Massively Parallel k-Means Clustering for Perturbation Resilient Instances

Vincent Cohen-addad Vahab Mirrokni ***Peilin Zhong***

Google Research

Euclidean k-Means Clustering

- Unsupervised learning
  - Partition points into k groups
  - Similar points are in the same group

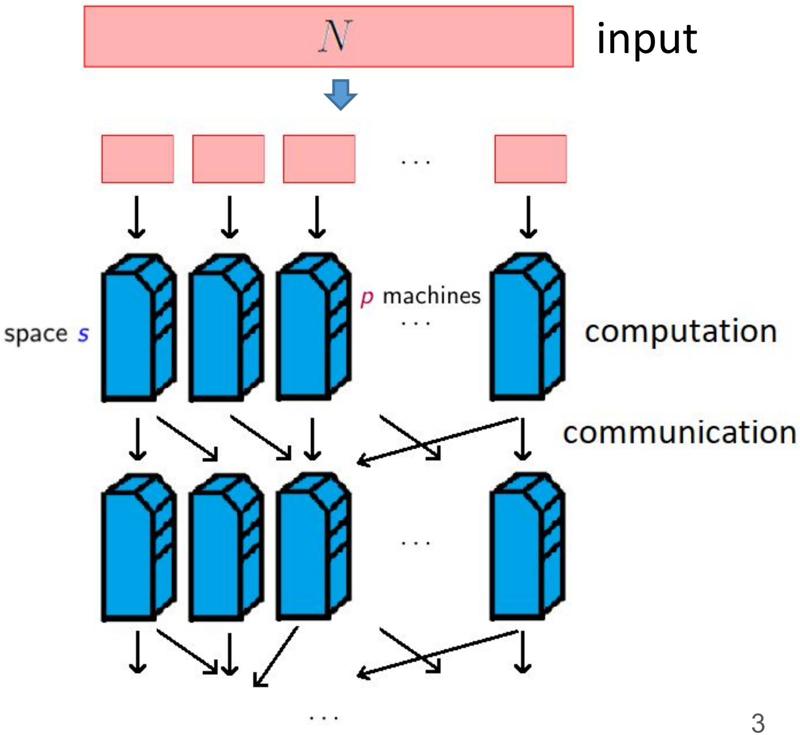
- Euclidean k-means clustering
  - Input: n points $p_1, p_2, \ldots, p_n \in \mathbf{R}^d$
  - Goal: find centers $c_1, c_2, \ldots, c_k \in \mathbf{R}^d$ s.t. the clustering cost $\Sigma_{i\in[n]}\min_{j\in[k]}|| p_i - c_j ||_2^2$ is minimized

- Scalable parallel/distributed algorithms are desired to handle massive data

Massively Parallel Computation (MPC)

- MPC model
  - An abstraction of MapReduce
  - Sublinear local memory
  - Computation proceeds in rounds
  - Bounded communication

- Efficiency Measure
  - Number of rounds (parallel time)
  - Total space
  - Local memory



MPC k-Means Clustering

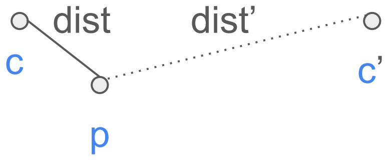
- **Input:** n-point set P in $\mathbf{R}^d$ distributed on several machines
- **Output:** k center points distributed on several machines

- Previous results
  - Small # of rounds & local space but large $\Omega(\log n)$ approximation
  - Small approximation factor & # of rounds but large $\Omega(k)$ local space
  - Small approximation factor & local space but large $\Omega(\log n)$ number of rounds
  - $O(1)$ approximation, $o(\log n)$ rounds, $o(k)$ local space is impossible under certain conditions

- Our result
  - Consider natural well-structured point set
  - $O(1)$ rounds, n^δ local space for any constant $\delta>0$, $1+\varepsilon$ approximation, near linear total space
  - If local space is $\Omega(k)$, the **exact** optimal k-means solution is obtained

Perturbation Resilient Instances

- α-Perturbation resilience → α-center proximity
  - Let C be the optimal solution
  - If p is in a cluster with center c∈C, then $\alpha \cdot ||p - c||_2 \leq ||p - c'||_2$ for any other center c'∈C



α·dist ≤ dist'

Our Techniques

- Candidate clusters via locality sensitive hashing (LSH)
  - LSH → near neighbor graph for different scales
  - Optimal cluster → connected component
  - Candidate clusters → Hierarchical tree structure

- O(1)-round dynamic programming over small depth tree
  - A novel task scheduling process via subtree generation