# Average Sensitivity of Hierarchical $k$-Median Clustering

Shijie Li[1] Weiqiang He[1] Ruobing Bai[1] Pan Peng[1]

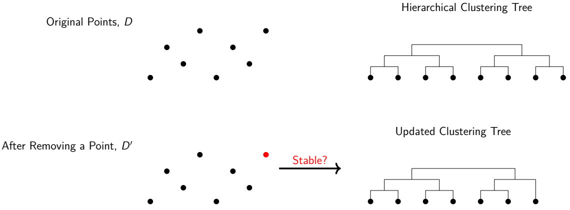
[1] School of Computer Science and Technology,
University of Science and Technology of China

ICML 2025

# Motivation

- Hierarchical clustering reveals data structure at multiple scales
- But classic methods can be highly sensitive to small input changes [BLG14][1]
- This instability harms interpretability and reliability

Our goal: Analyze and minimize the expected change in clustering output (symmetric difference) under perturbation



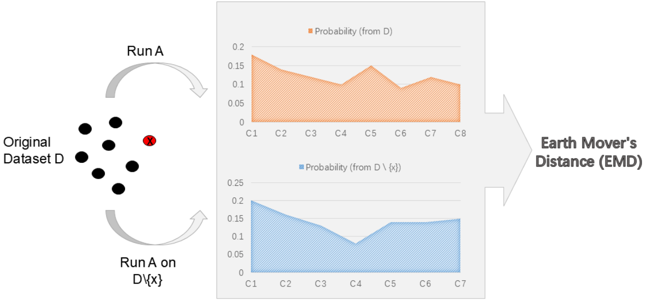

[1] Balcan, Liang, and Gupta. "Robust hierarchical clustering". In JMLR 2014

# Average Sensitivity of Randomized Algorithms

**Setup:** Randomized algorithm $A$ and dataset $P \subseteq [0, \Lambda]^d$

**Average sensitivity** [VY21][2]:

$$\mathrm{avg}_{p \in P} \left[ d_{\mathrm{EM}}(A(P), A(P \setminus \{p\})) \right]$$



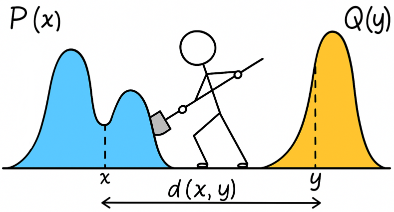

[2] Varma and Yoshida. "Average Sensitivity of Graph Algorithms". In SODA 2021

# Average Sensitivity of Randomized Algorithms

**Earth Mover's Distance (EMD)**:

$$d_{\mathrm{EM}}(A(P), A(P \setminus \{p\})) = \min_{D} \mathbb{E}_{(S,S')\sim D}\left[|S \triangle S'|\right]$$

- $A(P)$, $A(P \setminus \{p\})$: distributions over clustering outputs
- $D$: joint distribution with marginals matching each algorithm output



# Problem Formulation

## Hierarchical $k$-median clustering

- Input: Dataset $P \subseteq [0, \Lambda]^d$
- Output:
  - Centers: $c_1, \ldots, c_n$
  - Clusters: $P_1, \ldots, P_k$ minimizing the $k$-median cost for all $k \in [n]$, i.e.

$$\mathrm{COST}(P, \{c_1, \ldots, c_k\}) = \min_{P_1, \ldots, P_k} \sum_{i \in [k]} \sum_{p \in P_i} \|p - c_i\|$$

- Perturbation model: Uniformly at random delete one point $p \in P$

**Goal:** Design hierarchical $k$-median algorithms with **provably low average sensitivity**

# Our Contributions

## Theorem (Main Theorem, informal)

*Given a point set $P$ of size $n$ and a parameter $\varepsilon > 0$, our algorithm computes a hierarchical $k$-median clustering for all $k \in \{1, \dots, n\}$ with:*

- *Expected cost:*

$$\mathbb{E}[\mathrm{COST}_T(P, S_k)] \le O(d \log \Lambda \cdot (1+\varepsilon)^k) \cdot \mathrm{OPT}(P, k)$$

- *Average sensitivity:* $O\left(\frac{k \ln n}{\varepsilon}\right)$
- *Success probability:* $\ge 1 - \frac{k}{n^2}$
- *Running time:* $O(dn \log \Lambda + n^3)$

Here, $S_k$ is the level-$k$ center set, and $\mathrm{COST}_T(P, S_k)$ is the clustering cost on RHST tree $T$.

# Our Contributions

We prove lower bounds on the average sensitivity of Single Linkage and deterministic CLNSS [CLN+21][3].

**Lemma (Single Linkage)**

*The average sensitivity of Single Linkage is at least $\Omega(n)$.*

**Lemma (Deterministic CLNSS)**

*The average sensitivity of the deterministic CLNSS algorithm is at least $\Omega(n)$.*

[3] Cohen-Addad et al. "Parallel and efficient hierarchical k-median clustering". In NeurIPS 2021

# Our Algorithm

## Low-Sensitivity Hierarchical $k$-Median Algorithm

**Input:** A set of points $P$

**Output:** Centers $c_1, \ldots, c_n$, clusterings $\mathcal{P}_1, \ldots, \mathcal{P}_n$

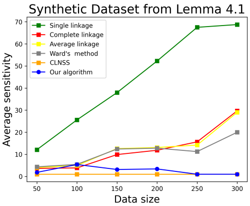
1. Apply a random shift to each point in $P$.
2. Construct a 2-RHST[a] tree $T$.
3. Initialize $S_0 \leftarrow \emptyset$, $\mathcal{P}_0 \leftarrow \{P\}$.
4. Label all internal nodes of the RHST as unlabelled.
5. **For** $t = 1$ to $n$, do the following:
    1. Sample $\lambda$ from a dataset-dependent interval.
    2. Sample $c_t$ with probability $\propto \exp\left(-\mathrm{COST}_T(P, x \cup S_{t-1})/\lambda\right)$.
    3. Label the highest unlabelled ancestor of $c_t$ with $c_t$.
    4. Update $S_t \leftarrow c_t \cup S_{t-1}$.
    5. Define $\mathcal{P}_t$ by assigning points to closest labelled ancestor.

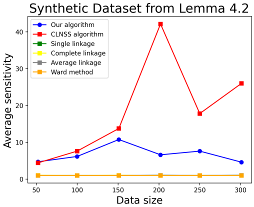
[a] Restricted 2-hierarchically well-separated tree (2-RHST)

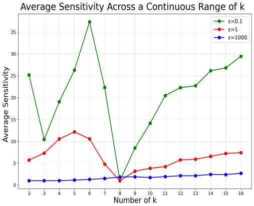
# Experimental Setup

- **Datasets:** Synthetic and real-world (Scikit-learn , UCI Repository )
- **Baselines:**
  - Hierarchical methods: single, complete, average, Ward's
  - CLNSS algorithm
- **Metrics:**
  - Average sensitivity (robustness to deletions)
  - Clustering cost (e.g., $k$-median)
  - Effect of $\varepsilon$ (randomness impact)

# Experimental Results on Synthetic Datasets

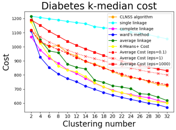


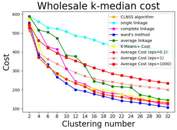



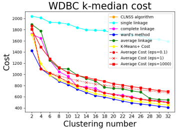



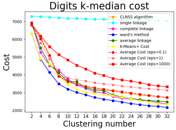

- Left and middle: average sensitivity of Single Linkage and CLNSS (vs others) on synthetic datasets to show the lower bounds.
- Right: results on a synthehtic regression dataset with 500 points.

# Experimental Results on Real-World Datasets

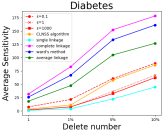


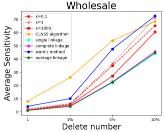



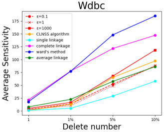



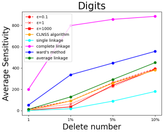




- $k$-**Median Cost:** Comparison across algorithms for varying $k$ on real datasets.




- **Average Sensitivity ($k = 4$):** Slightly worse than single linkage, but better than all other methods.

# Thank you!