Objective: Study hierarchical clustering when only similarity comparisons are available, that is without features nor explicit similarities.

**Comparison-Based Machine Learning**

Humans are bad at giving unbiased, quantitative information. Better at giving relative information.

**Example:** The left vehicles are more similar to each other than the right vehicles.

Given an unknown similarity function $w$, the corresponding quadruplet is $w$ (SUV left, SUV right) $>$ $w$ (Sport car, Tractor).

**Challenging problem:** No features (coordinates), not even distances! Given a list of quadruplets, can we solve standard machine learning tasks such as clustering?

**Example:** Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a set of $N$ cars. Can we build a dendrogram that reflects their similarities using only a limited set of quadruplets $\mathcal{Q}$?

**Existing solutions:**
- **Embedding-based methods:** Retrieve a Euclidean representation of the objects that respects the quadruplets, then use standard machine learning methods.
- **Direct methods:** Design learning algorithms that directly handle the quadruplets to solve a specific task.

**Obtaining the comparisons:**
- **Actively:** quadruplets chosen by the algorithm.
- **Passively:** quadruplets given to the algorithm with no way to make new queries.

**Contributions:**
We propose new algorithms for hierarchical clustering that directly use quadruplets.

We derive sufficient conditions that guarantee exact recovery of a planted model.

**Hierarchical Clustering**

**Goal:** Construct a dendrogram that reflects the similarities between the objects.

**Idea:** Iteratively group clusters together using a linkage function. Given two clusters $G$ and $G'$, $w_{ij}$ is the distance between $i \in G$ and $j \in G'$.

- **Single Linkage (SL):** $W(G, G') = \min_{i \in G, j \in G'} w_{ij}$.
- **Complete Linkage (CL):** $W(G, G') = \max_{i \in G, j \in G'} w_{ij}$.
- **Average Linkage (AL):** $W(G, G') = \frac{1}{|G||G'|} \sum_{i \in G, j \in G'} w_{ij}$.

**Planted Model**

**Summary:**
We study hierarchical clustering under a noisy hierarchical block matrix. The model complexity is controlled by the size $N_0$ of the pure clusters, by the number of levels $L$ in the hierarchy, and by the signal to noise ratio $\sigma$.

**Expected similarities.**

**Hierarchical structure.**

**Guarantees:**
We empirically verify our theoretical findings: SL and CL only recover the hierarchy for large signal to noise ratios while 4–AL exactly recovers the hierarchy for smaller signal to noise ratios.

**Evaluation:**
We use the Average Adjusted Rand Index (AARI, higher is better).

**Quadruplets-Based Average Linkage (4–AL)**

**Summary:**
We use passive comparisons to define a cluster-level similarity function. If sufficiently large initial clusters are provided, 4–AL obtains better guarantees than 4K–AL.

Cluster-level similarity: Clusters $G_1, G_2$ are more similar to each other than $G_1, G_3$ if their objects are, on average, more similar to each other than the objects of $G_1$ and $G_3$.

$$W_4(G_1, G_2 | G_3) = \frac{1}{N_0^2 - 1} \sum_{i \in G_1} \sum_{j \in G_2} w_{ij} - \frac{1}{N_0^2} \sum_{i \in G_1} \sum_{j \in G_3} w_{ij}$$

**Averaging over all cluster pairs gives rise to the following linkage function:**

$$W(G_1, G_2) = \frac{1}{N(2N - 1)} \sum_{i \neq r, l} w_{ij}$$

**Guarantees:**
With sufficiently large initial clusters, a constant signal to noise ratio, and a sufficient number of comparisons, 4–AL exactly recovers the planted hierarchy.

**Theorem (Exact recovery of planted hierarchy by 4–AL):**

- With $L = \Theta(1)$, $\sigma = \Omega(1)$, and an initial partition of the examples in pure clusters of sizes $N$ in the range $[m, 2m]$ for some $m = \frac{1}{N_0}$, 4–AL exactly recovers the planted hierarchy with high probability using $O\left(\frac{N^{\frac{1}{2}}}{\sigma}\right)$ passive quadruplets.

- With $L = \Theta(1)$, $\sigma = \Omega(1)$, and $\Theta(N_0)$-sized initial clusters, 4–AL exactly recovers the planted hierarchy with high probability using only $O\left(\frac{N^{\frac{1}{2}}}{\sigma}\right)$ passive quadruplets.

**Experiments: Planted Model**

**Summary:**
We empirically verify our theoretical findings: SL and CL only recover the hierarchy for large signal to noise ratios while 4–AL and 4–AL exactly recover the hierarchy for smaller signal to noise ratios.

**Evaluation:**
We use the Average Adjusted Rand Index (AARI, higher is better).