Novelty Assessment Report

Paper: Distributed Algorithms for Euclidean Clustering
PDF URL: https://openreview.net/pdf?id=17h5Sl2EaK
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

We study the problem of constructing $(1+\varepsilon)$-coresets for Euclidean (k,z)-clustering in the distributed setting, where N data points are partitioned across S sites. We focus on two prominent communication models: the coordinator model and the blackboard model. In the coordinator model, we design a protocol that achieves a $(1+\varepsilon)$-strong coreset with total communication complexity $\tilde{O}\left(sk + \frac{dk}{\min(\varepsilon^4,\varepsilon^{2+z})} + dk\log(n\Delta)\right)$ bits, improving upon prior work (Chen et al., NeurIPS 2016) by eliminating the need to communicate explicit point coordinates in-the-clear across all servers. In the blackboard model, we further reduce the communication complexity to $\tilde{O}\left(s\log(n\Delta) + dk\log(n\Delta) + \frac{dk}{\min(\varepsilon^4,\varepsilon^{2+z})}\right)$ bits, achieving better bounds than previous approaches while upgrading from constant-factor to $(1+\varepsilon)$-approximation guarantees. Our techniques combine new strategies for constant-factor approximation with efficient coreset constructions and compact encoding schemes, leading to optimal protocols that match both the communication costs of the best-known offline coreset constructions and existing lower bounds (Chen et al., NeurIPS 2016, Huang et. al., STOC 2024), up to polylogarithmic factors.

Core Task Landscape

This paper addresses: **Constructing Coresets for Euclidean Clustering in Distributed Settings**
A total of **30 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Distributed Coreset Construction Protocols**
- **Federated Learning with Coresets**
- **Coreset Construction Algorithms and Theory**
- **Application-Driven Coreset Systems**

Complete Taxonomy Tree

- Constructing Coresets for Euclidean Clustering in Distributed Settings Survey Taxonomy
- Distributed Coreset Construction Protocols
  - Coordinator-Based Communication Models ★ (3 papers)
  - [0] Distributed Algorithms for Euclidean Clustering (Anon et al., 2026) View paper
  - [14] Distributed PCA and k-means clustering (Yingyu Liang, 2013) View paper
  - [16] Accurate MapReduce Algorithms for -median and -means in General Metric Spaces (A Mazzetto, 2019) View paper
  - Decentralized Communication Models (2 papers)
  - [11] Scalable decision fusion algorithm for enabling decentralized computation in distributed, big data clustering problems (H. S. Jennath, 2024) View paper
  - [19] Coreset-sharing based Collaborative Model Training among Peer Vehicles (Han Zheng, 2024) View paper
  - MapReduce and Parallel Frameworks (3 papers)
  - [3] Distributed -means and -median Clustering on General Topologies (MFF Balcan, 2013) View paper
  - [6] Distributed k-means with outliers in general metrics (Enrico Dandolo, 2022) View paper
  - [8] Distributed balanced clustering via mapping coresets (MohammadHossein Bateni, 2014) View paper
- Federated Learning with Coresets
  - Horizontal Federated Learning (3 papers)
  - [1] Fedcore: Straggler-free federated learning with distributed coresets (Hongpeng Guo, 2024) View paper
  - [22] Coreset-Based Data Reduction for Machine Learning at the Edge (Hanlin Lu, 2023) View paper
  - [26] Machine Learning at the Edge (Marian Verhelst, 2023) View paper
  - Vertical Federated Learning (2 papers)
  - [9] Coresets for Vertical Federated Learning: Regularized Linear Regression and -Means Clustering (L Huang, 2022) View paper
  - [28] Coresets for Vertical Federated Learning: Regularized Linear Regression and K-Means Clustering (Huang, 2022) View paper
- Coreset Construction Algorithms and Theory
  - Lightweight and Additive-Error Coresets (2 papers)
  - [4] Scalable k-means clustering via lightweight coresets (Olivier Bachem, 2018) View paper
  - [23] Scalable and Distributed Clustering via Lightweight Coresets (Bachem, 2022) View paper
  - Deterministic and Sparse Data Coresets (1 papers)
  - [5] Deterministic Coresets for k-Means of Big Sparse Data (Artem Barger, 2020) View paper
  - Composable and Streaming Coresets (2 papers)
  - [7] Fair Coresets and Streaming Algorithms for Fair k-means (Melanie Schmidt, 2019) View paper

  - [15] A composable coreset for k-center in doubling metrics (Sepideh Aghamolaei, 2019) View paper
  - Metric-Specific Coreset Constructions (2 papers)
  - [2] Coreset for line-sets clustering (S Lotan, 2022) View paper
  - [24] Coresets for Clustering in Graphs of Bounded Treewidth (Baker, 2019) View paper
  - Bayesian and Probabilistic Coreset Methods (1 papers)
  - [20] Distributed Bayesian Coresets (Vladimir Omelyusik, 2023) View paper
- Application-Driven Coreset Systems
  - Mixture Model Training (3 papers)
  - [10] Training gaussian mixture models at scale via coresets (Lucic, 2018) View paper
  - [18] Scalable training of mixture models via coresets (Dan Feldman, 2011) View paper
  - [25] Training mixture models at scale via coresets (Mario Lucic, 2017) View paper
  - Communication-Efficient Machine Learning (2 papers)
  - [17] Joint Coreset Construction and Quantization for Distributed Machine Learning (Lu Hanlin, 2022) View paper
  - [21] Robust Coreset Construction for Distributed Machine Learning (Hanlin Lu, 2019) View paper
  - Sensor Networks and Streaming Data (2 papers)
  - [12] Towards faster big data analytics for antiâ□□jamming applications in vehicular adâ□□hoc network (Hind Bangui, 2021) View paper
  - [29] An effective coreset compression algorithm for large scale sensor networks (Dan Feldman, 2012) View paper
  - Robust and Specialized Clustering Tasks (3 papers)
  - [13] Unsupervised Clustering for Condition Monitoring of Electric Drives (Perumal, 2025) View paper
  - [27] Anomaly detection in vertically partitioned data by distributed core vector machines (Marco Stolpe, 2013) View paper
  - [30] Efficient Implementation of Coreset-based K-Means Methods (Xiaobo, 2021) View paper

Narrative

Core task: constructing coresets for Euclidean clustering in distributed settings. The field organizes around four main branches that reflect different emphases in how coresets are built and deployed. Distributed Coreset Construction Protocols focus on communication models and coordination strategies, examining how nodes exchange summaries under various network topologies—ranging from coordinator-based schemes like those in Distributed k-means Topologies[3] and MapReduce k-median[16] to fully decentralized approaches. Federated Learning with Coresets adapts coreset techniques to privacy-sensitive scenarios, as seen in Fedcore[1] and Vertical Federated Coresets[9], where data remains partitioned across clients. Coreset Construction Algorithms and Theory develops the mathematical foundations and sampling strategies that guarantee approximation quality, including work on lightweight constructions like Lightweight Coresets[4] and robust methods such as Robust Coreset Construction[21]. Application-Driven Coreset Systems translate these ideas into domain-specific deployments, from edge computing platforms to vehicular networks, exemplified by Edge Coresets[22] and Peer Vehicle Coresets[19].

Recent lines of work highlight trade-offs between communication overhead, approximation guarantees, and scalability. Many studies explore how hierarchical aggregation or merge-and-reduce paradigms can minimize rounds of communication while preserving clustering quality, whereas others investigate robustness to outliers or fairness constraints. Distributed Euclidean Clustering[0] sits squarely within the Coordinator-Based Communication Models branch, sharing the emphasis on centralized aggregation seen in Distributed PCA k-means[14] and MapReduce k-median[16]. Compared to these neighbors, Distributed Euclidean Clustering[0] appears to refine protocols for merging local coresets under a coordinator, balancing theoretical guarantees with practical efficiency in settings where a central node orchestrates the summarization process.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Distributed PCA and k-means clustering

Authors: Yingyu Liang, Maria-Florina Balcan | **Year/Venue**: 2013 | **URL**: View paper

Abstract

This paper proposes a distributed PCA algorithm, with the theoretical guarantee that any good approximation solution on the projected data for k-means clustering is also a good approximation on the original data, while the projected dimension required is independent of the original dimension. When combined with the dis-tributed coreset-based clustering approach in [3], this leads to an algorithm in which the number of vectors communicated is independent of the size and the di-mension of the orig...

Relationship Analysis

Both papers belong to the coordinator-based communication models category, focusing on protocols where a central coordinator aggregates information from distributed sites to construct coresets for Euclidean clustering. The candidate paper addresses distributed k-means clustering by combining distributed PCA for dimension reduction with coreset construction, achieving communication cost independent of original data dimension, while the original paper focuses on constructing $(1+\varepsilon)$-coresets through adaptive sampling and lazy update mechanisms without explicit dimension reduction preprocessing. The key difference is that the candidate paper uses PCA as a preprocessing step before coreset construction, whereas the original paper directly constructs coresets through novel adaptive sampling protocols and compact encoding schemes.

---

2. Accurate MapReduce Algorithms for -median and -means in General Metric Spaces

Authors: A Mazzetto, A Pietracaprina, G Pucci | **Year/Venue**: 2019 | **URL**: View paper

Abstract

â¦ a clustering of P. While scarcely meaningful for general metric spaces, for Euclidean spacesâ¦ We devise new distributed coreset constructions and show how to employ them to yield accuâ¦

Relationship Analysis

Both papers belong to the coordinator-based communication models category, focusing on protocols where a central coordinator aggregates information from distributed sites to construct coresets for Euclidean clustering. The original paper develops protocols for (k,z)-clustering in both coordinator and blackboard models with communication complexity $\tilde{O}(sk + dk/\min(\varepsilon^4,\varepsilon^{2+z}) + dk\log(n\Delta))$ bits, while the candidate paper presents MapReduce algorithms for k-median and k-means in general metric spaces with focus on doubling dimension D, achieving $O(|P|^{2/3}k^{1/3}(c/\varepsilon)^{2D}\log^2|P|)$ local memory in 3 rounds. The key difference is that the original paper optimizes total communication bits across all sites with lazy adaptive sampling and coordinate-wise techniques, whereas the candidate paper emphasizes local memory efficiency and coreset composability through CoverWithBalls construction in the MapReduce framework.

Contributions Analysis

Overall novelty summary. The paper contributes communication-optimal protocols for constructing (1+ε)-coresets in distributed Euclidean (k,z)-clustering, targeting both coordinator and blackboard models. Within the taxonomy, it resides in the Coordinator-Based Communication Models leaf, which contains only three papers total. This leaf sits under Distributed Coreset Construction Protocols, one of four major branches in the field. The small sibling count suggests this is a relatively focused research direction rather than a densely populated area, though the broader Distributed Coreset Construction Protocols branch encompasses multiple communication paradigms.

The taxonomy reveals neighboring leaves addressing Decentralized Communication Models (peer-to-peer and blackboard without central coordination) and MapReduce and Parallel Frameworks (multi-worker paradigms). The paper's dual focus on coordinator and blackboard models bridges these directions: the coordinator protocol aligns with centralized aggregation strategies seen in sibling work, while the blackboard protocol connects to decentralized approaches. The taxonomy's scope notes clarify that coordinator-based methods exclude peer-to-peer architectures, positioning this work at the boundary between centralized and decentralized paradigms within the distributed coreset landscape.

Among thirteen candidates examined, the analysis identifies one refutable pair for the blackboard model contribution, while the coordinator model protocol shows no clear refutation across ten candidates examined. The lazy adaptive sampling and coordinate-wise sampling techniques were not evaluated against prior work in this limited search. The blackboard model contribution, despite one overlapping candidate, still claims improved bounds and upgraded approximation guarantees. The coordinator model contribution appears more distinct within the examined literature, though the search scope remains modest relative to the broader field captured in the thirty-paper taxonomy.

Given the limited search scale—thirteen candidates from semantic matching—the analysis provides a snapshot rather than exhaustive coverage. The paper's position in a small taxonomy leaf with only two siblings suggests it addresses a specific niche within distributed coreset construction. The contribution-level statistics indicate varying degrees of prior work overlap, with the coordinator protocol appearing more novel among examined candidates while the blackboard protocol engages more directly with existing approaches.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Communication-optimal protocol for coordinator model

Description: The authors present a distributed clustering protocol for the coordinator model that constructs (1+ε)-coresets using $\tilde{O}(sk + dk/\min(\varepsilon^4,\varepsilon^{2+z}) + dk\log(n\Delta))$ bits of communication. This improves upon prior work by eliminating the need to communicate explicit point coordinates across all servers and matches known lower bounds up to polylogarithmic factors.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. FedCode: Communication-Efficient Federated Learning via Transferring Codebooks

URL: View paper

Brief Assessment

FedCode[31] addresses federated learning communication efficiency through weight clustering and codebook transfer, not distributed clustering protocols for constructing coresets in coordinator models as described in the original paper.

2. Gradient coreset for federated learning

URL: View paper

Brief Assessment

Gradient Coreset Federated[33] focuses on federated learning with coreset selection for machine learning models, not distributed clustering protocols or communication-optimal coreset construction in coordinator models for Euclidean clustering.

3. Distributed k-means with outliers in general metrics

URL: View paper

Brief Assessment

k-means Outliers Metrics[6] focuses on distributed k-means with outliers in general metrics using MapReduce, not on communication-optimal protocols for coordinator models achieving strong coresets as described in the original paper.

4. CBFL: A Communication-Efficient Federated Learning Framework From Data Redundancy Perspective

URL: View paper

Brief Assessment

CBFL[34] focuses on federated learning with coreset-based model compression for mobile devices, not distributed clustering protocols in coordinator models with strong coreset guarantees.

5. Distributed -means and -median Clustering on General Topologies

URL: View paper

Brief Assessment

Distributed k-means Topologies[3] focuses on general communication topologies and coreset construction via local approximations, not on the specific coordinator model protocol with the claimed communication bounds.

6. Robustð-Center Clustering for Contin-uous Monitoring

URL: View paper

Brief Assessment

Robust k-Center Monitoring[36] focuses on the k-center problem with outliers in continuous monitoring settings, not general (k,z)-clustering protocols. The technical approaches differ fundamentally in problem formulation and objectives.

7. Coresets for Vertical Federated Learning: Regularized Linear Regression and K-Means Clustering

URL: View paper

Brief Assessment

Vertical Federated Regression[28] focuses on vertical federated learning with feature partitioning across parties, not general distributed clustering with data partitioning across sites in coordinator models.

8. k-center clustering with outliers in the MPC and streaming model

URL: View paper

Brief Assessment

k-center MPC Streaming[32] focuses on k-center clustering with outliers in MPC and streaming models, not general (k,z)-clustering protocols. The candidate addresses different problem variants and computational models than the original paper's distributed Euclidean clustering framework.

9. Coresets for Vertical Federated Learning: Regularized Linear Regression and -Means Clustering

URL: View paper

Brief Assessment

Vertical Federated Coresets[9] focuses on vertical federated learning with feature partitioning across parties, not general distributed clustering with data partitioning across sites in coordinator models.

10. Randomized greedy algorithms and composable coreset for k-center clustering with outliers

URL: View paper

Brief Assessment

Greedy k-center Outliers[35] focuses on k-center clustering with outliers using randomized greedy algorithms and coreset construction in doubling metrics. The candidate does not address general Euclidean (k,z)-clustering protocols or communication-optimal distributed frameworks for coordinator models as described in the original paper.

Contribution 2: Communication-optimal protocol for blackboard model

Description: The authors develop a distributed clustering protocol for the blackboard model that achieves $(1+\varepsilon)$-approximation guarantees using $\tilde{O}(s\log(n\Delta) + dk\log(n\Delta) + dk/\min(\varepsilon^4,\varepsilon^{2+z}))$ bits of communication. This improves upon previous constant-factor approximations while reducing communication costs and matching existing lower bounds.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Communication-optimal distributed dynamic graph clustering

URL: View paper

Prior Art Analysis

Distributed Graph Clustering[37] demonstrates prior work on communication-optimal protocols for the blackboard model with similar approximation guarantees and communication complexity. The candidate paper presents a protocol achieving $\tilde{O}(s\log(n\Delta) + dk\log(n\Delta) + dk/\min(\varepsilon^4,\varepsilon^{2+z}))$ bits for distributed clustering in the blackboard model, while the original paper claims this as a novel contribution. The candidate's algorithm d2-cabl achieves $\tilde{O}(n + s)$ communication cost for graph clustering in the blackboard model, which when adapted to the original paper's setting would yield comparable bounds. Both papers address distributed clustering with approximation guarantees in the blackboard communication model, though applied to different problem domains (graph clustering vs. Euclidean clustering).

Evidence

Evidence 1 - **Rationale**: Both papers present formal theorems establishing communication-optimal protocols for the blackboard model with approximation guarantees. The candidate's theorem demonstrates that communication-optimal bounds with polylogarithmic factors were established for distributed clustering in the blackboard model before the original paper. - **Original**: theorem 1.2 (communication-optimal clustering in the blackboard model, informal). there exists a protocol on n points distributed across s sites that produces a (1 + ε)-strong coreset for (k, z)- clustering that uses ˜o s log(nΔ) + dk log(nΔ) + dk min(ε4,ε2+z) total bits of communication in the bl... - **Candidate**: theorem 4 (the blackboard model). for every time point τ ∈ [1,t], suppose that gτ satisfies υ(k) = ω(k3) and there is an optimal partitionp1,··· ,pk which achieves ρ(k) for some positive integer k, w.h.p. d2-cabl can output partition a1,··· ,ak at the coordinator such that for every i ∈[1,k], v...

Evidence 2 - **Rationale**: Both papers compare their blackboard model algorithms to Chen et al. 2016 as prior work, and both claim improvements in communication complexity. The candidate paper demonstrates that communication-optimal protocols for the blackboard model improving upon Chen et al. 2016 existed prior to the original paper's contribution. - **Original**: by comparison, the state-of-the-art protocol achieves a constant factor approximation using ˜o (s + dk) log2(nΔ) total communication (chen et al., 2016). thus compared to the work of chen et al. (2016), not only do we achieve a (1 + ε)-coreset construction in the blackboard setting, but also we r... - **Candidate**: d2-cabl can also work in the distributed static setting by considering that there is only one time point, at which all graph information comes together. as mentioned earlier, it is a brand new algorithm with nearly-optimal communication complexity, the same as the state-of-the-art algorithm (chen et...

2. Communication-Efficient Distributed Graph Clustering and Sparsification Under Duplication Models

URL: View paper

Brief Assessment

Graph Clustering Duplication[38] focuses on graph clustering and sparsification problems in distributed settings, not on Euclidean clustering with coresets. The technical approaches and problem domains are fundamentally different.

3. Communication-Efficient Algorithms for Distributed Computation and Machine Learning

URL: View paper

Brief Assessment

Communication-Efficient Distributed[39] focuses on general distributed computation and machine learning algorithms, not specifically on distributed clustering protocols in the blackboard model with the approximation guarantees claimed by the original paper.

Contribution 3: Lazy adaptive sampling and coordinate-wise sampling techniques

Description: The authors introduce lazy adaptive sampling where sites update the blackboard only when local weight estimates change significantly, combined with coordinate-wise sampling that decomposes points along coordinates and samples dimensions based on significance. These techniques enable compact summaries and efficient distributed protocols without transmitting full high-dimensional centers.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Distributed Algorithms for Euclidean Clustering View paper
- [1] Fedcore: Straggler-free federated learning with distributed coresets View paper

- [2] Coreset for line-sets clustering View paper
- [3] Distributed -means and -median Clustering on General Topologies View paper
- [4] Scalable k-means clustering via lightweight coresets View paper
- [5] Deterministic Coresets for k-Means of Big Sparse Data View paper
- [6] Distributed k-means with outliers in general metrics View paper
- [7] Fair Coresets and Streaming Algorithms for Fair k-means View paper
- [8] Distributed balanced clustering via mapping coresets View paper
- [9] Coresets for Vertical Federated Learning: Regularized Linear Regression and -Means Clustering View paper
- [10] Training gaussian mixture models at scale via coresets View paper
- [11] Scalable decision fusion algorithm for enabling decentralized computation in distributed, big data clustering problems View paper
- [12] Towards faster big data analytics for antiâ□□jamming applications in vehicular adâ□□hoc network View paper
- [13] Unsupervised Clustering for Condition Monitoring of Electric Drives View paper
- [14] Distributed PCA and k-means clustering View paper
- [15] A composable coreset for k-center in doubling metrics View paper
- [16] Accurate MapReduce Algorithms for -median and -means in General Metric Spaces View paper
- [17] Joint Coreset Construction and Quantization for Distributed Machine Learning View paper
- [18] Scalable training of mixture models via coresets View paper
- [19] Coreset-sharing based Collaborative Model Training among Peer Vehicles View paper
- [20] Distributed Bayesian Coresets View paper
- [21] Robust Coreset Construction for Distributed Machine Learning View paper
- [22] Coreset-Based Data Reduction for Machine Learning at the Edge View paper
- [23] Scalable and Distributed Clustering via Lightweight Coresets View paper
- [24] Coresets for Clustering in Graphs of Bounded Treewidth View paper
- [25] Training mixture models at scale via coresets View paper
- [26] Machine Learning at the Edge View paper
- [27] Anomaly detection in vertically partitioned data by distributed core vector machines View paper
- [28] Coresets for Vertical Federated Learning: Regularized Linear Regression and K-Means Clustering View paper
- [29] An effective coreset compression algorithm for large scale sensor networks View paper
- [30] Efficient Implementation of Coreset-based K-Means Methods View paper
- [31] FedCode: Communication-Efficient Federated Learning via Transferring Codebooks View paper
- [32] k-center clustering with outliers in the MPC and streaming model View paper
- [33] Gradient coreset for federated learning View paper
- [34] CBFL: A Communication-Efficient Federated Learning Framework From Data Redundancy Perspective View paper
- [35] Randomized greedy algorithms and composable coreset for k-center clustering with outliers View paper
- [36] Robustð□□□-Center Clustering for Contin-uous Monitoring View paper
- [37] Communication-optimal distributed dynamic graph clustering View paper
- [38] Communication-Efficient Distributed Graph Clustering and Sparsification Under Duplication Models View paper
- [39] Communication-Efficient Algorithms for Distributed Computation and Machine Learning View paper