CORESETS FOR EUCLIDEAN CLUSTERING PROBLEMS WITH CONCENTRATED GEO-PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Coresets are compact summaries of datasets such that models trained on a coreset can (provably) approximate those trained on the entire dataset. As such, they have been widely used to scale up clustering problems to massive data. In many real-world scenarios, clustering involves sensitive data, making the release of coresets subject to privacy concerns. Existing differentially private (DP) methods for coreset construction focus predominantly on the centralized model, while extending them to the local model of DP is still largely unexplored. In this paper, we study how to construct coresets for Euclidean clustering problems with concentrated geo-privacy (CGP), a local DP model that quantifies distinguishability based on the distance between data points as defined by a distance function (i.e., the Euclidean distance in our formulation). We first propose an algorithm to build coresets with CGP for the 1-center problem (a.k.a. minimum enclosing ball, MEB) with a $(1+\gamma)$ relative approximation and an additive error. We then extend the scheme to the k-center problem with any $k = O(1)$. Furthermore, we propose constructing approximate coresets with CGP for general (k, z)-clustering (of which k-median and k-means are special cases when $z = 1$ and $z = 2$, respectively). We implement our algorithms and conduct extensive experiments to demonstrate their effectiveness in practice.

1 INTRODUCTION

Clustering is a fundamental unsupervised machine learning problem. The (k, z)-clustering (Huang et al., 2024) is a common class of clustering problems that finds a set of k points as centers and partitions the data into k clusters by assigning each point to its nearest center, in order to minimize an ℓ_z-cost function defined on the distances between centers and data points. Classic formulations of (k, z)-clustering include k-center ($z = \infty$), k-median ($z = 1$), and k-means ($z = 2$), which correspond to minimizing the maximum distance, the sum of distances, and the sum of squared distances, respectively. Since these problems are generally NP-hard (Megiddo & Supowit, 1984; Aloise et al., 2009), extensive efforts have been devoted to designing approximation algorithms for them (Gonzalez, 1985; Arthur & Vassilvitskii, 2007).

Among the approximation approaches, coreset-based methods (Jubran et al., 2021) have become a standard paradigm for scalable (k, z)-clustering. These methods construct substantially smaller summaries of datasets, called *coresets*, such that clustering costs evaluated on the coresets approximate those on the original datasets with provable guarantees. For the 1-center (minimum enclosing ball, MEB) and k-center problems, dimension-independent coresets have been established (Badoiu et al., 2002; Badoiu & Clarkson, 2003; 2008; Larsson & Källberg, 2013). For general (k, z)-clustering problems, importance sampling-based weighted coresets have also become a standard approach (Feldman & Langberg, 2011; Bachem et al., 2017; 2018; Huang & Vishnoi, 2020; Schwiegelshohn & Sheikh-Omar, 2022).

In many real-world scenarios, clustering should be performed on datasets with sensitive information. Releasing either cluster centers or coresets directly may reveal information about individual records. Differential privacy (DP) (Dwork et al., 2006) has therefore become a standard framework for privacy-preserving data analysis. Accordingly, private coreset construction and clustering algorithms have been extensively studied in the centralized DP model (Feldman et al., 2009; 2017; Balcan et al., 2017; Jones et al., 2021; Cohen-Addad et al., 2022), where a trusted curator has access

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to the original dataset and perturbs the output so that no single record is distinguishable. Trusted curators, however, may not be available in practice; thus, a local DP (LDP) model that requires each user to perturb their own input before sending it to the curator, so that it is indistinguishable from any other possible input, is needed (Kasiviswanathan et al., 2011). Nevertheless, coreset construction in the LDP model is substantially more challenging because the geometric structure of the raw data may be overwhelmed by noise. There are also a few attempts at k-means or MEB problems with LDP (Nissim & Stemmer, 2018; Chang et al., 2021; Stemmer, 2021; Chaturvedi et al., 2022; Mahpud & Sheffet, 2022). But these methods can achieve only weaker theoretical bounds and empirical performance than methods in the centralized DP model.

To address the limitations of LDP, metric differential privacy (MDP) (Chatzikokolakis et al., 2013), an LDP model whose indistinguishability level is tied to a distance metric, has been considered. In particular, MDP makes nearby inputs less distinguishable than distant ones, thus preserving each record's exact information while better maintaining the dataset's geometric structure. Recent work has considered MDP for k-means in (Yang et al., 2022); however, it follows a distinct approach from one based on coresets and does not extend to other (k, z)-clustering problems.

Our Results In this paper, we investigate Euclidean (k, z)-clustering problems with concentrated geo-privacy (CGP) (Liang & Yi, 2023; 2024), a variant of MDP that provides an advanced composition property to allow the combination of geo-private mechanisms with bounded privacy loss.

We first propose an algorithm to build a coreset for 1-center with ρ-CGP. The algorithm draws on the non-private iterative farthest-point search method in (Badoiu & Clarkson, 2008; Larsson & Källberg, 2013) and utilizes noisy distance queries and point releases to guarantee ρ-CGP. For a dataset $P \subseteq \mathbb{R}^d$ of n points, it provides a coreset of size $O(1/\gamma)$, which, with probability at least $1-\beta$ for any $\beta \in (0,1)$, achieves a $(1+\gamma)$-approximation with an additive error of $\widetilde{O}((\sqrt{d}+\sqrt{\log(n/\gamma)})/\sqrt{\gamma})$ for 1-center.[1] We then extend the framework to k-center with $k = O(1)$ by adapting the farthest-first traversal procedure of (Gonzalez, 1985) to be ρ-CGP. This leads to the same approximation factor of 2 as (Gonzalez, 1985), along with an additive error of $\widetilde{O}(\sqrt{k}+\sqrt{d})$. With post-processing, we further build a coreset of size $O(k/\gamma)$ for k-center with an improved approximation factor of $1+\gamma$ and an additive error of $\widetilde{O}(2^{O((k\log k)/\gamma^2)}+\sqrt{d})$.

Furthermore, we propose an algorithm to build coresets for general (k, z)-clustering with $z \geq 1$ by incorporating CGP mechanisms into sensitivity sampling (Feldman & Langberg, 2011). The algorithm provides a weighted coreset of size $\widetilde{O}(k\eta^{-(2z+2)})$ with a $(1+O(\eta))$-approximation and an additive error of $\widetilde{O}(kN_{\max}(\eta(\sqrt{k}+\sqrt{d})^z+(\sqrt{d})^z/\eta^{z-1}))$, where $N_{\max}$ is the maximum weight of any single point in the coreset. For the canonical cases, setting $z=1$ results in an $\widetilde{O}(k\eta^{-4})$-size $(1+O(\eta))$-approximation coreset with an additive error of $\widetilde{O}(kN_{\max}(\eta\sqrt{k}+\sqrt{d}))$ for Euclidean k-median; and setting $z=2$ leads to an $\widetilde{O}(k\eta^{-6})$-size $(1+O(\eta))$-approximation coreset with an additive error of $\widetilde{O}(kN_{\max}(\eta k + d/\eta))$ for Euclidean k-means.

Finally, we implement the proposed algorithms and evaluate them on four synthetic and real-world datasets. The results show that the constructed coresets effectively approximate MEB, k-center, and k-means under a mild privacy budget. Moreover, our k-means algorithm significantly outperforms that in (Yang et al., 2022) in clustering quality under the same privacy budget.

2 PRELIMINARIES

Euclidean Clustering Problems Let $P = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ be a set of n points in the d-dimensional Euclidean space. For a set $C \subseteq \mathbb{R}^d$ of (center) points, define $d(x, C) = \min_{c\in C} \|x - c\|_2$. For a center $c \in \mathbb{R}^d$ and radius $r \geq 0$, a ball is defined as $B(c, r) = \{x \in \mathbb{R}^d : \|x - c\|_2 \leq r\}$.

The Euclidean k-center problem seeks a set $C \subseteq \mathbb{R}^d$ of at most k centers that minimize the maximum distance from any point $p \in P$ to its nearest center in C. We write it as

$$\mathrm{OPT}_k(P) = \min_{C\subseteq\mathbb{R}^d:|C|\leq k} \max_{x\in P} d(x, C). \tag{1}$$

[1] The $\widetilde{O}(\cdot)$ notation suppresses the dependency on β, ρ, and their polylogarithmic factors.

The special case $k = 1$, i.e., Euclidean 1-center, is also called the minimum enclosing ball (MEB) problem, where $\mathrm{OPT}_1(P)$ is the radius of the smallest ball containing all points in P. In particular, let $c^\star$ denote the optimal center. $B(c^\star, \mathrm{OPT}_1(P))$ is thus the MEB of P. A ball $B(c, r)$ is $(1+\gamma)$-approximate for MEB if $P \subseteq B(c, r)$ and $r \leq (1+\gamma)\mathrm{OPT}_1(P)$. There have been coreset-based approximation schemes for Euclidean 1-center and k-center, e.g., (Badoiu et al., 2002; Badoiu & Clarkson, 2003; Kumar et al., 2003; Badoiu & Clarkson, 2008). A formal definition of *coresets* for k-center is given as follows:

Definition 2.1 (Coreset for k-Center). *For $\gamma \in (0, 1)$, a set $S \subseteq \mathbb{R}^d$ is a γ-coreset for k-center on $P \subseteq \mathbb{R}^d$ if the optimal solution C_S to k-center on S satisfies $\max_{x \in P} d(x, C_S) \leq (1+\gamma)\mathrm{OPT}_k(P)$. For $k = 1$, this corresponds to the coreset definition for 1-center (i.e., MEB).*

For any $z \geq 1$ and a center set $C \subseteq \mathbb{R}^d$ with $|C| \leq k$, we define the Euclidean (k, z)-clustering cost and the corresponding optimal (k, z)-clustering solution as

$$\mathrm{cost}_z(P, C) = \sum_{x \in P} d(x, C)^z, \quad \mathrm{OPT}_{k,z}(P) = \min_{C \subseteq \mathbb{R}^d : |C| \leq k} \mathrm{cost}_z(P, C). \tag{2}$$

The cases $z = 1$ and $z = 2$ correspond to k-median and k-means, respectively. Note that k-center can also be considered as (k, z)-clustering with $z = \infty$. For a weighted point set (S, w), where $S \subseteq \mathbb{R}^d$ and $w : S \to \mathbb{R}_{\geq 0}$, define

$$\mathrm{cost}_z(S, w, C) = \sum_{s \in S} w(s) d(s, C)^z. \tag{3}$$

There have been several methods to build the above-defined weighted coresets via importance and sensitivity sampling for (k, z)-clustering, e.g., (Har-Peled & Mazumdar, 2004; Feldman & Langberg, 2011; Huang & Vishnoi, 2020). A formal definition is given as follows:

Definition 2.2 (Coreset for (k, z)-Clustering). *For $\eta \in (0, 1)$, a weighted set (S, w) is an η-coreset for (k, z)-clustering if, for every $C \subseteq \mathbb{R}^d$ with $|C| \leq k$,*

$$|\mathrm{cost}_z(S, w, C) - \mathrm{cost}_z(P, C)| \leq \eta\, \mathrm{cost}_z(P, C). \tag{4}$$

Concentrated Geo-Privacy Local differential privacy (LDP) perturbs each user's record before sending it to an untrusted curator. A local randomizer $R : \mathcal{X} \to \mathcal{Y}$ is (ε, δ)-LDP if, for all $x, x' \in \mathcal{X}$ and measurable $A \subseteq \mathcal{Y}$, $\Pr[R(x) \in A] \leq e^\varepsilon \Pr[R(x') \in A] + \delta$ (Dwork et al., 2006; Kasiviswanathan et al., 2011). Metric differential privacy (MDP) replaces the neighboring relation in LDP with a metric on the input domain (Chatzikokolakis et al., 2013). Rather than treating all input pairs equally, it scales the privacy loss by their distance under the metric, i.e.,

Definition 2.3 (MDP). *Let $(\mathcal{X}, d)$ be a metric space. A mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ satisfies $(\varepsilon, \delta, \Delta)$-MDP if, for all $x, x' \in \mathcal{X}$ with $d(x, x') \leq \Delta$ and all measurable $A \subseteq \mathcal{Y}$, $\Pr[\mathcal{M}(x) \in A] \leq e^{\varepsilon \cdot d(x, x')} \Pr[\mathcal{M}(x') \in A] + \delta$.*

For Euclidean clustering, we consider $d(\cdot, \cdot)$ to be the Euclidean distance, i.e., $d(x, x') = \|x - x'\|_2$.

For metric spaces $(\mathcal{X}, d)$ and $(\mathcal{Y}, d_\mathcal{Y})$, a query $g : \mathcal{X} \to \mathcal{Y}$ is L-Lipschitz if $d_\mathcal{Y}(g(x), g(x')) \leq L d(x, x')$ for all $x, x' \in \mathcal{X}$. Let $D_\alpha(P \| Q)$ denote the Rényi divergence of order $\alpha > 1$ (van Erven & Harremoës, 2014). Concentrated geo-privacy (CGP) defines privacy through Rényi divergence with respect to the metric distance between inputs (Liang & Yi, 2023), i.e.,

Definition 2.4 (CGP). *Let $\rho \geq 0$. A mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ satisfies ρ-CGP w.r.t. $d(\cdot, \cdot)$ if, for all $x, x' \in \mathcal{X}$ and all $\alpha > 1$, $D_\alpha(\mathcal{M}(x) \,\|\, \mathcal{M}(x')) \leq \alpha\rho\, d(x, x')^2$.*

For an L-Lipschitz query with Euclidean output, the following Gaussian mechanism satisfies CGP.

Theorem 2.5 (Gaussian Mechanism). *Let $g : \mathcal{X} \to \mathbb{R}^d$ be L-Lipschitz. For $\rho > 0$, the mechanism $G(x) = g(x) + \frac{L}{\sqrt{2\rho}} Z$, where $Z \sim \mathcal{N}(0, I_d)$, satisfies ρ-CGP.*

An important property of CGP is that it satisfies the following adaptive composition.

Theorem 2.6 (Adaptive Composition). *Let $M_1 : \mathcal{X} \to \mathcal{Y}_1$ satisfy ρ_1-CGP and $M_2 : \mathcal{X} \times \mathcal{Y}_1 \to \mathcal{Y}_2$ satisfy ρ_2-CGP. Then, the mechanism $M(x) = (M_1(x), M_2(x, M_1(x)))$ satisfies $(\rho_1 + \rho_2)$-CGP. By induction, a T-fold adaptive composition with parameters $\rho_1, \ldots, \rho_T$ satisfies $\sum_{t=1}^T \rho_t$-CGP.*

Algorithm 1 MEBREFINE

Input: Proxy set $S \subseteq \mathbb{R}^d$, center θ, radius r, refinement accuracy $\alpha \in (0, 1)$.
Output: Refined ball $B(\theta, r)$.
while true **do**
  Choose $q \leftarrow \arg\max_{u \in S} \|u - \theta\|_2$ and compute $h \leftarrow \|q - \theta\|_2$.
  if $h \leq (1 + \alpha)r$ **then return** (θ, r).
  Update $r^+ \leftarrow (r^2/h + h)/2$, $\theta^+ \leftarrow q + (r^+/h)(\theta - q)$, and $(\theta, r) \leftarrow (\theta^+, r^+)$.

3 CORESETS FOR MEB AND K-CENTER WITH CGP

We first propose an interactive coreset construction for MEB under CGP and then apply the same framework to k-center. Both objectives are defined by worst-case distances and therefore admit a common farthest-point view: users answer noisy distance queries, while noisy vectors are released to the curator only for selected geometric witnesses and cached for reuse. The coreset construction for MEB adapts the iterative farthest-point paradigm that grounds the coreset-based approximation (Badoiu & Clarkson, 2003; 2008; Larsson & Källberg, 2013). For k-center, we adapt Gonzalez's farthest-first traversal and then provide an exhaustive coreset refinement for better approximation. Note that all proofs are deferred to Appendix B due to space limitations.

3.1 MINIMUM ENCLOSING BALL

Interactive Farthest-Point Construction Our construction is built on two primitives. The first is *private distance query*: for a point or center set determined by the public transcript, every user reports a Gaussian-perturbed scalar distance. These reports identify geometric witnesses and, for MEB, certify the coverage of the original dataset. The second is *private point release*: only a selected user releases a Gaussian-perturbed vector, which is stored by the curator and will be reused whenever needed. All subsequent procedures are curator-side post-processing.

For MEB, fix $\gamma, \beta \in (0, 1)$ and $\rho > 0$, and set $K = \lceil 2/\gamma \rceil$. The construction executes at most $K + 2$ scalar distance rounds, so each scalar report uses Gaussian variance $\sigma_d^2 = (K + 2)/\rho$. A uniform budget assignment over all scalar reports gives $\tau_d = \sqrt{2(K + 2)\log(4n(K + 2)/\beta)}/\sqrt{\rho}$. Each selected user releases at most one vector with noise $\mathcal{N}(0, I_d/\rho)$; Gaussian norm concentration over the at most $K+2$ released witnesses gives $\tau_x = (\sqrt{d} + \sqrt{2\log(4(K + 2)/\beta)})/\sqrt{\rho}$. The scalar primitive returns both a noisy maximizer and the certificate $U(\theta) = \max\{0, M(\theta) + \tau_d\}$, whereas the released proxies are refined by Algorithm 1. Their composition gives Algorithm 2.

Theoretical Analysis First, every call to Algorithm 1 returns a well-defined refined $B(\theta, r)$.

Lemma 3.1. *For any finite proxy set S, Algorithm 1 terminates after a finite number of steps.*

The scalar distance transcript and the cached vector release are calibrated so that their adaptive composition remains within the prescribed CGP budget.

Lemma 3.2. *For any $\rho > 0$, Algorithm 2 satisfies ρ-CGP.*

Next, we analyze the approximation bound of Algorithm 2. The analysis conditions on a simultaneous concentration event for all scalar and vector perturbations. Let $c^\star$ be an optimal 1-center of P, and define $h(\theta) = \max_{i \in [n]} \|x_i - \theta\|_2$.

Lemma 3.3. *Let $\mathcal{E}$ denote the event that every executed scalar report satisfies $|Z_{i,t}| \leq \tau_d$ and every released proxy satisfies $\|\widetilde{x}_j - x_j\|_2 \leq \tau_x$. Then, $\Pr(\mathcal{E}) \geq 1 - \beta$.*

Condition on $\mathcal{E}$ and define $\tau_{\text{fp}} = 2\tau_d + \tau_x$. The next lemma transfers the noisy maximization step to the underlying farthest-point geometry. It simultaneously controls the selected proxy and the covering certificate derived from the same scalar transcript.

Lemma 3.4. *Consider any distance query at center θ, and let j and M denote its noisy maximizer and noisy maximum. Then, $h(\theta) - \tau_d \leq M \leq h(\theta) + \tau_d$. Moreover, $\|\widetilde{x}_j - \theta\|_2 \geq h(\theta) - \tau_{\text{fp}}$, and $h(\theta) \leq U(\theta)$. Thus, $B(\theta, U(\theta))$ contains P.*

Lemma 3.4 provides a per-round approximation to the true farthest witness. To control the full refinement trajectory, set $R = \text{OPT}_1(P) + \tau_x$. On $\mathcal{E}$, every released proxy lies in $B(c^\star, R)$. The

Algorithm 2 CGP-MEB-CORESET

Input: Dataset $P \subseteq \mathbb{R}^d$, accuracy parameter γ, privacy parameter ρ, failure probability β.
Output: Coreset $\widehat{S}$ and ball $B(\widehat{\theta}, \widehat{r})$.
Set $K \leftarrow \lceil 2/\gamma \rceil$, $\sigma_d^2 \leftarrow (K+2)/\rho$, and τ_d, τ_x accordingly.
Initialize $\mathcal{M} \leftarrow \emptyset$ and pick a public anchor $a \in \mathbb{R}^d$.
for $i = 1, \ldots, n$ **in parallel do**
  User i sends $Y_i(a) \leftarrow \|x_i - a\|_2 + Z_i$ with $Z_i \sim \mathcal{N}(0, \sigma_d^2)$ to the curator.
Let $j_a \leftarrow \arg\max_{i \in [n]} Y_i(a)$.
if $j_a \notin \mathcal{M}$ **then**
  User j_a releases $\widetilde{q}_a \leftarrow x_{j_a} + \xi_{j_a}$ with $\xi_{j_a} \sim \mathcal{N}(0, I_d/\rho)$ to the curator and store $\mathcal{M}(j_a) \leftarrow \widetilde{q}_a$.
else
  Set $\widetilde{q}_a \leftarrow \mathcal{M}(j_a)$.
for $i = 1, \ldots, n$ **in parallel do**
  User i sends $Y_i(\widetilde{q}_a) \leftarrow \|x_i - \widetilde{q}_a\|_2 + Z_i$ with $Z_i \sim \mathcal{N}(0, \sigma_d^2)$ to the curator.
Let $j_b \leftarrow \arg\max_{i \in [n]} Y_i(\widetilde{q}_a)$.
if $j_b \notin \mathcal{M}$ **then**
  User j_b releases $\widetilde{q}_b \leftarrow x_{j_b} + \xi_{j_b}$ with $\xi_{j_b} \sim \mathcal{N}(0, I_d/\rho)$ to the curator and store $\mathcal{M}(j_b) \leftarrow \widetilde{q}_b$.
else
  Set $\widetilde{q}_b \leftarrow \mathcal{M}(j_b)$.
Set $S_0 \leftarrow \{\widetilde{q}_a, \widetilde{q}_b\}$, $\theta_0 \leftarrow (\widetilde{q}_a + \widetilde{q}_b)/2$, and $r_0 \leftarrow \|\widetilde{q}_a - \widetilde{q}_b\|_2/2$.
Set $(\widehat{\theta}, \widehat{r}) \leftarrow (\theta_0, +\infty)$.
for $t = 0, \ldots, K-1$ **do**
  for $i = 1, \ldots, n$ **in parallel do**
    User i sends $Y_{i,t} \leftarrow \|x_i - \theta_t\|_2 + Z_{i,t}$ with $Z_{i,t} \sim \mathcal{N}(0, \sigma_d^2)$ to the curator.
  Let $j_t \leftarrow \arg\max_{i \in [n]} Y_{i,t}$.
  Set $M_t \leftarrow \max_{i \in [n]} Y_{i,t}$ and $U_t \leftarrow \max\{0, M_t + \tau_d\}$.
  if $U_t < \widehat{r}$ **then**
    Set $(\widehat{\theta}, \widehat{r}) \leftarrow (\theta_t, U_t)$.
  if $U_t \leq (1+\gamma) r_t$ **then**
    return $(S_t, B(\theta_t, U_t))$.
  if $j_t \notin \mathcal{M}$ **then**
    User j_t releases $\widetilde{q}_t \leftarrow x_{j_t} + \xi_{j_t}$ with $\xi_{j_t} \sim \mathcal{N}(0, I_d/\rho)$ to the curator and store $\mathcal{M}(j_t) \leftarrow \widetilde{q}_t$.
  else
    Set $\widetilde{q}_t \leftarrow \mathcal{M}(j_t)$.
  Set $S_{t+1} \leftarrow S_t \cup \{\widetilde{q}_t\}$.
  Set $(\theta_{t+1}, r_{t+1}) \leftarrow \text{MEBREFINE}(S_{t+1}, \theta_t, r_t, \gamma/2)$.
return $(S_K, B(\widehat{\theta}, \widehat{r}))$.

following invariant shows that every state produced by the deterministic refinement remains feasible with respect to this enlarged optimal ball.

Lemma 3.5. *Every internal state (θ, r) generated by the initialization or by Algorithm 1 satisfies $\|\theta - c^\star\|_2^2 + r^2 \leq R^2$. In particular, every outer state satisfies $r_t \leq R$.*

The viability invariant bounds the refinement radius globally. The next lemma complements this bound with a progress condition: whenever the current center remains above the target scale, the approximate farthest witness forces a strict increase in the refinement radius.

Lemma 3.6. *Define $R_\gamma = (1+\gamma)R$. If an outer-round center satisfies $h(\theta_t) > R_\gamma + \tau_{\text{fp}}$, then the refinement in that round satisfies $r_{t+1} > (r_t^2/R_\gamma + R_\gamma)/2$. Consequently, among the queried centers $\theta_0, \ldots, \theta_{K-1}$, there exists an index $t^\star$ satisfying $h(\theta_{t^\star}) \leq (1+\gamma)R + \tau_{\text{fp}}$.*

Lemma 3.6 guarantees the existence of a queried center with a near-optimal true farthest distance, while Lemma 3.4 converts its scalar transcript into a valid covering certificate. Since Algorithm 2 returns the smallest certificate encountered during the execution, these yield the utility guarantee.

Theorem 3.7. *For any $\gamma, \beta \in (0,1)$ and $\rho > 0$, with probability at least $1-\beta$, Algorithm 2 returns a proxy set S and a certified ball $B(\widehat{\theta}, \widehat{r})$ such that $P \subseteq B(\widehat{\theta}, \widehat{r})$ and*

$$\widehat{r} \leq (1+\gamma)\text{OPT}_1(P) + O\left(\frac{\sqrt{d} + \sqrt{\log(1/(\gamma\beta))}}{\sqrt{\rho}} + \sqrt{\frac{\log(n/(\gamma\beta))}{\gamma\rho}}\right). \tag{5}$$

More precisely, $\widehat{r} \le (1+\gamma)\mathrm{OPT}_1(P) + (2+\gamma)\tau_x + 4\tau_d$. Moreover, $|S| \le K+2 = O(1/\gamma)$, the algorithm uses at most $K+2 = O(1/\gamma)$ scalar distance-report rounds, and at most $K+2 = O(1/\gamma)$ distinct privatized vectors are released.

3.2 K-CENTER

We next extend the same two local primitives to Euclidean k-center. Recall that $\mathrm{OPT}_k(P)$ denotes the optimal k-center radius. We first introduce CGP-GMM, a private analog of the farthest-first traversal of Gonzalez (1985), which preserves its factor-2 approximation up to an additive privacy error. We then combine the same private farthest-point interface with the dimension-independent coreset construction of Badoiu et al. (2002) to obtain a $(1+\gamma)$ multiplicative guarantee.

Private Gonzalez's Traversal CGP-GMM directly reuses the distance-query and cached witness-release primitives from the MEB construction. A uniformly sampled user serves only as an initialization anchor. The distance-query primitive selects the noisy farthest user from the privatized anchor as the first center and subsequently selects the noisy farthest user from the current private center set. Each selected center is represented through the cached witness-release primitive, so its privatized vector is released at most once.

The algorithm uses exactly k scalar rounds. Accordingly, setting $T_c = k$ and using scalar noise variance $\sigma_c^2 = T_c/\rho$ calibrates the scalar transcript within the CGP budget, while the cached vector release uses the same Gaussian calibration as in the MEB construction. Uniform concentration gives $\tau_c = \sqrt{2T_c\rho^{-1}\log(4nT_c/\beta)}$ for the scalar reports and $\tau_x = (\sqrt{d} + \sqrt{2\log(4(k+1)/\beta)})/\sqrt{\rho}$ for the released vectors. The pseudocode of CGP-GMM is deferred to Appendix E.

The packing argument underlying the farthest-first traversal of Gonzalez (1985) remains valid after accounting for the perturbation in witness selection and center release. This yields the following guarantee for CGP-GMM.

Theorem 3.8. *For any $\rho > 0$ and $\beta \in (0,1)$,* CGP-GMM *satisfies ρ-CGP and, with probability at least $1-\beta$, returns a center set $\widehat{C}$ with $|\widehat{C}| \le k$ such that*

$$\max_{x\in P} d(x,\widehat{C}) \le 2\mathrm{OPT}_k(P) + \widetilde{O}\left(\sqrt{\frac{k}{\rho}} + \sqrt{\frac{d}{\rho}}\right). \tag{6}$$

Exhaustive Coreset Refinement The factor 2 in Theorem 3.8 is inherited from the farthest-first traversal of Gonzalez (1985). To obtain a $(1+\gamma)$ multiplicative guarantee, we combine the same private distance-query and cached witness-release primitives with the dimension-independent Euclidean k-center coreset construction of Badoiu et al. (2002). Fix an optimal k-center solution and let $P_1^\star, \dots, P_k^\star$ denote its induced clusters. If the optimal-cluster identity of each selected witness were known, the coreset refinement could be applied independently within every $P_\ell^\star$. Since these identities are unavailable, the refinement enumerates all possible assignments of each selected witness to the k clusters. The resulting search tree contains an oracle branch in which every witness is assigned to its corresponding optimal cluster. Along this oracle branch, the classical Euclidean k-center construction requires $O(1/\gamma^2)$ witnesses per optimal cluster to obtain a $(1+\gamma)$ approximation (Badoiu et al., 2002). Hence, the total number of witnesses is $O(k/\gamma^2)$, and the branch depth is $H = \Theta(k/\gamma^2)$. Enumerating all witness-to-cluster assignments induces $Q_{\mathrm{ref}} = \sum_{h=0}^{H} k^h = 2^{O((k\log k)/\gamma^2)}$ branch-specific violation queries. Each branch-specific violation query is 1-Lipschitz, conditioned on the current transcript. Calibrating the scalar Gaussian mechanisms over the complete search tree therefore gives high-probability error $\tau_{\mathrm{ref}} = \widetilde{O}(\sqrt{Q_{\mathrm{ref}}/\rho})$. The cached witness-release primitive contributes $\widetilde{O}(\sqrt{d/\rho})$, since each user's vector is released at most once and subsequently reused throughout the search tree. All branch enumeration and center computation are post-processing. The refinement procedure is deferred to Appendix E. Along the oracle branch, the non-private coreset refinement applies independently to the optimal clusters. The private interaction affects the construction only through the approximation of the selected witnesses and the perturbation of their released locations. Combining these errors with the $(1+\gamma)$ coreset guarantee yields the following result.

Theorem 3.9. *For any $\gamma, \beta \in (0,1)$ and $\rho > 0$, the exhaustive k-center refinement satisfies ρ-CGP and, with probability at least $1-\beta$, returns a proxy set S_{ref} of size $O(k/\gamma^2)$ and a center set*

$C_{\mathrm{ref}} \subseteq \mathbb{R}^d$ *with* $|C_{\mathrm{ref}}| \le k$ *such that*

$$\max_{x \in P} d(x, C_{\mathrm{ref}}) \le (1+\gamma)\mathrm{OPT}_k(P) + \widetilde{O}\left(\sqrt{\frac{2^{O((k\log k)/\gamma^2)}}{\rho}} + \sqrt{\frac{d}{\rho}}\right). \tag{7}$$

4 CORESETS FOR (K,Z)-CLUSTERING WITH CGP

We extend our construction to the (k, z)-clustering objectives defined in Section 2, where $z \ge 1$ is fixed and all distances are Euclidean. The cases $z = 1$ and $z = 2$ recover Euclidean k-median and k-means, respectively. Our construction combines a private first-stage sensitivity sampler with the two-stage importance-sampling framework of Huang & Vishnoi (2020). The first stage interacts with the original users, whereas the second stage operates only on its private weighted output and is therefore post-processing. This yields a final coreset of size $\widetilde{O}_z(k\eta^{-(2z+2)})$ without additional privacy cost.

Private Construction Let $b = \min\{n, \lceil c_z k \rceil\} = O_z(k)$ denote the number of bicriteria seeds. The construction reuses the private distance-query and cached witness-release primitives developed above. A uniformly sampled user serves only as an initialization anchor. The first seed is selected by noisy D^z-sampling from the privatized anchor, the second seed is obtained by noisy farthest-point selection from the first seed, and the remaining seeds are generated by noisy D^z-sampling with respect to the current private seed set. All selected users share the same release cache, so each user releases at most one Gaussian-perturbed vector.

Seed construction and private assignment each use at most b scalar rounds, giving $T_z = 2b$. We therefore use scalar Gaussian variance $\sigma_z^2 = T_z/\rho$. Uniform concentration over the scalar transcript gives $\tau_{\mathrm{sc}} = \sqrt{2T_z\rho^{-1}\log(4nT_z/\beta)}$, while the cached vector releases satisfy $\tau_x = (\sqrt{d} + \sqrt{2\log(4n/\beta)})/\sqrt{\rho}$ with high probability.

After constructing the private seed set $A = \{a_1, \dots, a_b\}$, users privately report their distances to the seeds. Let $\widehat{a}(i)$ denote the resulting private assignment, let $\overline{D}_i = \max\{0, Y^{\mathrm{asg}}_{i,\widehat{a}(i)} + \tau_{\mathrm{sc}}\}$, and let N_ℓ be the number of users assigned to a_ℓ. Writing $\widehat{\Phi} = \sum_{i=1}^n \overline{D}_i^z$, we define the private analog of the first-stage importance score of Huang & Vishnoi (2020) by $s_i^{(z)} = \kappa_z\left(\frac{\overline{D}_i^z}{\widehat{\Phi}} + \frac{1}{N_{\widehat{a}(i)}}\right)$ and $\kappa_z = 2^{2z+2}\alpha_z^2$, when $\widehat{\Phi} > 0$, and set $s_i^{(z)} = \kappa_z/N_{\widehat{a}(i)}$ otherwise. The first-stage weighted sample is drawn with probabilities proportional to $s_i^{(z)}$. We then apply the second-stage reduction of Huang & Vishnoi (2020) entirely to this private weighted output. The pseudocode is deferred to Appendix E.

The privacy guarantee follows from the same composition principle used for MEB and k-center. Scalar interactions account for the adaptive distance transcript, while the shared cache ensures that each user incurs the vector-release cost at most once. The second-stage reduction is post-processing.

Lemma 4.1. *For any $\rho > 0$, the private (k, z) coreset construction satisfies ρ-CGP.*

For utility, the main step is to show that the private scores retain the sensitivity domination required by importance sampling, up to an additive term induced by the privatized bicriteria solution. Let $N_{\max}$ denote the maximum weight of a single point in the resulting weighted coreset.

Lemma 4.2. *With probability at least $1-\beta$, $\sum_{i=1}^n s_i^{(z)} = O_z(k)$ and, for every $C \subseteq \mathbb{R}^d$ with $|C| \le k$, $\frac{d(x_i,C)^z}{\mathrm{cost}_z(P,C)+A_z} \le s_i^{(z)}$ and $A_z = \widetilde{O}_z\left(kN_{\max}(\tau_{\mathrm{sc}} + \tau_x)^z\right)$.*

Lemma 4.2 allows the first-stage private sample to satisfy the same importance-sampling argument as in the non-private construction, with A_z accounting for the perturbation introduced before sampling. Applying the second-stage reduction of Huang & Vishnoi (2020) to this private weighted sample preserves the multiplicative approximation while reducing the support size.

Theorem 4.3. *For any fixed $z \ge 1$, $\eta, \beta \in (0,1)$, and $\rho > 0$, with probability at least $1-\beta$, outputs a weighted private coreset (S_z, w_z) of size $\widetilde{O}_z(k\eta^{-(2z+2)})$ such that, for every $C \subseteq \mathbb{R}^d$*

with $|C| \le k$, $|\mathrm{cost}_z(S_z, w_z, C) - \mathrm{cost}_z(P, C)|$ *is at most*

$$\eta\,\mathrm{cost}_z(P,C) + \widetilde{O}_z\left(kN_{\max}\left(\eta\left(\sqrt{\frac{k}{\rho}} + \sqrt{\frac{d}{\rho}}\right)^z + \eta^{1-z}\left(\frac{d}{\rho}\right)^{z/2}\right)\right). \tag{8}$$

The two canonical objectives follow directly from setting $z = 1$ and $z = 2$.

Corollary 4.4. *With probability at least* $1-\beta$*, the construction yields an* $\widetilde{O}(k\eta^{-4})$*-size* k*-median* η*-coreset with additive error* $\widetilde{O}(N_{\max}k(\eta\sqrt{k/\rho} + \sqrt{d/\rho}))$*, and an* $\widetilde{O}(k\eta^{-6})$*-size* k*-means* η*-coreset with additive error* $\widetilde{O}(N_{\max}k(\eta k/\rho + d/(\eta\rho)))$.

5 Experiments

5.1 Experimental Setup

Datasets The real-world datasets in the experiments are *Seeds*, *Travel Review*, and *3D Road Network*, all publicly available from the UCI Machine Learning Repository, containing 210 points in 7 dimensions, 980 points in 10 dimensions, and 500 points in 3 dimensions, respectively. We also generate synthetic datasets for evaluation. The generation procedures are described in Appendix C.

Evaluation Metrics For MEB and k-center, we use the *Approx. Ratio* to measure the radius/cost achieved by the private coreset relative to a non-private reference. For MEB, we use the exact MEB radius as the reference. For k-center, because it is NP-hard, we use the cost of the GMM solution as the reference. Specifically, for a MEB center $\widehat{\theta}$ and a k-center solution $\widehat{C}$, we compute

$$\text{Approx. Ratio} = \begin{cases} \dfrac{\max_{x\in P}\|x - \widehat{\theta}\|_2}{\mathrm{OPT}_1(P)}, & \text{for MEB,} \\ \dfrac{\max_{x\in P} d(x, \widehat{C})}{\max_{x\in P} d(x, C_{\mathrm{GMM}})}, & \text{for } k\text{-center,} \end{cases} \tag{9}$$

where C_{GMM} denotes the center set returned by GMM.

For k-means, we report two complementary metrics, *Cost Ratio* and *Distortion* (Schwiegelshohn & Sheikh-Omar, 2022). For a center set C, let $\Phi_P(C) = \sum_{x\in P} d(x, C)^2$ denote its cost on P. We run the same weighted k-means solver on the coreset to obtain C_{alg} and on P to obtain a reference solution C_{ref}. The *Cost Ratio* is defined as $\Phi_P(C_{\mathrm{alg}})/\Phi_P(C_{\mathrm{ref}})$. To further evaluate the coreset w.r.t. its definition, for a weighted coreset (S, w) we write $\Phi_{S,w}(C) = \sum_{s\in S} w(s)d(s, C)^2$. We evaluate a collection $\mathcal{C}$ of candidate center sets and report

$$\text{Distortion} = \max_{C\in\mathcal{C}} \max\left\{\frac{\Phi_{S,w}(C)}{\Phi_P(C)}, \frac{\Phi_P(C)}{\Phi_{S,w}(C)}\right\}. \tag{10}$$

Implementation Details We set $\gamma = 0.1$ throughout all the experiments. For MEB, we use a coreset size of $m = 10$ and a privacy parameter of $\rho = 3$ by default. For k-center, we use $k = 10$ and $\rho = 3$ by default. For k-means, the default setting is $\rho = 10$, $k = 10$, $m = 20$, and $|C| = 5$; when varying the privacy budget ρ, we instead set the coreset size to $m = 0.2n$. Unless otherwise specified, we fix all remaining parameters to their default values. For each experiment, we perform 100 independent runs with different random seeds. We report the mean, standard deviation, minimum, and maximum of each measure. All experiments were conducted on an Intel(R) Xeon(R) Silver 4210R server with 40 threads and 251 GiB of memory.

5.2 Experimental Results

Results for MEB Figure 1 reports the Approx. Ratio of CGP-MEB-Coreset under different privacy budgets. Across all datasets, increasing ρ consistently moves the returned radius closer to the non-private optimum. The variation across repeated runs also decreases, showing that the solution becomes more stable as the perturbation is reduced.

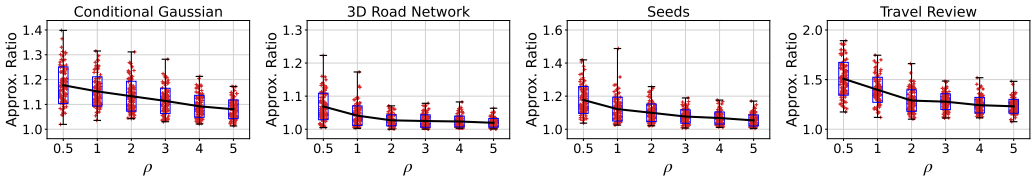

Figure 1: Approx. Ratio of CGP-MEB-Coreset by varying the privacy parameter ρ.

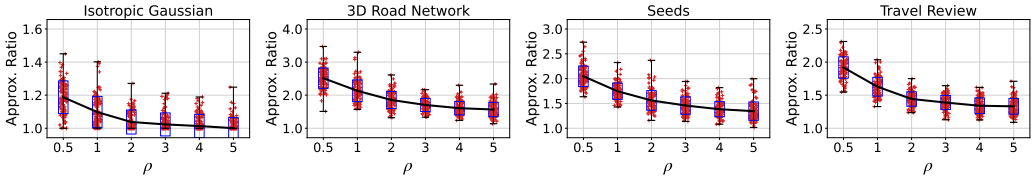

Figure 2: Approx. Ratio of CGP-GMM by varying the privacy parameter ρ.

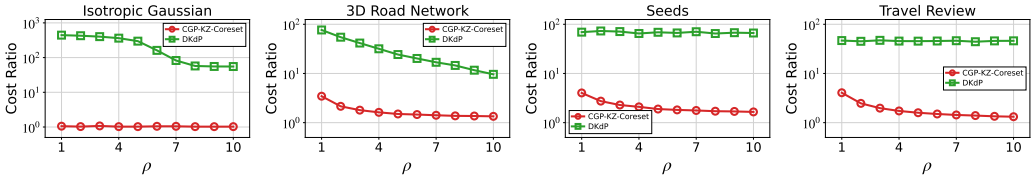

Figure 3: Cost Ratio of CGP-KZ-Coreset and DKdP by varying the privacy parameter ρ.

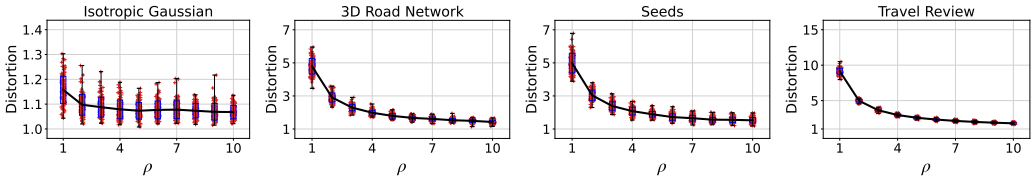

Figure 4: Distortion of CGP-KZ-Coreset by varying the privacy parameter ρ.

Results for k-Center Figure 2 shows a similar privacy–utility trend for CGP-GMM. The Approx. Ratio decreases steadily as ρ increases, and the dispersion across repeated runs becomes smaller. This indicates that the private farthest-point traversal approaches its non-private counterpart as the scalar and vector perturbations decrease.

Results for k-Means Figure 4 further evaluates how well the private coreset preserves the clustering objective. Distortion moves toward 1 as ρ increases, consistent with the Cost Ratio results. We further evaluate coreset distortion. Figure 4 varies ρ with $k = 10$ and $m = 0.2n$. The distortion approaches 1 as the privacy budget increases. The agreement between Cost Ratio and Distortion indicates that improved clustering quality is accompanied by better preservation of the clustering objective. Additional experiments varying the number of centers k and the coreset size m are reported in Appendix D.

6 Conclusion

In this paper, we propose a unified framework for private coreset construction for Euclidean clustering with concentrated geo-privacy (CGP). For MEB, our framework yields an $O(1/\gamma)$-size coreset with a $(1+\gamma)$ approximation up to an additive error. We extend it to k-center, obtaining both a practical private GMM traversal with an approximation factor of 2 and an exhaustive refinement achieving a $(1+\gamma)$ multiplicative approximation. For general (k, z)-clustering, we combine a private bicriteria construction with sensitivity sampling to obtain weighted private coresets with multiplicative approximation and additive privacy errors. Experiments on synthetic and real-world datasets demonstrate the effectiveness of our algorithms.

AI USE STATEMENT

In this work, we used generative AI tools only for minor language editing of author-written text. We have not used generative AI tools for generating research ideas, theoretical results, proofs, experimental designs, code, or data analysis, and other required disclosure tasks are not applicable to this work. All AI-assisted edits were reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide complete descriptions of the proposed algorithms and their theoretical guarantees in the main text, with proofs provided in appendix. Section 5 describes the datasets, evaluation metrics, implementation details, and experimental protocol. The appendix provides the synthetic data generation procedures and additional experimental results. These materials specify the algorithmic, theoretical, and experimental details needed to reproduce the results reported in this work.

REFERENCES

Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine learning*, 75(2):245–248, 2009. doi: 10.1007/s10994-009-5103-0.

Miguel E. Andrés, Nicolás Emilio Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, pp. 901–914, 2013. doi: 10.1145/2508859.2516735.

David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning. *arXiv:1703.06476*, 2017. URL https://arxiv.org/abs/1703.06476.

Olivier Bachem, Mario Lucic, and Andreas Krause. Scalable k-means clustering via lightweight coresets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1119–1127, 2018. doi: 10.1145/3219819.3219973.

Mihai Badoiu and Kenneth L. Clarkson. Smaller core-sets for balls. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 801–802, 2003.

Mihai Badoiu and Kenneth L. Clarkson. Optimal core-sets for balls. *Comput. Geom.*, 40(1):14–22, 2008. doi: 10.1016/J.COMGEO.2007.04.002.

Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing*, pp. 250–257, 2002. doi: 10.1145/509907.509947.

Maria-Florina Balcan, Travis Dick, Yingyu Liang, Wenlong Mou, and Hongyang Zhang. Differentially private clustering in high-dimensional euclidean spaces. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 322–331, 2017. URL http://proceedings.mlr.press/v70/balcan17a.html.

Nicolás Emilio Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Optimal geo-indistinguishable mechanisms for location privacy. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 251–262, 2014. doi: 10.1145/2660267.2660345.

Timothy M. Chan and Vinayak Pathak. Streaming and dynamic algorithms for minimum enclosing balls in high dimensions. *Comput. Geom.*, 47(2):240–247, 2014. doi: 10.1016/J.COMGEO.2013.05.007.

Alisa Chang, Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. Locally private k-means in one round. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1441–1451, 2021. URL http://proceedings.mlr.press/v139/chang21a.html.

Anamay Chaturvedi, Matthew Jones, and Huy Le Nguyen. Locally private k-means clustering with constant multiplicative approximation and near-optimal additive error. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pp. 6167–6174, 2022. doi: 10.1609/AAAI.V36I6.20565.

Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *Privacy Enhancing Technologies - 13th International Symposium, PETS 2013, Bloomington, IN, USA, July 10-12, 2013. Proceedings*, pp. 82–102, 2013. doi: 10.1007/978-3-642-39077-7_5.

Vincent Cohen-Addad, Alessandro Epasto, Vahab Mirrokni, Shyam Narayanan, and Peilin Zhong. Near-optimal private and scalable k-clustering. In *Advances in Neural Information Processing Systems 35*, pp. 10462–10475, 2022. doi: 10.52202/068431-0760.

Hu Ding. A sub-linear time framework for geometric optimization with outliers in high dimensions. In *28th Annual European Symposium on Algorithms (ESA 2020)*, pp. 38:1–38:21, 2020. doi: 10.4230/LIPICS.ESA.2020.38.

Hu Ding, Haikuo Yu, and Zixiu Wang. Greedy strategy works for k-center clustering with outliers and coreset construction. In *27th Annual European Symposium on Algorithms (ESA 2019)*, pp. 40:1–40:16, 2019. doi: 10.4230/LIPICS.ESA.2019.40.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, pp. 265–284, 2006. doi: 10.1007/11681878_14.

Alessandro Epasto, Vahab Mirrokni, Shyam Narayanan, and Peilin Zhong. k-means clustering with distance-based privacy. In *Advances in Neural Information Processing Systems 36*, pp. 19570–19593, 2023. doi: 10.52202/075280-0859.

Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing*, pp. 569–578, 2011. doi: 10.1145/1993636.1993712.

Dan Feldman, Amos Fiat, Haim Kaplan, and Kobbi Nissim. Private coresets. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pp. 361–370, 2009. doi: 10.1145/1536414.1536465.

Dan Feldman, Chongyuan Xiang, Ruihao Zhu, and Daniela Rus. Coresets for differentially private k-means clustering and applications to privacy in mobile sensor networks. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pp. 3–15, 2017. doi: 10.1145/3055031.305509.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the Thirteenth ACM International Conference on Web Search and Data Mining*, pp. 178–186, 2020. doi: 10.1145/3336191.3371856.

Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. Differentially private clustering: Tight approximation ratios. In *Advances in Neural Information Processing Systems 33*, pp. 4040–4054, 2020.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. doi: 10.1016/0304-3975(85)90224-5.

Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pp. 291–300, 2004. doi: 10.1145/1007352.1007400.

Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in euclidean spaces: Importance sampling is nearly optimal. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1416–1429, 2020. doi: 10.1145/3357713.3384296.

Lingxiao Huang, Jian Li, and Xuan Wu. On optimal coreset construction for Euclidean (k,z)-clustering. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 1594–1604, 2024. doi: 10.1145/3618260.3649707.

Matthew Jones, Huy L. Nguyen, and Thy Dinh Nguyen. Differentially private clustering via maximum coverage. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, pp. 11555–11563, 2021. doi: 10.1609/AAAI.V35I13.17375.

Ibrahim Jubran, Alaa Maalouf, and Dan Feldman. Overview of accurate coresets. *WIREs Data Mining and Knowledge Discovery*, 11(6):e1429, 2021. doi: 10.1002/widm.1429.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, 2011. doi: 10.1137/090756090.

Amer Krivosija and Alexander Munteanu. Probabilistic smallest enclosing ball in high dimensions via subgradient sampling. In *35th International Symposium on Computational Geometry (SoCG 2019)*, pp. 47:1–47:14, 2019. doi: 10.4230/LIPICS.SOCG.2019.47.

Piyush Kumar, Joseph S. B. Mitchell, and E. Alper Yildirim. Approximate minimum enclosing balls in high dimensions using core-sets. *ACM J. Exp. Algorithmics*, 8, 2003. doi: 10.1145/996546.996548.

Thomas Larsson and Linus Källberg. Fast and robust approximation of smallest enclosing balls in arbitrary dimensions. *Comput. Graph. Forum*, 32(5):93–101, 2013. doi: 10.1111/CGF.12176.

Yuting Liang and Ke Yi. Concentrated geo-privacy. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1934–1948, 2023. doi: 10.1145/3576915.3623068.

Yuting Liang and Ke Yi. Smooth sensitivity for geo-privacy. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 333–347, 2024. doi: 10.1145/3658644.3690365.

Bar Mahpud and Or Sheffet. A differentially private linear-time fptas for the minimum enclosing ball problem. In *Advances in Neural Information Processing Systems 35*, pp. 31640–31652, 2022. doi: 10.52202/068431-2294.

Nimrod Megiddo and Kenneth J. Supowit. On the complexity of some common geometric location problems. *SIAM Journal on Computing*, 13(1):182–196, 1984. doi: 10.1137/0213014.

Kobbi Nissim and Uri Stemmer. Clustering algorithms for the centralized and local models. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory*, pp. 619–653, 2018. URL http://proceedings.mlr.press/v83/nissim18a.html.

Kobbi Nissim, Uri Stemmer, and Salil P. Vadhan. Locating a small cluster privately. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 413–427, 2016. doi: 10.1145/2902251.2902296.

Chris Schwiegelshohn and Omar Ali Sheikh-Omar. An empirical evaluation of k-means coresets. In *30th Annual European Symposium on Algorithms (ESA 2022)*, pp. 84:1–84:17, 2022. doi: 10.4230/LIPICS.ESA.2022.84.

Moshe Shechner, Or Sheffet, and Uri Stemmer. Private k-means clustering with stability assumptions. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 2518–2528, 2020. URL http://proceedings.mlr.press/v108/shechner20a.html.

Uri Stemmer. Locally private k-means clustering. *Journal of Machine Learning Research*, 22(176):1–30, 2021. URL http://jmlr.org/papers/v22/20-721.html.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *IEEE Trans. Inf. Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.

Emo Welzl. Smallest enclosing disks (balls and ellipsoids). In *New Results and New Trends in Computer Science, Graz, Austria, June 20-21, 1991, Proceedings [on occasion of H. Maurer's 50th birthday]*, pp. 359–370, 1991. doi: 10.1007/BFB0038202.

Mengmeng Yang, Ivan Tjuawinata, and Kwok-Yan Lam. K-means clustering with local d_χ-privacy for privacy-preserving data analysis. *IEEE Trans. Inf. Forensics Secur.*, 17:2524–2537, 2022. doi: 10.1109/TIFS.2022.3189532.