Distributed *k*-**Clustering** with Heavy Noise (NeurlPS 2018 Poster)

Introduction

- Input: Data set P from metric space (X, d) of size n. P is distributed onto m different machines as $P = P_1 \cup P_2 \cup \ldots \cup P_m.$
- **Output:** A set $C \subset P$ of k centers and a set Zoutliers, that minimize some cost function cost(P)• (k, z)-center: $cost(A, B) := max_{p \in A} d(p, B)$
- (k, z)-median: $cost(A, B) := \sum_{p \in A} d(p, B)$
- (k, z)-means: $cost(A, B) := \sum_{p \in A} d^2(p, B)$
- Communication model: The MapReduce model. There exists a single coordinator S, and only communication between the coordinator and the machines are allowed.
- Major concerns:
- Clustering quality: The solution (C, Z) should achieve O(1)-approximation, i.e. $cost(P \setminus Z, C) \leq O(1) \cdot OPT$.
- Communication cost: Focus on the case when data is heavily noisy, i.e., $z \gg k, m.$

Motivating Question

Can we achieve constant approximation with communication cost o(z)?

- No: Any O(1)-approximation algorithm needs $\Omega(z)$ communication cost.
- Yes: If we allow removing slightly more than z outliers: **Def.** (C, Z) is an (α, β) -approximation if $\cot(P \setminus Z, C) \leq \alpha \cdot \mathsf{OPT} \text{ and } |Z| \leq \beta z$

Two-Levels Clustering Framework

- Each machine *i* construct a local summary P'_i and send to the coordinator machine S
- **2** The coordinator S solves a single (k, z)-clustering over the aggregated summaries $\bigcup_{i \in [m]} P'_i$ to get final solution (C, Z).
- Folklore: view each summary P'_i as local clustering centers on P_i , then if each P'_i incurs small clustering cost, S can find a good global centers by clustering $\bigcup_{i \in [m]} P'_i$.

Xiangyu Guo and Shi Li

{xiangyug,shil}@buffalo.edu

Department of Computer Science & Engineering, State University of New York at Buffalo

Distributed (k, z)-Center

Results

(O(1),1)
$(O(1), 2+\epsilon)$
$(O(1), 1+\epsilon)$

Local Summary Construction

Parameter: A number L > 0.

- **1 While** $\exists p \in P_i \text{ s.t. } |\mathsf{ball}(p, 2L) \cup P_i| > \frac{\epsilon z}{km}$: • Add p to P'_i and set $w'_p = |\mathsf{ball}(p, 4L) \cup P_i|$ • Remove $\mathsf{ball}(p, 4L)$ from P_i
- Lemma. If $L \ge \mathsf{OPT}$, then $\sum_{i \in [m]} |P'_i| \le mk (1 + \epsilon^{-1})$ and $\sum_{i \in [m]} \sum_{p \in P'_i} w_p \ge n - (1 + \epsilon) z.$

• Remark.

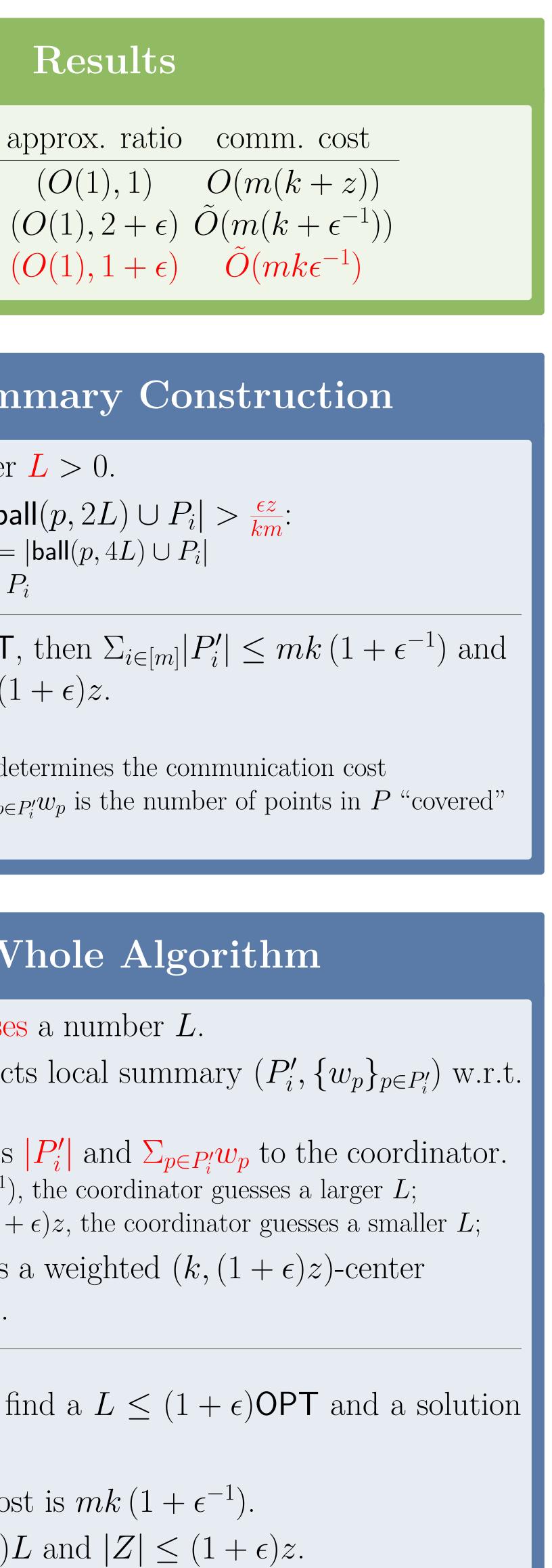
• the total size $\sum_{i \in [m]} |P'_i|$ determines the communication cost • the total weight $\sum_{i \in [m]} \sum_{p \in P'_i} w_p$ is the number of points in P "covered" by the summary

The Whole Algorithm

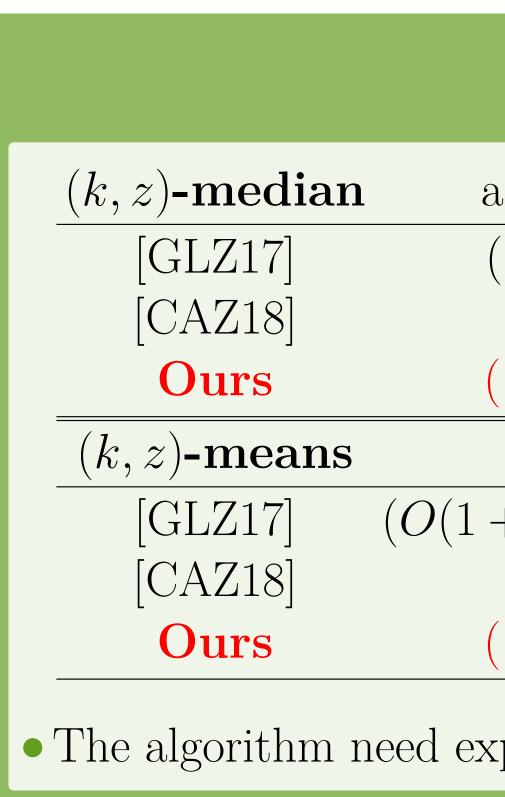
- **1** The coordinator guesses a number L.
- **2** Each machine constructs local summary $(P'_i, \{w_p\}_{p \in P'_i})$ w.r.t. L.
- **3** Each machine sends its $|P'_i|$ and $\sum_{p \in P'_i} w_p$ to the coordinator.
- if $\sum_{i \in [m]} |P'_i| > mk(1 + \epsilon^{-1})$, the coordinator guesses a larger L;
- if $\sum_{i \in [m]} \sum_{p \in P'_i} w_p < n (1 + \epsilon)z$, the coordinator guesses a smaller L;
- The coordinator solves a weighted $(k, (1 + \epsilon)z)$ -center problem over $\bigcup_{i \in [m]} P'_i$.
- **Theorem 1.** We can find a $L \leq (1 + \epsilon)\mathsf{OPT}$ and a solution (C, Z) s.t.
- The communication cost is $mk(1 + \epsilon^{-1})$. • $\operatorname{cost}(P \setminus Z, C) = O(1)L$ and $|Z| \leq (1 + \epsilon)z$.

$$\subset P \text{ of } \mathbf{z}$$

 $\langle Z, C \rangle$:



Distributed (k, z)-Median/Means



Local Summary Construction

Parameter: A number L > 0.

- defined by threshold distance: (k, z)-median/means respectively.
- converted to a k-clustering problem.

- **1** Discretize the set of all possible L as each $L \in \mathbb{L}$.

Results	
approx. ratio	comm. cost
$(O(1), 2 + \epsilon)$	$\tilde{O}(m(\epsilon^{-1}+k))$
(O(1),1)	$O(k \log n + z)$
$(1+\epsilon, 1+\epsilon)$	$\tilde{O}\left(k\epsilon^{-3} + mk\epsilon^{-1}\right)$
$(+1/\delta), 2+\delta+\epsilon$) $\tilde{O}(m(\delta^{-1}+k))$
(O(1),1)	$O(k \log n + z)$
$(1+\epsilon,1+\epsilon)$	$\tilde{O}\left(k\epsilon^{-5} + mk\epsilon^{-1}\right)$
xponential running time (in m, k, ϵ^{-1}).	

• Each machine i samples a coreset Q_i^L w.r.t. a cost function $\operatorname{cost}_L(P, C) := \sum_{p \in P} d_L(p, C)^l - zL^l$, where $d_L(p, C) := \min\{L, d(p, C)\}, \text{ and } l = 1, 2 \text{ for } l$ • Lemma. Let (C^*, Z^*) denote the optimal solution, then

 $\operatorname{cost}(P \setminus Z^*, C^*) = \sup_{L>0} \{ \Sigma_{p \in P} d_L(p, C^*) - zL \}$ • **Remark.** For each fixed L, the (k, z)-clustering problem is

The Whole Algorithm

 $\mathbb{L} = \{L_{\min}, (1+\epsilon)L_{\min}, (1+\epsilon)^2 L_{\min}, \cdots L_{\max}\}$ **2** Each machine i creates multiple local summaries, one for

3 The coordinator solves a min-max k-clustering problem on the aggregated coresets: $\min_C \sup_{L \in \mathbb{L}} \operatorname{cost}_L(P, C)$.