise similarities are 1 (Dasgupta, 2016, Theorem 3). Hence, the claim. ☐

C Proof of Theorem 3 and Corollary 4

We first state and prove two lemmas that are essential for the proof of Theorem 3. The first lemma shows that $Trev(H_0,\mathcal{T}_0)=\Omega(n^4)$. The second lemma derives concentration inequalities for the AddS3 similarities s_{ij}, which is then used in the proof of Theorem 3 to derive bound on $|Trev(H,\mathcal{T})-Trev(H,\mathcal{T}_0)|$ for all H, and subsequently arrive at the claim.

Lemma 7. *For every $\epsilon\in(0,1)$, there exists $n_0>0$ such that the following holds for all $n>n_0$. If H_0 is a hierarchy on $[n]$ and $\mathcal{T}_0$ is the set of triplets induced by H_0,*

$$Trev(H_0,\mathcal{T}_0)\geq\frac{(1-\epsilon)n^4}{12}.$$

Proof. We start with the equivalence in Theorem 1 to write the revenue as $Trev(H_0,\mathcal{T}_0)=-\sum_{i<j}s_{0ij}|H_0(i\vee j)|$, where s_{0ij} is the pairwise AddS3 similarity induced by $\mathcal{T}_0$. Due to the definition of $\mathcal{T}_0$, we note that, for any $k\neq i,j$, the term $(\mathbb{I}_{(i,j,k)\in\mathcal{T}_0}-\mathbb{I}_{(i,k,j)\in\mathcal{T}_0}+\mathbb{I}_{(j,i,k)\in\mathcal{T}_0}-\mathbb{I}_{(j,k,i)\in\mathcal{T}_0})$ either takes the value 2 if $k\notin(i\vee j)$—

that is, i, j is merged in H_0 before k—or the value -1 if k is merged to either i or j before $(i \vee j)$. Summing over all $k \neq i, j$ gives

$$\begin{aligned} s_{0ij} &= 2(n - |H_0(i \vee j)|) - (|H_0(i \vee j)| - 2) \\ &= 2n + 2 - 3|H_0(i \vee j)| \end{aligned}$$

for every i, j. Let $N = (i \vee j)$ denote the least common ancestor of i, j in H_0, and N_1, N_2 be the two children of N. Note that $|N_1| \cdot |N_2|$ pairs of i, j are merged at N. Hence, we can rewrite the revenue as

$$\begin{aligned} &Trev(H_0, \mathcal{T}_0) \\ &= -\sum_{i<j} s_{0ij} |H_0(i \vee j)| \\ &= \sum_{i<j} |H_0(i \vee j)| \big(3|H_0(i \vee j)| - 2n - 2\big) \\ &= \sum_{N \in H_0} |N_1| \cdot |N_2| \cdot |N| \cdot \big(3|N| - 2n - 2\big) \\ &= 3 \sum_{N \in H_0} |N_1||N_2||N|^2 - (2n+2) \sum_{N \in H_0} |N_1||N_2||N| \end{aligned}$$

where the summations are over all internal nodes N in the tree H_0, with N_1, N_2 deenoting the two children of N. The second summation is Dasgupta's cost for any tree on $[n]$ with all pairwise similarities as 1, and evaluates to $\frac{n^3-n}{3}$ (Dasgupta, 2016, Theorem 3). On the other hand, we claim that the first sum has lower bound $\sum\limits_{N \in H_0} |N_1||N_2||N|^2 \geq \frac{n^4}{4}$.

We prove this claim through induction on n. The claim is easy to verify for $n = 2, 3$. For $n \geq 4$, we assume that claim holds for any H_0 with k leaves, when $k < n$ (equivalently, H_0 on $[k]$). Consider the tree H_0 on $[n]$ such that the root node is split into two nodes of size $n_1, n_2 < n$ (note $n_1 + n_2 = n$). From our inductive hypothesis,

$$\begin{aligned} &\sum_{N \in H_0} |N_1||N_2||N|^2 \\ &\qquad \geq n_1 n_2 n^2 + \frac{n_1^4}{4} + \frac{n_2^4}{4} \\ &\qquad = \frac{1}{4}\left(4n_1^3 n_2 + 8n_1^2 n_2^2 + 4n_1 n_2^3 + n_1^4 + n_2^4\right) \\ &\qquad \geq \frac{1}{4}(n_1 + n_2)^4 = \frac{n^4}{4} \end{aligned}$$

which proves the claim for any n. Combining all terms, we have

$$\begin{aligned} Trev(H_0, \mathcal{T}_0) &\geq \frac{3n^4}{4} - \frac{(2n+2)(n^3-n)}{3} \\ &= \frac{n^4}{12} - \frac{2}{3}(n^3 - n^2 - n). \end{aligned}$$

Given $\epsilon \in (0, 1)$, we can choose n_0 such that for every $n > n_0$, the negative term is smaller that $\frac{\epsilon n^4}{12}$, which proves the statement of the lemma. ☐

We now state and prove the concentration results for the AddS3 similarity computed from the sampled triplet set $\mathcal{T}$. We first recall the sampling and introduce some notations. For any n, with probability $p_n \in (0, 1)$, a pair of triplets $(i, j, k), (j, i, k) \in \mathcal{T}_0$ is included in $\mathcal{T}$, independent of other pairs. To formalise this, we define the random variable $\chi_{ijk} = \chi_{jik} \sim \text{Bernoulli}(p_n)$ such that the collection $\{\chi_{ijk} \ : \ i < j, k \neq i, j\}$ are mutually independent. If s_{ij} denotes the pairwise AddS3 similarity, computed using $\mathcal{T}$, then observe that

$$s_{ij} = \sum_{k \neq i,j} \chi_{ijk} (\mathbb{I}_{(i,j,k) \in \mathcal{T}_0} - \mathbb{I}_{(i,k,j) \in \mathcal{T}_0} + \mathbb{I}_{(j,i,k) \in \mathcal{T}_0} - \mathbb{I}_{(j,k,i) \in \mathcal{T}_0}) \tag{7}$$

Hence, for a fixed H_0—and $\mathcal{T}_0$—the similaritiy s_{ij} is a weighted sum of independent Bernoullis, with weights either 2 or -1 (cf. proof of Lemma 7). As a consequence, $\mathbb{E}[s_{ij}] = p_n s_{0ij}$, where the expectation is with respect to sampling, and furthermore we can state the following concentration for all pairwise similarities.

Lemma 8. *Assume $p_n > c\log n/n$ for some constant $c > 0$. Let $\mathcal{T}$ denote a random subset of $\mathcal{T}_0$ (obtained from the aforementioned sampling), and $\{s_{ij}\}_{i<j}, \{s_{0ij}\}_{i<j}$ denote the pairwise AddS3 similarities computed using triplets in $\mathcal{T}$ and $\mathcal{T}_0$, respectively. There is an absolute constant $c' > 0$ such that with probability $1 - n^{-c'}$,*

$$|\mathcal{T}| = \Theta(p_n n^3) \quad \textit{and} \quad \max_{i<j} |s_{ij} - p_n s_{0ij}| = O\left(\sqrt{p_n n \log n}\right).$$

Proof. We first derive the bound on $\max\limits_{i<j} |s_{ij} - p_n s_{0ij}|$. From the expression of s_{ij}, mentioned above, we note that $s_{ij} - p_n s_{0ij}$ is a sum of $(n-2)$ independent mean zero random variables, with each term in $[-2, 2]$ and variance bounded by $4p_n$. By Bernstein's inequality,

$$\mathbb{P}\big(|s_{ij} - p_n s_{0ij}| > \delta\big) \leq 2\exp\left(-\frac{\delta^2}{8p_n(n-2) + \frac{4}{3}\delta}\right).$$

Assuming $p_n > c\log n/n$ and setting $\delta = c''\sqrt{p_n n \log n}$ for $c, c'' > 0$, we have $|s_{ij} - p_n s_{0ij}| > c''\sqrt{p_n n \log n}$ with probability $\leq n^{-c'''}$ for a constant that depends on c''. Using union bound over all $\binom{n}{2}$ i, j pairs, we have $\max\limits_{i<j} |s_{ij} - p_n s_{0ij}| > c''\sqrt{p_n n \log n}$ with probability at most $n^{-(c'''-2)}$, where $c''' > 2$ assuming c, c' is chosen large enough. This proves the second claim of the lemma.

The claim $|\mathcal{T}| = \Theta(p_n n^3)$ is proven as follows. From definition of $\mathcal{T}_0$, every internal node $N \in H_0$ contributes $|N_1||N_2|(|N| - 2)$ triplets to $\mathcal{T}_0$ since merger of every i, j contributes to $|N| - 2$ triplets, one for each k that is either merged with i or j at a lower level. Hence, $|\mathcal{T}_0| = \sum\limits_{N \in H_0} |N_1||N_2|(|N| - 2)$, which can be bounded from below by $n^3/9$, using induction on n. Finally, note that $\mathbb{E}[|\mathcal{T}|] = p_n|\mathcal{T}_0|$ and a Bernstein-type concentration inequality shows that $|\mathcal{T}| = (1 - o(1))p_n|\mathcal{T}_0|$ with probability $1 - n^{-c''''}$ for some constant c'''' for $p_n > c\log n/n$. $\square$

Below, we prove Theorem 3 using Lemmas 7–8.

Proof of Theorem 3. We first derive bounds on the deviation of the revenue $Trev$ of any tree H due to sampling. The concentration of $\{s_{ij}\}_{i<j}$ ensures that we can state a deviation bound that uniformly holds for all H, as shown below. From the equivalence in Theorem 1, we write for any H,

$$\begin{aligned}
|Trev(H, \mathcal{T}) - p_n Trev(H, \mathcal{T}_0)| &= \left|\sum_{i<j} (s_{ij} - p_n s_{0ij})|H(i \vee j)|\right| \\
&\leq \sum_{i<j} |s_{ij} - p_n s_{0ij}| \cdot |H(i \vee j)| \\
&\leq c'\sqrt{p_n n \log n} \sum_{i<j} |H(i \vee j)|,
\end{aligned}$$

where the last bound holds with probability $1 - n^{-c'}$ due to Lemma 8. Note that $\sum\limits_{i<j} |H(i \vee j)|$ is Dasgupta's cost of tree H on $[n]$ if all pairwise similarities are 1, and hence the summation is $\frac{n^3 - n}{3}$ (Dasgupta, 2016). We conclude that, with probability $1 - n^{-c'}$,

$$\max_H |Trev(H, \mathcal{T}) - p_n Trev(H, \mathcal{T}_0)| = O\left(\sqrt{p_n n^7 \log n}\right).$$

We now write

$$\begin{aligned} Trev(\widehat{H}, \mathcal{T}_0) &\geq \frac{1}{p_n} Trev(\widehat{H}, \mathcal{T}) - O\left(\sqrt{\frac{n^7 \log n}{p_n}}\right) \\ &\geq \frac{1}{p_n} Trev(H_0, \mathcal{T}) - O\left(\sqrt{\frac{n^7 \log n}{p_n}}\right) \\ &\geq Trev(H_0, \mathcal{T}_0) - O\left(\sqrt{\frac{n^7 \log n}{p_n}}\right), \end{aligned}$$

where the first and third inequalities follow from the deviation bound stated above, and the second inequality holds since $\widehat{H}$ maximises $Trev(H, \mathcal{T})$. For $p_n > c \log n / n\epsilon^2$, the second term is $O(\epsilon n^4/\sqrt{c})$. Choosing c large enough, the second term can be made smaller than $\epsilon(1-\epsilon)n^4/12 \leq \epsilon\, Trev(H_0, \mathcal{T}_0)$, where the last bound is due to Lemma 7. Hence the claim. ☐

Proof of Corollary 4. In the noisy setting, the random flipping can be modelled through the independent variables $\{\zeta^i_{jk} \;:\; i, j < k\}$ where $\zeta^i_{jk} \sim \text{Bernoulli}(\alpha)$ is the indicator for triplet (i,j,k) to be flipped with triplet (i,k,j). Using the notation of equation 7, the variable $\xi_{ijk}\zeta^i_{jk}$ indicates $(i,k,j) \in \mathcal{T}'$ (noisy triplet) whereas $\xi_{ijk}(1-\zeta^i_{jk})$ indicates correct triplet $(i,j,k) \in \mathcal{T}'$. Hence,

$$s_{ij} = \sum_{k \neq i,j} \chi_{ijk}(1-2\zeta^i_{jk})(\mathbb{I}_{(i,j,k)\in\mathcal{T}_0} - \mathbb{I}_{(i,k,j)\in\mathcal{T}_0}) + \chi_{ijk}(1-2\zeta^j_{ik})(\mathbb{I}_{(j,i,k)\in\mathcal{T}_0} - \mathbb{I}_{(j,k,i)\in\mathcal{T}_0}), \tag{8}$$

and $\mathbb{E}[s_{ij}] = (1-2\alpha)p_n s_{0ij}$. Following the arguments of Lemma 8, we can similarly state that, with probability $1 - n^{-c'}$,

$$\max_{i<j} |s_{ij} - (1-\alpha)p_n s_{0ij}| = O\left(\sqrt{p_n n \log n}\right)$$

where the deviation bound is same as in Lemma 8 since the same variance bound holds for the independent random terms in the summation in equation 8. Subsequently, following the proof of Theorem 3, we have with probability $1 - n^{-c'}$,

$$\max_{H} |Trev(H, \mathcal{T}) - (1-2\alpha)p_n Trev(H, \mathcal{T}_0)| = O\left(\sqrt{p_n n^7 \log n}\right).$$

and so

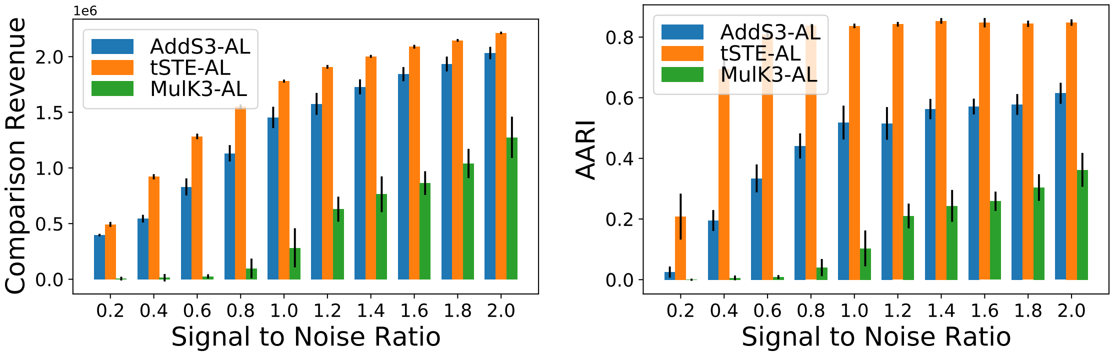
$$\begin{aligned} Trev(\widehat{H}, \mathcal{T}_0) \geq \frac{1}{(1-2\alpha)p_n} Trev(\widehat{H}, \mathcal{T}) - O\left(\sqrt{\frac{n^7 \log n}{p_n}}\right) &\geq \frac{1}{(1-2\alpha)p_n} Trev(H_0, \mathcal{T}) - O\left(\sqrt{\frac{n^7 \log n}{p_n}}\right) \\ &\geq Trev(H_0, \mathcal{T}_0) - O\left(\sqrt{\frac{n^7 \log n}{p_n}}\right), \end{aligned}$$

which leads to the same claim as Theorem 3. ☐

D Additional Results on the Planted Model

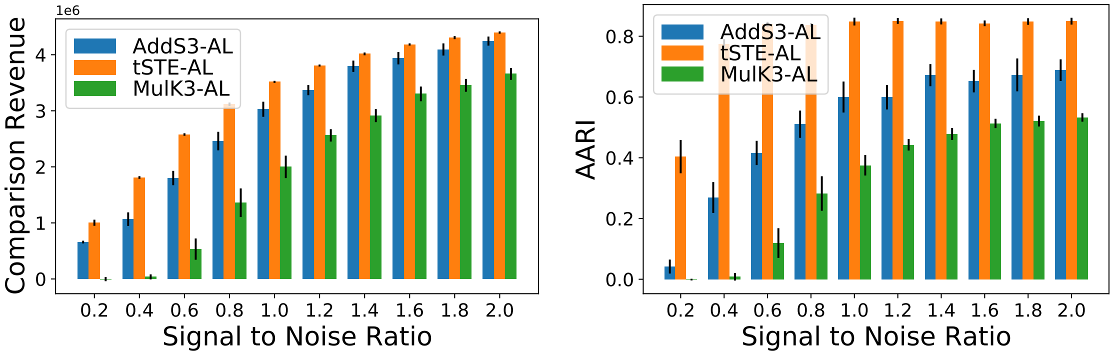
In this section, we provide additional results on the Planted Model presented in Section 7.1 of the main paper.

In Figure 3, we present the results obtained $n^2/2$, n^2 and $2n^2$ triplet comparisons respectively. Similarly, Figure 4 displays the results obtained using $n^2/2$, n^2 and $2n^2$ quadruplet comparisons respectively. In all these figures, we notice that, given a set signal to noise ratio, the ordering between the methods remains the same for the revenue and the AARI, that is the method with the highest revenue also has the highest AARI. In other words, a higher revenue indicates a better dendrogram.



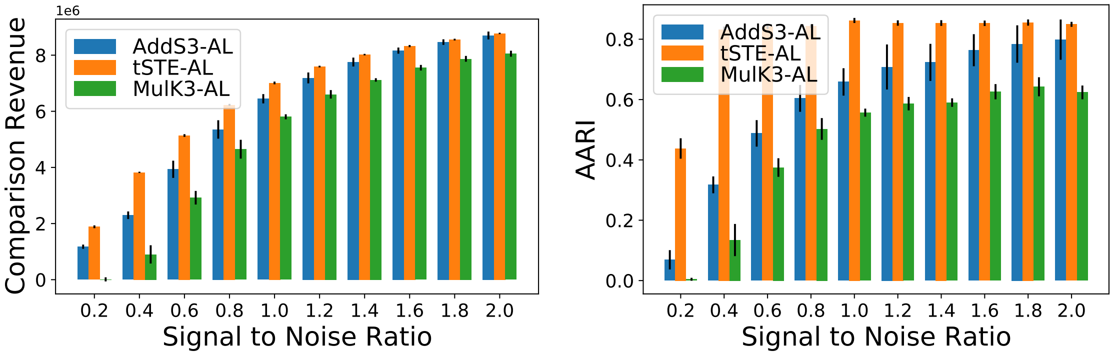
(a) Revenue ($n^2/2$)

(b) AARI ($n^2/2$)



(c) Revenue (n^2)

(d) AARI (n^2)

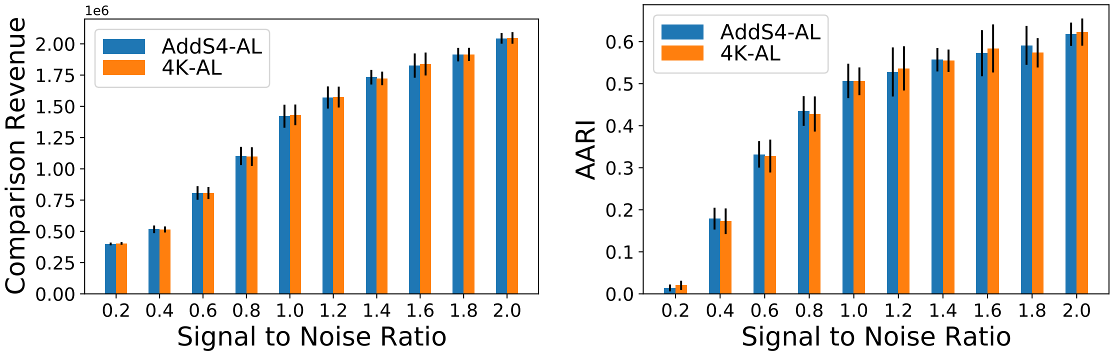


(e) Revenue ($2n^2$)

(f) AARI ($2n^2$)

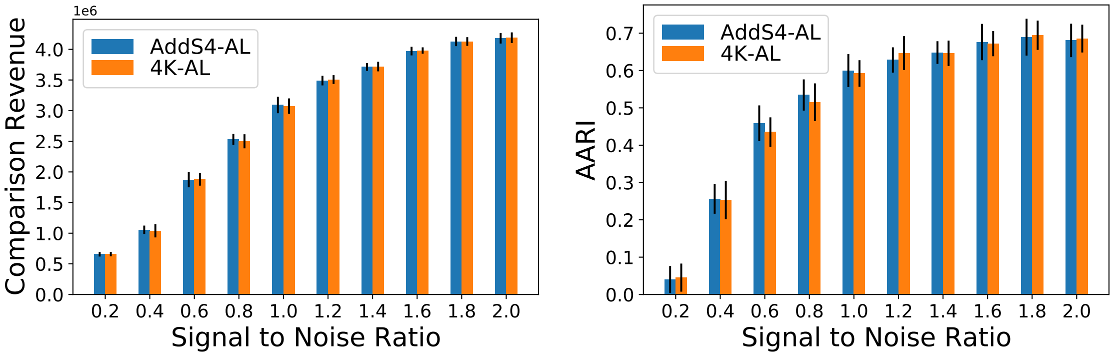
Figure 3: Revenue and AARI (higher is better) of several triplets-based methods using respectively $n^2/2$ comparisons (a-b), n^2 comparisons (c-d), and $2n^2$ comparisons (e-f). Given various signal to noise ratios, a higher revenue implies higher AARI values, that is better dendrograms.

In Table 4 we verify that this remains true for constant signal to noise ratios of 1.5, and halving number of comparisons. The highest revenue and AARI are underlined. We can notice that, when the revenue of AddS3-AL becomes higher than the revenue of tSTE-AL, the AARI also follows the same trend, thus confirming that selecting the dendrogram with the highest revenue is indeed a good way to select meaningful hierarchies. In Table 5 we provide repeatable results with seed varying from 0 to 9 across the 10 runs of each algorithm. This follows the same trend as seen in Table 4, thus demonstrating no dependence on seeding.



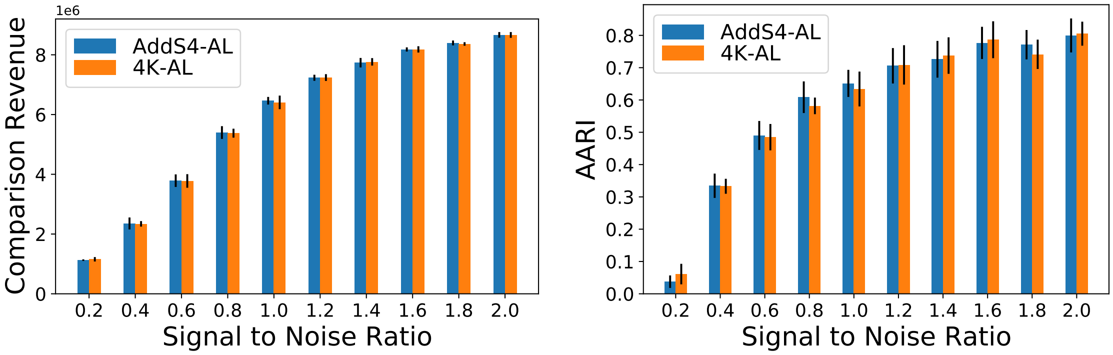
(a) Revenue ($n^2/2$)

(b) AARI ($n^2/2$)



(c) Revenue (n^2)

(d) AARI (n^2)



(e) Revenue ($2n^2$)

(f) AARI ($2n^2$)

Figure 4: Revenue and AARI (higher is better) of several quadruplets-based methods using respectively $n^2/2$ comparisons (a-b), n^2 comparisons (c-d), and $2n^2$ comparisons (e-f). Given various signal to noise ratios, a higher revenue implies higher AARI values, that is better dendrograms.

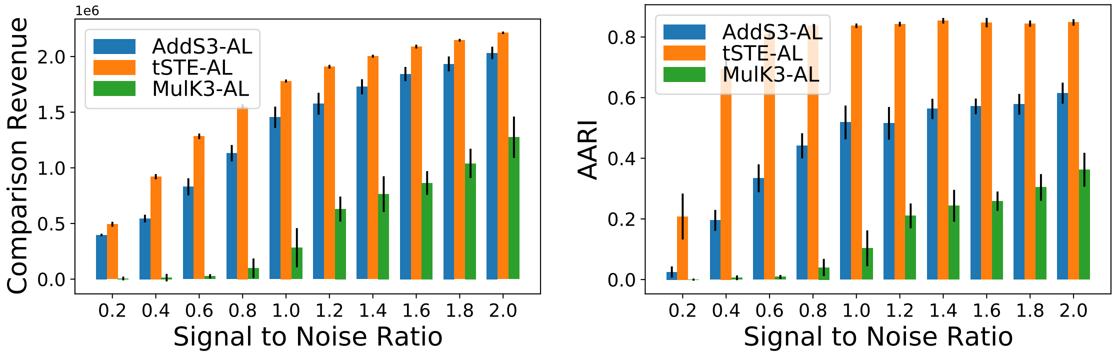
E Results on the Planted Model with Noisy Comparisons

In the main paper, we only used the planted model to generate comparisons with no noise. In this section, we show that our findings remain true even when some of the comparisons are noisy, that is randomly flipped with a probability of 5%.

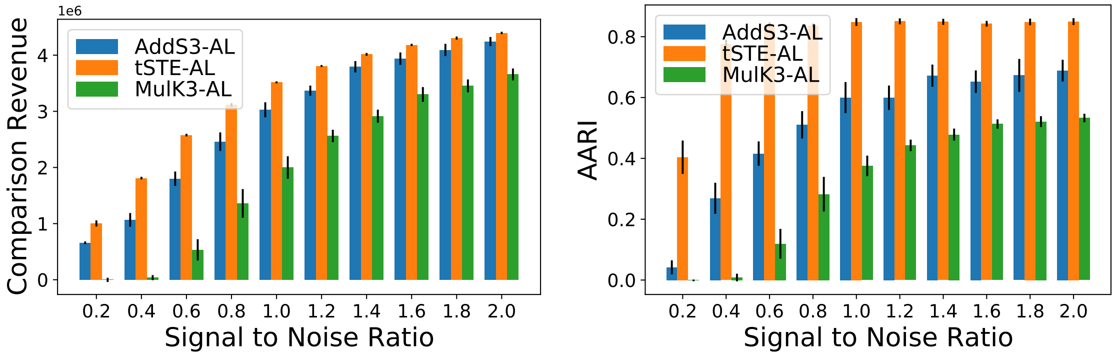
In Figure 5, we present the results obtained using $n^2/2$, n^2 and $2n^2$ noisy triplet comparisons respectively. In Figure 6, we present the results obtained using $n^2/2$, n^2 and $2n^2$ noisy quadruplet comparisons respectively. In all these figures, we notice that, given a set signal to noise ratio, the ordering between the methods remains the same for the revenue and the AARI, that is the method with the highest revenue is also the one with the highest AARI. In other words, a higher revenue indicates a better dendrogram.

F Standard Deviation on Real Data

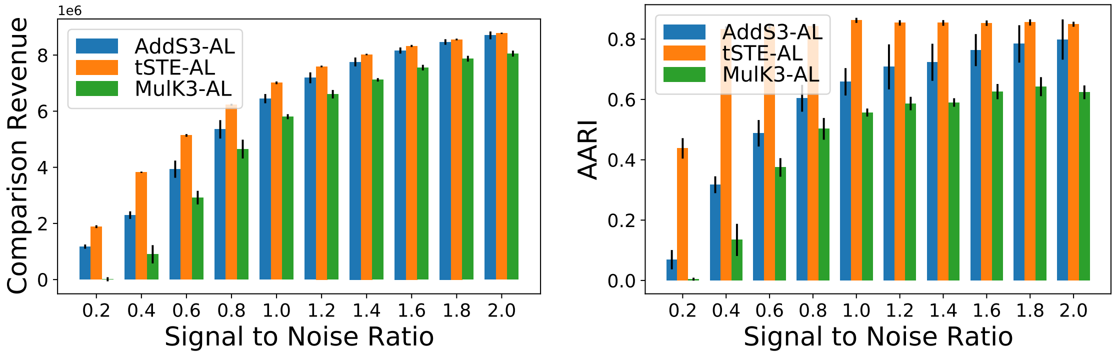
In Table 6, we provide the standard deviations for the real data experiments that were omitted in the main paper.



(a) Revenue ($n^2/2$) (b) AARI ($n^2/2$)

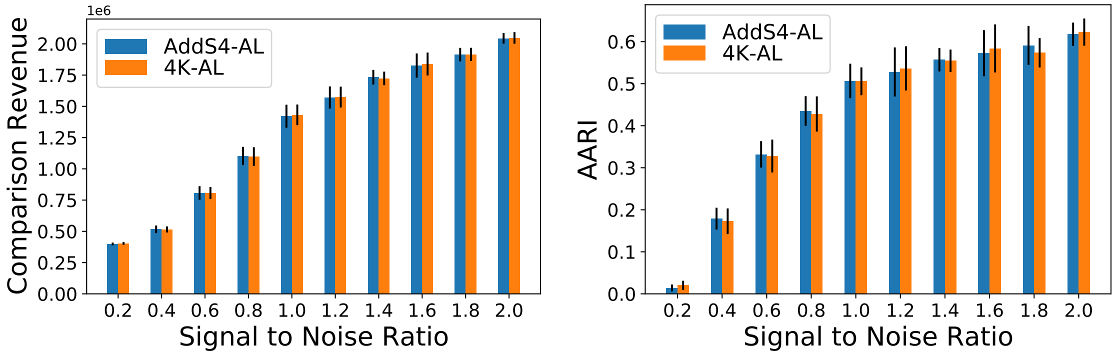


(c) Revenue (n^2) (d) AARI (n^2)



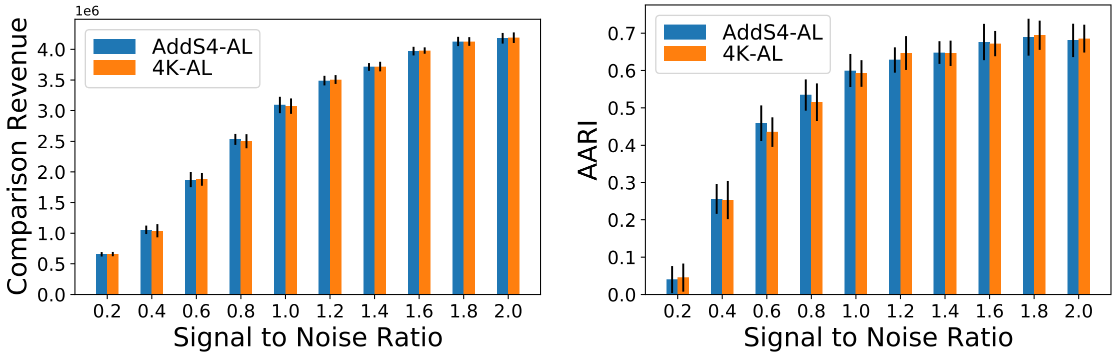
(e) Revenue ($2n^2$) (f) AARI ($2n^2$)

Figure 5: Revenue and AARI (higher is better) of several triplets-based methods using respectively $n^2/2$ comparisons (a-b), n^2 comparisons (c-d), and $2n^2$ comparisons (e-f) with 5% noise. Given various signal to noise ratios, a higher revenue implies higher AARI values, that is better dendrograms.



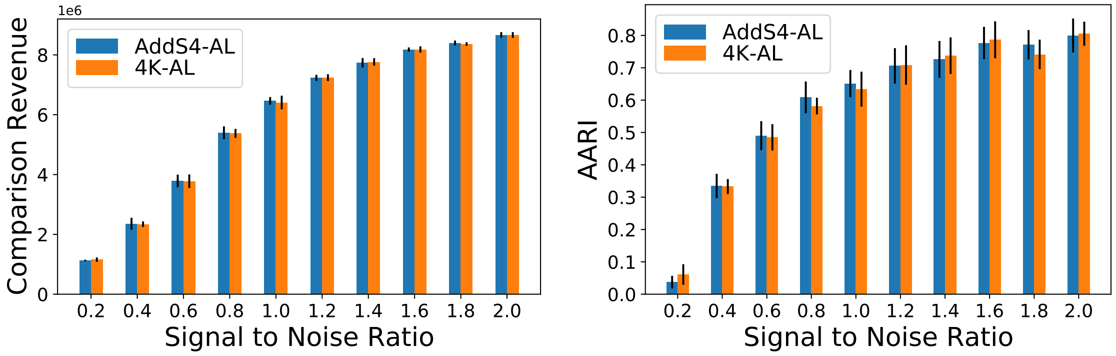
(a) Revenue ($n^2/2$)

(b) AARI ($n^2/2$)



(c) Revenue (n^2)

(d) AARI (n^2)



(e) Revenue ($2n^2$)

(f) AARI ($2n^2$)

Figure 6: Revenue and AARI (higher is better) of several quadruplets-based methods using respectively $n^2/2$ comparisons (a-b), n^2 comparisons (c-d), and $2n^2$ comparisons (e-f) with 5% noise. Given various signal to noise ratios, a higher revenue implies higher AARI values, that is better dendrograms.

Table 4: Revenue and AARI of various methods for a signal to noise ratio of 1.5 and halving of comparisons. In each line the highest revenue and the highest AARI are underlined (taking into account standard deviation), showing that the two measures are well aligned.

Number of triplets	AddS3-AL		tSTE-AL	
	Revenue	AARI	Revenue	AARI
$16n^2$	$\underline{7.336 \times 10^7 \pm 3.9 \times 10^5}$	$\underline{0.937 \pm 0.030}$	$7.286 \times 10^7 \pm 8.9 \times 10^4$	0.849 ± 0.009
$8n^2$	$\underline{3.667 \times 10^7 \pm 7.9 \times 10^4}$	$\underline{0.901 \pm 0.025}$	$3.647 \times 10^7 \pm 4.9 \times 10^4$	0.866 ± 0.010
$4n^2$	$\underline{1.820 \times 10^7 \pm 1.4 \times 10^5}$	$\underline{0.855 \pm 0.054}$	$\underline{1.816 \times 10^7 \pm 5.3 \times 10^4}$	$\underline{0.852 \pm 0.011}$
$2n^2$	$8.927 \times 10^6 \pm 1.4 \times 10^5$	0.778 ± 0.046	$\underline{9.093 \times 10^6 \pm 1.9 \times 10^4}$	$\underline{0.855 \pm 0.009}$
n^2	$4.333 \times 10^6 \pm 9.3 \times 10^4$	0.699 ± 0.063	$\underline{4.559 \times 10^6 \pm 2.5 \times 10^4}$	$\underline{0.856 \pm 0.008}$
$n^2/2$	$2.097 \times 10^6 \pm 6.6 \times 10^4$	0.615 ± 0.043	$\underline{2.289 \times 10^6 \pm 1.8 \times 10^4}$	$\underline{0.849 \pm 0.014}$
$n^2/4$	$9.363 \times 10^5 \pm 5.1 \times 10^4$	0.501 ± 0.040	$\underline{1.138 \times 10^6 \pm 1.0 \times 10^4}$	$\underline{0.847 \pm 0.010}$
$n^2/8$	$4.310 \times 10^5 \pm 2.4 \times 10^4$	0.403 ± 0.036	$\underline{5.710 \times 10^5 \pm 8.3 \times 10^3}$	$\underline{0.841 \pm 0.011}$
$n^2/16$	$2.128 \times 10^5 \pm 1.8 \times 10^4$	0.323 ± 0.038	$\underline{2.853 \times 10^5 \pm 7.3 \times 10^3}$	$\underline{0.667 \pm 0.046}$
$n^2/32$	$1.048 \times 10^5 \pm 5.8 \times 10^3$	0.151 ± 0.031	$\underline{1.438 \times 10^5 \pm 5.7 \times 10^3}$	$\underline{0.543 \pm 0.015}$

Number of triplets	MulK3-AL	
	Revenue	AARI
$16n^2$	$7.304 \times 10^7 \pm 8.4 \times 10^4$	0.863 ± 0.007
$8n^2$	$3.631 \times 10^7 \pm 4.8 \times 10^4$	0.854 ± 0.003
$4n^2$	$1.784 \times 10^7 \pm 7.9 \times 10^4$	0.821 ± 0.008
$2n^2$	$8.357 \times 10^6 \pm 7.7 \times 10^4$	0.660 ± 0.020
n^2	$3.754 \times 10^6 \pm 1.2 \times 10^5$	0.539 ± 0.012
$n^2/2$	$1.254 \times 10^6 \pm 1.7 \times 10^5$	0.353 ± 0.049
$n^2/4$	$1.680 \times 10^5 \pm 7.7 \times 10^4$	0.085 ± 0.044
$n^2/8$	$7.938 \times 10^3 \pm 1.4 \times 10^4$	0.010 ± 0.011
$n^2/16$	$5.156 \times 10^2 \pm 7.1 \times 10^3$	-0.0002 ± 0.003
$n^2/32$	$1.314 \times 10^3 \pm 3.271 \times 10^3$	-0.0004 ± 0.0009

Table 5: Revenue and AARI of various methods for a signal to noise ratio of 1.5 and halving of comparisons. In each line the highest revenue and the highest AARI are underlined, showing that the two measures are well aligned. Here the seed is varied from 0 to 9.

Number of triplets	AddS3-AL		tSTE-AL	
	Revenue	AARI	Revenue	AARI
$16n^2$	$\underline{7.347 \times 10^7 \pm 1.3 \times 10^5}$	$\underline{0.937 \pm 0.024}$	$7.300 \times 10^7 \pm 8.7 \times 10^4$	0.877 ± 0.007
$8n^2$	$\underline{3.667 \times 10^7 \pm 1.7 \times 10^5}$	$\underline{0.905 \pm 0.020}$	$3.656 \times 10^7 \pm 8.8 \times 10^4$	0.877 ± 0.007
$4n^2$	$\underline{1.823 \times 10^7 \pm 1.0 \times 10^5}$	$\underline{0.862 \pm 0.023}$	$\underline{1.825 \times 10^7 \pm 7.0 \times 10^4}$	$\underline{0.874 \pm 0.012}$
$2n^2$	$8.962 \times 10^6 \pm 9.8 \times 10^4$	0.782 ± 0.042	$\underline{9.130 \times 10^6 \pm 4.0 \times 10^4}$	$\underline{0.867 \pm 0.014}$
n^2	$4.315 \times 10^6 \pm 9.7 \times 10^4$	0.682 ± 0.047	$\underline{4.559 \times 10^6 \pm 2.0 \times 10^4}$	$\underline{0.868 \pm 0.012}$
$n^2/2$	$2.038 \times 10^6 \pm 6.8 \times 10^4$	0.593 ± 0.037	$\underline{2.277 \times 10^6 \pm 1.4 \times 10^4}$	$\underline{0.860 \pm 0.014}$
$n^2/4$	$9.268 \times 10^5 \pm 4.2 \times 10^4$	0.498 ± 0.035	$\underline{1.137 \times 10^6 \pm 1.0 \times 10^4}$	$\underline{0.851 \pm 0.065}$
$n^2/8$	$4.261 \times 10^5 \pm 2.5 \times 10^4$	0.396 ± 0.033	$\underline{5.728 \times 10^5 \pm 6.2 \times 10^3}$	$\underline{0.840 \pm 0.010}$
$n^2/16$	$2.015 \times 10^5 \pm 1.2 \times 10^4$	0.295 ± 0.041	$\underline{2.858 \times 10^5 \pm 3.3 \times 10^3}$	$\underline{0.720 \pm 0.057}$
$n^2/32$	$1.096 \times 10^5 \pm 8.5 \times 10^3$	0.192 ± 0.057	$\underline{1.450 \times 10^5 \pm 2.2 \times 10^3}$	$\underline{0.549 \pm 0.025}$

Number of triplets	MulK3-AL	
	Revenue	AARI
$16n^2$	$7.315 \times 10^7 \pm 1.3 \times 10^5$	0.861 ± 0.005
$8n^2$	$3.636 \times 10^7 \pm 9.7 \times 10^4$	0.855 ± 0.003
$4n^2$	$1.795 \times 10^7 \pm 9.3 \times 10^4$	0.830 ± 0.009
$2n^2$	$8.444 \times 10^6 \pm 1.3 \times 10^5$	0.677 ± 0.041
n^2	$3.728 \times 10^6 \pm 1.4 \times 10^5$	0.540 ± 0.016
$n^2/2$	$1.220 \times 10^6 \pm 1.7 \times 10^5$	0.347 ± 0.050
$n^2/4$	$1.531 \times 10^5 \pm 8.9 \times 10^4$	0.077 ± 0.047
$n^2/8$	$1.856 \times 10^4 \pm 1.2 \times 10^4$	0.011 ± 0.011
$n^2/16$	$4.026 \times 10^3 \pm 8.0 \times 10^3$	0.005 ± 0.004
$n^2/32$	$2.015 \times 10^3 \pm 3.5 \times 10^3$	0.0003 ± 0.001

Table 6: Experiments on real datasets. For the triplets-based methods, AddS3-AL tends to obtain the dendrograms with the best revenues. For the quadruplets-based approaches, AddS4-AL and 4K-AL obtain comparable results. Using the original Cosine similarities only yields slightly better hierarchies than comparison-based methods.

Dataset	Triplet			
	AddS3-AL	tSTE-AL	MulK3-AL	Cosine-AL
Zoo	$2.76 \times 10^5 \pm 5 \times 10^3$	$2.18 \times 10^5 \pm 9 \times 10^3$	$2.02 \times 10^5 \pm 2 \times 10^4$	$2.81 \times 10^5 \pm 2 \times 10^3$
Glass	$2.18 \times 10^6 \pm 4 \times 10^4$	$1.99 \times 10^6 \pm 3 \times 10^4$	$1.48 \times 10^6 \pm 8 \times 10^4$	$2.12 \times 10^6 \pm 1 \times 10^4$
MNIST	$1.88 \times 10^9 \pm 5 \times 10^7$	$2.03 \times 10^9 \pm 5 \times 10^7$	$1.75 \times 10^9 \pm 4 \times 10^7$	$2.03 \times 10^9 \pm 5 \times 10^7$
Car	1.52×10^5	$1.56 \times 10^5 \pm 2 \times 10^3$	1.26×10^5	-
Food	6.13×10^6	$5.99 \times 10^6 \pm 2 \times 10^4$	6.09×10^6	-
Vogue	2.72×10^4	$2.10 \times 10^4 \pm 1 \times 10^3$	3.02×10^3	-
Nature	2.65×10^5	$2.06 \times 10^5 \pm 8 \times 10^3$	1.23×10^5	-
Imagenet	7.18×10^7	$6.57 \times 10^7 \pm 8 \times 10^5$	3.44×10^7	-

Dataset	Quadruplet		
	AddS4-AL	4K-AL	Cosine-AL
Zoo	$2.86 \times 10^5 \pm 7 \times 10^3$	$2.86 \times 10^5 \pm 8 \times 10^3$	$2.96 \times 10^5 \pm 2 \times 10^3$
Glass	$2.43 \times 10^6 \pm 2 \times 10^4$	$2.45 \times 10^6 \pm 2 \times 10^4$	$2.49 \times 10^6 \pm 1 \times 10^4$
MNIST	$1.94 \times 10^9 \pm 3 \times 10^7$	$1.91 \times 10^9 \pm 3 \times 10^7$	$2.06 \times 10^9 \pm 4 \times 10^7$
Car	1.52×10^5	1.12×10^5	-
Food	6.13×10^6	6.13×10^6	-
Vogue	2.72×10^4	2.55×10^4	-
Nature	2.65×10^5	2.23×10^5	-
Imagenet	7.18×10^7	6.99×10^7	-