Private Coresets

Dan Feldman Amos Fiat Haim Kaplan
School of Computer Science
Tel Aviv University
{dannyf, fiat, haimk}@tau.ac.il

Kobbi Nissim
Department of Computer Science
Ben-Gurion University
kobbi@cs.bgu.ac.il

ABSTRACT

A coreset of a point set P is a small weighted set of points that captures some geometric properties of P. Coresets have found use in a vast host of geometric settings. We forge a link between coresets, and differentially private sanitizations that can answer any number of queries without compromising privacy. We define the notion of *private coresets*, which are simultaneously both coresets *and* differentially private, and show how they may be constructed.

We first show that the existence of a small coreset with low generalized sensitivity (*i.e.*, replacing a single point in the original point set slightly affects the quality of the coreset) implies (in an inefficient manner) the existence of a private coreset for the same queries. This greatly extends the works of Blum, Ligett, and Roth [STOC 2008] and McSherry and Talwar [FOCS 2007].

We also give an *efficient* algorithm to compute private coresets for k-median and k-mean queries in $\Re^d$, immediately implying efficient differentially private sanitizations for such queries. Following McSherry and Talwar, this construction also gives efficient coalition proof (approximately dominant strategy) mechanisms for location problems.

Unlike coresets which only have a multiplicative approximation factor, we prove that private coresets must exhibit additive error. We present a new technique for showing lower bounds on this error.

Categories and Subject Descriptors: F. [Theory of Computation]: F.2 [Analysis of Algorithms and Problem Complexity]: F.2.2 [Nonnumerical Algorithms and Problems]: Geometrical problems and computations.
General Terms: Algorithms, Security, Theory.
Keywords: Coresets, differential privacy, privacy.

1. INTRODUCTION

The notion of Differential privacy [18] has emerged in a recent line of work on private data analysis that seeks a rigorous treatment of privacy and its consequences (see [15, 22, 20, 7, 18, 17, 16, 40, 19, 37, 5, 41, 8, 6, 21, 35]). It captures a very strong notion of privacy: irrespective of what the attacker seeks to learn and of the attacker's prior knowledge, the attacker learns very little about any specific individual. This is formalized by requiring that for any two databases that differ only on the details of one individual, the probability distribution on the outputs of a differentially private analysis are very similar, *i.e.*, the probability of any specific outcome differs by a multiplicative factor very close to one.

Much of the work on differential privacy is in an interactive scenario where a central authority (called a *curator*) answers a small number of specific queries. In our scenario, one would like to publish a "sanitized" version of the data, the publication of which preserves differential privacy. Such a database could be queried *ad infinitum* without impacting privacy. A query to the sanitized database should return (approximately) the same answer as the same query on the *original* database.

In general, it is impossible to produce differentially private sanitized databases (see [15, 18, 19, 21]). Therefore, the interesting question is "*For what query classes can we construct useful sanitized databases?*". We show that it is possible to construct differentially private sanitized databases for a large class of important problems in metric spaces.

In particular, we consider k-median queries in which the query is a set Q of k points and the answer is the sum of the distances of each point in the database to its closest point in Q. We give *efficient* constructions of sanitized databases for k-median queries in low dimensional spaces that we term *private coresets*. These have direct applications in further analysis of the data (*e.g.*, clustering), geometric optimization problems (*e.g.*, facility location), etc.

Private Coresets.

In computational geometry, a coreset for a point set P is a small (possibly weighted) point set C that is useful in computing approximate solutions for queries on P. Coresets have been the subject of many recent papers [29, 34, 4, 3, 32, 30, 3, 10, 1, 31, 23, 11, 25] and several surveys [1, 14]. Coresets have been used to great effect for a vast host of geometric and graph problems, including k-median [11, 25, 31, 32], k-mean [25, 31, 32], k-center [33], subspace approximation [23, 30, 34], shape fitting [2], k-line median [23], k-line center [33, 34], moving points [29], max TSP [27], minimum spanning tree [26, 13], maximal margin learning, etc. Coresets also imply streaming algorithms for many of these problems [1, 25, 32, 27, 36, 26, 9].

We define *private coresets* as coreset schemes that also preserve differential privacy. As our main running example, we focus on k-median queries (*i.e.*, the sum distances from P to some k points). Our techniques extend to k-mean queries (sum of squares of the distances), and to other problems.

In the original context of computational geometry, privacy is not a consideration, and the entire justification for the study of coresets is because of their small size (small number of points). Small core-

STOC'09, May 31–June 2, 2009, Bethesda, Maryland, USA.

sets imply efficient approximation algorithms for some geometric optimization problems. In contrast, a private coreset would be of value even if it's size exceeded that of the original set[1]. Moreover, a private coreset can be reduced in size, without much compromise in the quality of approximation, by applying a (traditional — non private) coreset scheme to it. This has no privacy implications.

Another distinction between (traditional) coresets and private coresets is that (traditional) coresets typically give a multiplicative approximation (a $1 \pm \epsilon$ factor), we show that private coresets require an additive error term (β). The additive error grows at least linearly with the diameter of the metric space from which the points are taken. Thus, we restrict our study to point sets taken from a metric space of bounded diameter (e.g., the unit ball). The question of how β grows a a function of other problem parameters (e.g., dimension) is further discussed herein.

The existence of private coreset schemes for many problems of interest greatly extends the set of problems for which differentially private sanitized databases exist, and in particular, the recent *general release mechanism* of [8] for low VC dimension range space queries[2].

1.1 Our Contributions

1. **From (non-private) coresets to private coresets:** It is intuitively appealing to think that because a coreset is much smaller than the point set it was computed for, it also preserves privacy. However, many coreset schemes output points that are contained in the original point set, and hence are not differentially private. Our first result shows, however, that this intuition is not completely flawed: the mere existence of small coresets for a family of queries immediately implies private coresets for these queries (the resulting schemes are not necessarily efficient). This construction is similar to the private PAC learner of [35] and the general release mechanism of [8], and uses the *exponential mechanism* of [37].

2. **Efficient private coresets:** Our main technical contribution is a framework for constructing private coresets efficiently. Informally, we reduce the problem of constructing private coresets for k-median queries to the construction of a particular type private coresets for 1-median (that we call z-centered private 1-median coresets) and the construction of a private ϵ-net.

3. **A lower bound on the additive error of private coresets:** We present a new technique for proving lower-bounds on the accuracy of differentially private schemes. In our context of private coresets, this techniques yields a lower bound on the additive error β of coresets for k-median queries in the d-dimensional Euclidean space.

Potential Applications.

1. **Differentially private clustering:** Our work suggests a very effective approach to private clustering. After applying the private coreset scheme on the input set P and obtaining a coreset C, any clustering algorithm/heuristic for k-median can be applied on C. In particular, it is possible to further reduce the size of C by applying a (not necessarily private) coreset scheme to obtain a coreset with a very small (typically a constant which depends on $1/\varepsilon$ and d) number of points C', and then compute the optimal clustering for C'. As C' is very small, the algorithm for computing the optimal clustering may be allowed to run in super-polynomial time.

   This approach is very different from previous work on the construction of differentially private clustering algorithms [7, 40] which are more restricted. Namely, [7] present a private implementation of a specific heuristic (Lloyd's heuristic) with limited guarantee on the outcome, and the construction does not readily adapt to accommodate other heuristics or algorithms. The algorithm in [40] is based on the sample and aggregate framework, and a meaningful outcome is obtained *only* if repeated applications of the underlying clustering algorithm on samples taken at random from the input point set result in consistent outcomes.

2. **Comparing alternatives:** Being a non-interactive sanitization scheme, private coresets allow issuing many queries without a further effect on privacy and hence enable comparing alternatives. For instance, it is possible to use a private coreset to experiment with possible hospital locations (say, trying to minimize the sum of distances from individual to their closest hospital). This problem is not the same as clustering, as location may be subjected to restrictions (*e.g.*, the hospital should not be built in a restricted access area such as a military base).

3. **Constructing approximately truthful mechanisms:** McSherry and Talwar [37] prove a connection between incentive compatible mechanisms and differential privacy: if differential privacy is preserved then small coalitions of manipulators reporting false information cannot significantly change their utility. The constructions in [37] utilize the exponential mechanism, and hence the resulting mechanisms are not computationally efficient.

   Private coresets can be used to give coalition proof (approximately) incentive compatible mechanisms. In particular, our constructions directly yield mechanisms without money for location problems, that are efficiently computable.

4. **Private Streaming Algorithms:** Coresets have been used to derive streaming algorithms (see, *e.g.*, [32]) and the usage of private coresets can similarly lead to private streaming algorithms.

1.2 Related Work

Several papers give non-interactive differentially private sanitizations [20, 38, 5, 8]. Most closely related to our work are the results of Blum, Ligett, and Roth [8] who prove the existence of differentially private "sanitized" databases for range queries. Range queries are of the form "what is the number of individuals within a range?" where the range space from which the queries are taken is of low VC dimension. As [8] make use of the exponential mechanism of [37] this is not an efficient construction.

We use techniques introduced for constructing (non-private) coresets for k-median. Such coresets were introduced by Har-Peled and Mazumdar [32], and successive improvements were given in [31, 11]. Our construction is strongly related to that of [32, 31].

Another tool we use is "bi-criteria" approximations. Specifically, the bi-criteria construction of Feldman *et al.* [24]. We show how private ϵ-nets can be used to construct a differentially private version of the construction in [24].

[1] This may actually be the case for our construction.

[2] The results of [8] can be viewed as private coresets for count queries for range families that exhibit low VC dimension.

2. DEFINITIONS

Throughout this paper we consider algorithms that take as input a set P of n points in some metric space $\mathcal{M}$, usually a subset of the d-dimensional Euclidean space. We note that we will actually allow several points to have the same coordinates, hence our input sets may actually be multi-sets. We denote the distance between two points p, p' by $\text{dist}(p, p')$ and define the distance between a point p and a set of points P to be $\text{dist}(p, P) = \min_{p' \in P} \text{dist}(p, p')$. The *difference* between two multi-sets of n points P and P' is the number of points in P that need to be modified (*i.e.*, moved) to get the set P', *i.e.*, $\text{SD}(P, P') = |P \setminus P'| = n - |P \cap P'|$. If $\text{SD}(P, P') = 1$ we say that P and P' are *neighboring*. As our privacy criteria we use the following notion of *differential privacy* put forward by [18]:

DEFINITION 2.1. *An algorithm $\mathcal{A}$ preserves α-differential privacy, if for all neighboring P, P' and for all sets of possible outcomes $\mathcal{C}$,*

$$\Pr[\mathcal{A}(P) \in \mathcal{C}] \leq e^{\alpha} \cdot \Pr[\mathcal{A}(P') \in \mathcal{C}],$$

where the probability is taken over the randomness of $\mathcal{A}$.

See Appendix A for basic properties of differential privacy.

For the problems that we consider a query Q is specified by a set of k points. We denote by $\mathcal{Q}$ the collection of all possible queries, and by $Q(P)$ the output of a query Q on a point set P. Concrete examples are k-median, k-mean, and k-center queries, which return the sum of distances $Q(P) = \sum_{p \in P} \text{dist}(p, Q)$, the sum of the squares of the distances $Q(P) = \sum_{p \in P} (\text{dist}(p, Q))^2$, and the maximum distance $Q(P) = \max_{p \in P} \text{dist}(p, Q)$, respectively.

Our definition of a coreset scheme deviates from the usual definition in the literature:

- We allow both additive and multiplicative approximation, with parameters β, and ϵ, respectively (whereas, only multiplicative approximation is usually considered); We later show that this is unavoidable for private coresets.
- We specify a bound δ on the probability that the scheme fails[3] to achieve the required approximation. This is required for private coresets because standard techniques of reducing failure probability (repeating the computation several times) are not directly applicable when privacy is a concern (see Lemma A.1 in Appendix A).
- Atypically (for coresets), the weight of a point in the coreset may be negative.

DEFINITION 2.2. *An $(\epsilon, \beta, \delta)$ coreset scheme $\mathcal{A}$ for a class of queries $\mathcal{Q}$, is an algorithm that gets as input a set P of n points and outputs a set $\mathcal{A}(P) = C$ of weighted points such that with probability at least $1 - \delta$ over the randomness of $\mathcal{A}$, for every $Q \in \mathcal{Q}$:*

$$(1 - \epsilon) \cdot Q(P) - \beta \leq Q(\mathcal{A}(P)) \leq (1 + \epsilon) \cdot Q(P) + \beta.$$

The set C is called a coreset.

Combining definitions 2.1 and 2.2, we can now define private coresets:

DEFINITION 2.3. *An $(\alpha, \epsilon, \beta, \delta)$-private coreset scheme for a class of queries $\mathcal{Q}$ is an algorithm $\mathcal{A}$ satisfying:*

- **Privacy:** *Algorithm $\mathcal{A}$ preserves α-differential privacy; and*
- **Utility:** *Algorithm $\mathcal{A}$ is an $(\epsilon, \beta, \delta)$ coreset scheme for the class of queries $\mathcal{Q}$.*

We note that while the main objective in (non-private) coresets is minimizing the number of points in the coreset $C = \mathcal{A}(P)$, this is not necessarily the case with private coresets – we actually do not mind large coresets as long as α-differential privacy is preserved. We note however, that it is possible to construct a private coreset scheme with number of points that is comparable to that of non-private scheme by first applying a *private* coreset scheme on the input, and then applying on its outcome a (regular, *non-private*) coreset scheme where the number of points is small. By Part 1 of Lemma A.1 the resulting algorithm is also a private coreset scheme.

We end this section with a simple consequence of Definition 2.3 – a lower bound demonstrating that every private coreset scheme for k-median must exhibit *additive* approximation error which is proportional to the diameter of the input domain (all omitted proofs are deferred to the full version of this paper). A lower bound which increases with the dimension is shown in Section 6, where we already bound the diameter of the metric space.

CLAIM 2.4. *Let $\mathcal{A}$ be an $(\alpha, \epsilon, \beta, \delta)$-private coreset scheme for k-median queries where the points P reside in a Euclidean space of diameter Λ. Then, $\beta = \Omega(\Lambda \ln(1/\delta)/\alpha)$.*

Similarly, we get $\beta = \Omega(\Lambda^2 \ln(1/\delta)/\alpha)$ for k-mean and $\beta = \Omega(\Lambda)$ for k-center. We will hence restrict the input of all our constructions to reside in a subset of the metric space $\mathcal{M}$ of bounded diameter (typically the unit ball). We emphasize, however, that we do not restrict the query set Q.

Before we get to our constructions, we note that the lower bound on β for k-center renders coresets for k-center to be of very limited interest (if at all): if the point sets P are taken from a subspace of diameter Λ then varying P can cause a change of at most Λ in $Q_{\text{ctr}}(P)$, which is of the same magnitude as β. This is not the case for k-median and k-mean, as varying P can cause a change of $n\Lambda$ and $n\Lambda^2$ in $Q_{\text{med}}(P)$ and $Q_{\text{mean}}(P)$ respectively.

3. AN INFORMATION THEORETIC UPPER BOUND

We begin with a result that is similar in spirit to the private PAC learner of [35] and the general release mechanism of [8]. It captures the intuition that the existence of small (traditional) coresets implies private coresets (of related parameters). For this section, point sets P and coresets C are located in the unit sphere on a d-dimensional grid of granularity polynomial in $n = |P|$[4].

CLAIM 3.1. *If there exists a $(\epsilon, \beta, \delta)$ coreset scheme $\mathcal{A}$ for k-median that on a set P of n points in the d-dimensional unit sphere outputs a coreset C consisting of $m = m(n)$ points, then there exists a $(\alpha, \epsilon, \beta', \delta')$ private coreset scheme $\mathcal{A}'$ for k-median where $\beta' = \beta + O\left((md \log n + \log(1/\delta'))/\alpha\right)$.*

The proof of Claim 3.1 employs the exponential mechanism of [37], and hence produces a coreset scheme that may be inefficient to compute.

THEOREM 3.2 ([11]). *There exists a coreset scheme for k-median in the d-dimensional Euclidean space that outputs coresets of size $O((\frac{k}{\epsilon})^2 \log n(d \log(1/\epsilon) + \log k + \log \log n))$.*

[3] See also traditional coresets in [11, 27, 26] that also define such δ (for a different reason).

[4] Moving the points to this grid introduces an additive error that is significantly smaller than the additive error inherent in the scheme.

Applying Claim 3.1 to the above coreset, we get:

COROLLARY 3.3. *There exists a $(\alpha, \epsilon, \beta, \delta)$ private coreset scheme for k-median in the d-dimensional Euclidean space where the additive error β is*

$$O\left(\frac{k^2 \log^2 n}{\alpha\epsilon^2} d\,(d\log(1/\epsilon) + \log k + \log\log n) + \frac{\log(1/\delta)}{\alpha}\right).$$

The scheme is not necessarily efficient.

Corollary 3.3 shows that (ignoring computation costs) it is possible to construct private coresets schemes for k-median with additive error that grows polynomially with d. This matches qualitatively our lower bound on β (see Section 6), and sets a goal for efficient constructions.

4. EFFICIENT PRIVATE CORESETS FOR k-MEDIAN QUERIES

We describe an efficient algorithm for private coresets that approximate k-median queries for sets P of n points in $\Re^d$, *i.e.*, all queries of the form

$$Q(P) = \sum_{p\in P} \text{dist}(p, Q), \quad \text{where } Q \subset \Re^d, |Q| \le k. \qquad (1)$$

The coresets we produce are weighted sets of points, C, where for every point $c \in C$, we denote its weight by $w^*(c) \in \Re$ (the $w^*(c)$ notation is to indicate that the exact weight $w(c)$ has been modified for privacy). Analogously to Eq.(1), we define

$$Q(C) = \sum_{c\in C} w^*(c) \cdot \text{dist}(c, Q).$$

For a set of points P, integer $k \ge 1$, we define $\text{opt}(P, k)$ to be the minimum of $Q(P)$ for sets Q of k points in $\Re^d$ (this is often referred to as the value of an optimal solution of the k-median problem, with an input P). That is

$$\text{opt}(P,k) = \min_{\substack{Q\subset\Re^d \\ |Q|=k}} Q(P) = \min_{\substack{Q\subset\Re^d \\ |Q|=k}} \sum_{p\in P} \text{dist}(p, Q). \qquad (2)$$

z-Centered Private Coresets.

Our basic building block is a z-centered $(\alpha, \epsilon, \beta, \delta)$-private 1-median coreset for a set of points P on a line in $\Re^d$. This construct preserves α-differential privacy and satisfies the following property:

DEFINITION 4.1. *A z-centered $(\alpha, \epsilon, \beta, \delta)$-private 1-median coreset scheme $\mathcal{A}$ is an α-differentially private algorithm that takes as input a set $P \subset \Re^d$ of n points and outputs a set of weighted points $\mathcal{A}(P)$ such that with probability at least $1-\delta$ over the randomness of $\mathcal{A}$, for every arbitrary (finite) set of points Q in $\Re^d$,*

$$|Q(P) - Q(\mathcal{A}(P))| \le \epsilon \cdot \sum_{p\in P} \textit{dist}(p, z) + \beta(\epsilon, \alpha, \delta).$$

We remark that taking z to be the mean of the points of P (plus some appropriate noise for diffential privacy) results in a 1-median private coreset. We also remark that the query, Q, for a z-centered private 1-median coreset, may be an arbitrary set of points in $\Re^d$ and is not restricted to being a single point in $\Re$.

We first construct a z centered 1-median private corset for points $P \subset \Re$ (*i.e.*, points on the real line), and then use this construction to build a z-centered 1-median private coreset for $P \subset \Re^d$, where $d > 1$ is a constant (Section 4.2). This is done by projecting P onto a small set of lines and taking the union of multiple z-centered private 1-median coresets for these projections.

Private Bi-Criteria Approximation.

Our second tool is a private bi-criteria approximation. Informally, the outcome of a bi-criteria approximation is a small collection of points $B \subset \Re^d$ such that $\sum_{p\in P}(p, B)$ is (roughly) bounded by $\text{opt}(P, k)$. More formally,

DEFINITION 4.2. *A $(\alpha, c, \beta, \delta)$-private bi-criteria approximation scheme $\mathcal{B}$ of opt(P, k) is an α-differentially private algorithm that takes as input a set $P \subset \Re^d$ of n points on the real line and outputs a (small) set of points $\mathcal{B}(P)$ such that with probability at least $1-\delta$ over the randomness of $\mathcal{B}$,*

$$\sum_{p\in P} \textit{dist}(p, \mathcal{B}(P)) \le c \cdot \textit{opt}(P, k) + \beta.$$

Note that unlike in coresets (and z-centered 1-median coresets) the requirement is only for the set $\mathcal{B}(P)$, and furthermore, we do not require that $\sum_{p\in P} \text{dist}(p, B)$ be bounded from below. In fact it could be much smaller than $\text{opt}(P, k)$. We show how to construct private bi-criteria approximations in Section 5.

Overview of the Construction.

Differentially private k-median coresets, for $P \subset \Re^d$, are constructed in Section 4.3 as follows:

1. Let $B = \{b_i\}$ (called "bi-criteria centers") where $|B| \approx k^d \log n$ be a private bi-criteria approximation on P.
2. Consider the Voronoi regions induced by the points in B. They partition P into $|B|$ subsets, $P_1, P_2, \ldots, P_{|B|}$, some of which may be empty.
3. For every P_i (even if empty) we build a b_i-centered private 1-median coreset. The final coreset is the union of these coresets.

NOTATION 4.3. *Our constructions use the noisy version of many calculations. To avoid confusion, we use the notation $Y(v)$ to denote the random variable corresponding to the noise added to variable v.*

4.1 An Efficient z-Centered Private 1-Median Coreset for Points on the Line

Given a set of points $P \subset [0, 1]$ and a point $z \in \Re$, we construct a z-centered private 1-median coreset for P. We note that for this construction the parameter z is a public, *i.e.*, the privacy requirement applies to the point set P but not to z.

Denote by n the number of points in P and let $n_0 \ge n$. Consider the following sequence of exponentially growing intervals around z. Define

$$r_0 = \frac{\ln(1/\delta)}{\alpha n_0} \quad \text{and} \quad t = \left\lceil \log_{1+\epsilon}\left(\frac{1}{r_0}\right)\right\rceil, \qquad (3)$$

where $0 < \epsilon \le 1$ is our multiplicative error parameter. Consider the exponentially growing sequence $r_i = (1+\epsilon)^i \cdot r_0$ where $1 \le i \le t$ and define the points $c_0^+ = c_0^- = z$, $c_i^+ = z + r_{i-1}$, and $c_i^- = z - r_{i-1}$ where $1 \le i \le t$. Define the intervals $I_i^+ = (c_{i-1}^+, c_i^+]$ and $I_i^- = [c_i^-, c_{i-1}^-)$. Note that the intervals are disjoint and $r_0 \cdot (1+\epsilon)^t = 1$, hence, every point of P lies in exactly one of these intervals. We take the coreset C, to be the point z, the right

endpoints of the I_i^+ intervals, $C^+ = \{c_i^+\}$ and the left endpoints of the I_i^- intervals, $C^- = \{c_i^-\}$.

We now define the weights for the points in C. We use the notation $w(c)$ for the "exact" weight of point $c \in C$ and $w^*(c)$ for its "noisy" weight. Let $w(c_0^+) = w(c_0^-) = w(z) = 0$ and, for $1 \le i \le t$, let $w(c_i^+) = |P \cap I_i^+|$ and $w(c_i^-) = |P \cap I_i^-|$. Assign the noisy weights $w^*(c_i^+) = w(c_i^+) + Y(c_i^+)$ and $w^*(c_i^-) = w(c_i^-) + Y(c_i^-)$ for $1 \le i \le t$ where the random noise variables $Y(\cdot)$ are chosen i.i.d. from Lap($\frac{1}{\alpha}$) (See definition A.3). Finally, define $Y(c_0^+) = -\sum_{i=1}^t Y(c_i^+)$, $Y(c_0^-) = -\sum_{i=1}^t Y(c_i^-)$, and assign the noisy weight for the point z as $w^*(z) = Y(c_0^+) + Y(c_0^-)$.

The private z-centered coreset consists of the set C and the weights $w^*(c)$. We emphasize that $w^*(z)$ is published as part of the the private coreset but $Y(c_0^+), Y(c_0^-)$ are not, as we refer to $Y(c_0^+)$ and $Y(c_0^-)$ only in the analysis. The properties of this construction are summaized in Claim 4.4 (follows from Corollary A.7) and Theorem 4.5.

CLAIM 4.4. *The algorithm above preserves 2α-differential privacy.*

THEOREM 4.5. *For any set $P = \{p_1, p_2, \ldots, p_n\} \subset [0,1]$, with probability at least $1 - O(\delta)$ for every query $Q \subset \Re^d$,*

$$|Q(P) - Q(C)| = \epsilon \cdot \sum_{p \in P} \mathit{dist}(p, z) + O(1) \cdot \frac{\ln(1/\delta)}{\alpha\epsilon}.$$

PROOF. Our analysis[5] bounds the difference between the estimated cost of a query and its real cost. We begin by splitting this difference into two sums and will then bound each sum separately. Informally, the first of these sums (Eq. (4)) is similar to a 'standard' error term that appears in the analysis of many coresets, and the second (Eq. (5)) is an error term resulting from the noisy count of the number of points that fall within each interval.

$$\begin{aligned}
&|Q(P) - Q(C)| \\
&= \left|\sum_{p\in P} \text{dist}(p,Q) - \sum_{c\in C} w^*(c)\cdot\text{dist}(c,Q)\right| \\
&= \left|\sum_{p\in P} \text{dist}(p,Q) - \sum_{c\in C} w(c)\cdot\text{dist}(c,Q)\right. \\
&\quad \left. - \sum_{c\in C} Y(c)\cdot\text{dist}(c,Q)\right| \\
&\le \left|\sum_{p\in P} \text{dist}(p,Q) - \sum_{c\in C} w(c)\cdot\text{dist}(c,Q)\right| \qquad (4) \\
&\quad + \left|\sum_{c\in C} Y(c)\cdot\text{dist}(c,Q)\right|. \qquad (5)
\end{aligned}$$

Deriving an upper bound on Eq. (4).

For $p \in P$, let $c(p)$ be the right (resp. left) endpoint of the interval I_i^+ (resp. I_i^-) that contains p. We can now express $\sum_{c\in C} w(c)\cdot\text{dist}(c,Q)$ as a sum over points in P:[6]

$$\begin{aligned}
&\left|\sum_{p\in P} \text{dist}(p,Q) - \sum_{c\in C} w(c)\cdot\text{dist}(c,Q)\right| \\
&= \left|\sum_{p\in P} \text{dist}(p,Q) - \sum_{p\in P} \text{dist}(c(p),Q)\right| \\
&\le \sum_{p\in P} |\text{dist}(p,Q) - \text{dist}(c(p),Q)|.
\end{aligned}$$

By the triangle inequality we have that

$$\begin{aligned}
\text{dist}(p,Q) - \text{dist}(c(p),Q) &\le \text{dist}(c(p),p), \quad \text{and} \qquad (6)\\
\text{dist}(c(p),Q) - \text{dist}(p,Q) &\le \text{dist}(c(p),p). \qquad (7)
\end{aligned}$$

By equations (6) and (7) we get that $|\text{dist}(p,Q) - \text{dist}(c(p),Q)| \le \text{dist}(c(p),p)$. Also, by construction, for $p \notin I_1^+ \cup I_1^-$, $\text{dist}(c(p),p) \le \epsilon \cdot \text{dist}(p,z)$. Hence,

$$\begin{aligned}
&\left|\sum_{p\in P} \text{dist}(p,Q) - \sum_{c\in C} w(c)\cdot\text{dist}(c,Q)\right| \\
&\le \sum_{p\in P} \text{dist}(c(p),p) \\
&\le \left(w(c_1^+) + w(c_1^-)\right)\cdot r_0 + \sum_{p\in P\setminus I_1^+\setminus I_1^-} \text{dist}(c(p),p) \\
&\le n\cdot\frac{\ln(1/\delta)}{\alpha n_0} + \sum_{p\in P} \epsilon\cdot\text{dist}(p,z) \\
&\le \frac{\ln(1/\delta)}{\alpha} + \sum_{p\in P} \epsilon\cdot\text{dist}(p,z). \qquad (8)
\end{aligned}$$

Deriving an upper bound on Eq. (5).

We begin by splitting the sum into sums over positive and negative intervals. We analyze the sum over positive intervals, the negative intervals are analogous.

$$\begin{aligned}
&\left|\sum_{c\in C} Y(c)\cdot\text{dist}(c,Q)\right| \\
&\le \left|\sum_{i=0}^t Y(c_i^+)\cdot\text{dist}(c_i^+,Q)\right| + \left|\sum_{i=0}^t Y(c_i^-)\cdot\text{dist}(c_i^-,Q)\right|.
\end{aligned}$$

Noting that $\sum_{i=0}^t Y(c_i^+) = 0$ (by our choice of $Y(c_0^+)$) we get

$$\begin{aligned}
&\left|\sum_{i=0}^t Y(c_i^+)\cdot\text{dist}(c_i^+,Q)\right| \\
&= \left|\sum_{i=0}^t Y(c_i^+)\cdot\text{dist}(c_i^+,Q) - \sum_{i=0}^t Y(c_i^+)\cdot\text{dist}(z,Q)\right| \\
&= \left|\sum_{i=1}^t Y(c_i^+)\cdot(\text{dist}(c_i^+,Q) - \text{dist}(z,Q))\right| \\
&\le \sum_{i=1}^t \left|Y(c_i^+)\right|\cdot\left|\text{dist}(c_i^+,Q) - \text{dist}(z,Q)\right| \\
&\le \sum_{i=1}^t \left|Y(c_i^+)\right|\text{dist}(c_i^+,z) = \sum_{i=1}^t \left|Y(c_i^+)\right|(1+\epsilon)^i r_0,
\end{aligned}$$

[5] A better bound can be proved for a single specific query. Details are deferred to the full version.

[6] Note that since $w(z) = 0$ it does not affect the sum.

where we omitted the case of $i = 0$ from the last sum as $c_0^+ = z$. We use the following lemma:

LEMMA 4.6. *Let $S = \sum_{i=1}^{t} Y_i \cdot (1+\epsilon)^i$ be the weighted sum of t random variables Y_i distributed i.i.d. according to the exponential distribution with mean $1/\alpha$. Then*

$$\Pr\left[|S| \geq 8\ln(\frac{1}{\delta}) \cdot \frac{(1+\epsilon)^t}{\alpha\epsilon}\right] \leq 2\delta.$$

We get that

$$\left|\sum_{i=0}^{t} Y(c_i^+) \cdot \text{dist}(c_i^+, Q)\right| \leq 8\ln(\frac{1}{\delta}) \cdot \frac{(1+\epsilon)^t}{\alpha\epsilon} \cdot r_0 = \frac{8\ln(\frac{1}{\delta})}{\alpha\epsilon},$$

where the last equality is by substituting $(1+\epsilon)^t = 1/r_0$ per Eq. (3). Combining this bound and Eq. (8) we get that with probability $1 - O(\delta)$

$$\begin{aligned}&\left|\sum_{p\in P} \text{dist}(p, Q) - \sum_{c\in C} w^*(c) \cdot \text{dist}(c, Q)\right| \\ &= O(1) \cdot \frac{\ln(1/\delta)}{\alpha\epsilon} + \epsilon \cdot \text{cost}(P, z)\end{aligned}$$

as required. □

4.2 An Efficient z-Centered Private 1-median Coreset for points in $\Re^d$

Using our construction from the previous section it is possible to construct z-centered private 1-median coresets for point sets taken from the d-dimensional unit ball. We sketch the scheme for $d = 2$, that can be easily generalized to any constant d.

We draw $t = O(1/\epsilon)$ lines $L_1, \ldots, L_t$ through z, that divide the plane into into $O(1/\epsilon)$ equal-sized sections, centered at z. Given the set P of points taken from the unit circle, we first construct a set P' that for every $p \in P$ contains the projection $\ell(p)$ of p on its nearest line (ties broken arbitrarily). Note that for every $p \in P$, this projection incurs a movement $\text{dist}(p, \ell(p)) = O(\epsilon) \cdot \text{dist}(p, z)$.

For each line L_i, we apply the algorithm presented in the previous section to construct a z-centered private 1-median coreset for the points in $P'_i = P' \cap L_i$, taking the privacy parameter to be α, the upper bound on the number of points n_0 to be $|P|$, and the failure probability to be $\delta/t = O(\epsilon\delta)$. Denote the coreset output for line L_i by C_i. Our z-centered coreset is the union $C_1, \ldots, C_t$.

With probability at least $1 - \delta$, all the underlying constructions succeed. We analyze the approximation quality of the resulting coreset if this is the case. For every query $Q \subset \Re^d$ we have

$$\left|\sum_{p\in P} \text{dist}(p, Q) - \sum_{c\in C} w^*(c) \cdot \text{dist}(c, Q)\right| \leq$$
$$\left|\sum_{p\in P} \text{dist}(p, \ell(p))\right| + \left|\sum_{p'\in P'} \text{dist}(p', Q) - \sum_{c\in C} w^*(c) \cdot \text{dist}(c, Q)\right|.$$

The first term is bounded by noting that the total displacement from P to P' is low:

$$\sum_{p\in P} |\text{dist}(p, \ell(p))| \leq \sum_{p\in P} O(\epsilon) \cdot \text{dist}(p, z)\ .$$

For the second term, we use our guarantees on z-centered coresets:

$$\begin{aligned}&\left|\sum_{p'\in P'} \text{dist}(p', Q) - \sum_{c\in C} w^*(c) \cdot \text{dist}(c, Q)\right| \\ &= \left|\sum_{i=1}^{t}\left[\sum_{p'\in P'_i} \text{dist}(p', Q) - \sum_{c\in C_i} w^*(c) \cdot \text{dist}(c, Q)\right]\right| \\ &\leq \sum_{i=1}^{t}\left|\sum_{p'\in P'_i} \text{dist}(p', Q) - \sum_{c\in C_i} w^*(c) \cdot \text{dist}(c, Q)\right| \\ &\leq \sum_{i=1}^{t}\left[\epsilon \sum_{p\in P'_i} \text{dist}(p, z) + O(1) \cdot \frac{\ln(1/\delta\epsilon)}{\alpha\epsilon}\right] \\ &\leq O(1) \cdot \left(\epsilon \sum_{p\in P} \text{dist}(p, z) + \frac{\ln(1/\delta\epsilon)}{\alpha\epsilon^2}\right).\end{aligned}$$

The last inequality follows since $\sum_{p\in P'} \text{dist}(p, z)$ is bounded by $(1+\epsilon)\sum_{p\in P} \text{dist}(p, z)$.

For the case where P lies in the d-dimensional unit ball, we get the bound $O(1) \cdot \left(\sum_{p\in P} \epsilon \cdot \text{cost}(p, z) + \frac{\ln(1/\delta\epsilon^d)}{\alpha\epsilon^{2d}}\right)$.

4.3 Private Coresets for k-median Queries

For this section, we assume that a differentially private bi-criteria approximation for the point set P was computed (with privacy parameter α_B), and we have a set $B = \{b_1, \ldots, b_{|B|}\}$ of points such that

$$\sum_{p\in P} \text{dist}(p, B) \leq O(1) \cdot \text{opt}(P, k) + \beta_B. \tag{9}$$

We defer the construction of a private bi-criteria approximation to Section 5 and show now how the set B can be used to construct a k-median coreset for P.

We consider the Voronoi regions $V_1, V_2, \ldots, V_{|B|}$ induced by B, where V_i is the Voronoi cell of point $b_i \in B$. Let P_i be the subset of P contained in V_i (P_i may be empty). For every P_i (even if empty) we build an $(\alpha, \epsilon, \beta, \delta)$ b_i-centered private 1-median coreset C_i as shown in Section 4.2 above. The final coreset C is simply the union of the coresets C_i.

As the sets P_i are disjoint, we get by part (2) of Lemma A.1 that this construction is $(\alpha_B + 2\alpha)$-differentially private. We now analyze its approximation guarantee.

Let Q be a query set of centers, $|Q| = k$. We have that

$$\begin{aligned}&|Q(P) - Q(C)| \\ &= |\sum_{i=1}^{|B|} Q(P_i) - Q(C_i)| \leq \sum_{i=1}^{|B|} |Q(P_i) - Q(C_i)| \\ &\leq \sum_{i=1}^{|B|}(\varepsilon \sum_{p\in P_i} \text{dist}(p, b_i) + \beta) \qquad (10) \\ &= \varepsilon \sum_{i=1}^{|B|} \sum_{p\in P_i} \text{dist}(p, b_i) + \beta|B| \\ &= \varepsilon \sum_{i=1}^{|B|} \sum_{p\in P_i} \text{dist}(p, B) + \beta|B| \qquad (11) \\ &= \varepsilon \sum_{p\in P} \text{dist}(p, B) + \beta|B| \\ &\leq \epsilon \cdot O(1)\text{opt}(P, k) + \epsilon \cdot \beta_B + \beta \cdot |B|\ . \qquad (12)\end{aligned}$$

Inequality (10) follows since C_i is a private b_i-centered 1-median coreset for P_i. Eq. (11) follows since by our construction, point $b_i \in B$ is the closest to all points in P_i, and hence $\text{dist}(P_i, B) = \text{dist}(P_i, b_i)$. Inequality (12) follows from the guarantee of Eq. (9) satisfied by the Bi-criteria approximation algorithm.

5. PRIVATE BI-CRITERIA APPROXIMATIONS

5.1 Private approximate weak ϵ-nets

Let $P \subset \Re^d$ be a point set of n points, and $k > 0$. We first describe a deterministic algorithm (hence, not differentially private) to compute a set E of $f(k,d)$ points, within the unit cube, such that every cube which contains at least $n/(8k)$ points must contains a point from E.

For $d = 1$ (*i.e.*, $P \subset \Re$), we partition the interval $[0,1]$ into $f(k,1) = 16k$ segments, $I_1, I_2, \ldots, I_{16k}$ each of which contains $n/16k$ points from P. Let m_j be the mean of the points in $I_j \cap P$, and take $E = \{m_j \mid j = 1, \ldots, 16k\}$. Any line segment that contains at least $n/(8k)$ points of P must contain a whole segment I_j, and therefore also m_j and hence it has a non-empty intersection with E.

For $d \geq 1$ we partition the unit cube of $\Re^d$ into c "slabs" ($a \leq x_1 \leq b$), each containing n/c points of P. We continue to partition such slabs along successive dimensions, the i'th partition is into "slabs", each of which contains n/c^i points of P. This construction yields a partition of the unit cube into c^d rectilinear *basic boxes*, R_j, each containing n/c^d points of P. We can prove the following lemma.

LEMMA 5.1. *A rectilinear box which does not contain one of the basic boxes contains at most $\frac{2dn}{c}$ points of P.*

PROOF. Let $e(d,n)$ be the maximum number of points contained in a rectilinear box which does not contain a basic box. For $d = 1$, the c basic boxes are intervals, each containing n/c points. An interval that does not contain a basic box is contained in the union of two basic boxes, and hence $e(1,n) \leq 2n/c$. For $d > 1$, we have that

$$e(d,n) \leq 2\frac{n}{c} + c \cdot e(d-1, \frac{n}{c}) \,. \tag{13}$$

To see this observe that a box which does not contain a basic box contains at most n/c points in the leftmost and rightmost slab which it intersects in the x_1 direction, and at most $e(d-1, \frac{n}{c})$ in every other slab. One can easily verify by induction that Eq. (13) implies that $e(d,n) \leq \frac{2dn}{c}$. □

The set E consists of c^d mean values, of the points in each such basic box. If we take $c = 16kd$ then by Lemma 5.1 any box with at least $n/8k$ points of P contains a basic box and therefore a point from E. This gives $f(k,d) = (16kd)^d$.

To get an α-differentially private version of the above, we make the following modifications:

1. We construct E^* – a noisy version of E – by replacing every point $e \in E$ by $e^* = e + Y(e)$ where $Y(e) \in \Re^d$ is chosen according to $\text{Lap}^d(df(k,d)/(n\alpha'))$ (*i.e.*, each coordinate of $Y(e)$ is chosen i.i.d. according to $\text{Lap}(df(k,d)/(n\alpha'))$).

   The mean e has global sensitivity $df(k,d)/n$ (by replacing a point in P each coordinate of e changes by $\pm f(k,d)/n$, hence the L_1 norm of the change is at most $df(k,d)/n$). Therefore, for every $e \in E$, e^* preserves α'-differential privacy by Theorem A.4. Setting $\alpha' = \alpha/f(k,d)$ we get that outputting e^* for all $e \in E$ preserves α-differential privacy.

2. By Fact A.5, the magnitude of each coordinate of $Y(e)$ is at most $\ln(1/\delta')f(k,d)/(n\alpha')$ with probability $1-\delta'$. Setting $\delta' = \delta/(d \cdot f(k,d))$ we get that with probability $1-\delta$, for all $e \in E$,

$$\text{dist}_\infty(e^*, e) \leq \ln\left(d \cdot f(k,d)/\delta\right) f(k,d)^2/(n\alpha),$$

and

$$\text{dist}(e^*, e) \leq \sqrt{d} \ln\left(d \cdot f(k,d)/\delta\right) f(k,d)^2/(n\alpha).$$

To simplify notation we denote $\ln(d \cdot f(k,d)/\delta) f(k,d)^2/(n\alpha)$ by $\gamma(k,d,\delta,\alpha)$. In summary, with confidence $1-\delta$, there exists an α-differentially private weak approximate $1/(8k)$-net with respect to boxes, where the distance between $e^* \in E^*$, and the associated $e \in E$, is bounded by $\gamma(k,d,\delta,\alpha)/n$.

5.2 Private Bi-criteria

Define $\text{OPT}(P,k)$ be a set of k points in $\Re^d$ attaining minimum cost (*i.e.*, $\sum_{p \in P} \text{dist}(p, \text{OPT}(P,k)) = \text{opt}(P,k)$, where $\text{opt}(P,k)$ is defined in Eq. (2)).

We give a variant of the bi-criteria approximation of [23] that makes use of the weak approximate $1/(8k)$ nets above. The process is iterative, with iterations indexed $i = 1, 2, \ldots, \log(n)$. Let P_i be the set of points at the beginning of iteration i ($P_1 = P$), and let $n_i = |P_i|$.

In iteration i, we build an α-private weak approximate $1/(8k)$ net E_i^* for P_i. We discard the $n_i/2$ points in P_i closest to E_i^*, and set $P_{i+1}$to be the set of the remaining points. The final bi-criteria points, E^*, are the union of all sets E_i^*, thus $|E^*| = f(k,d)\log n$.

5.3 Analysis

Consider iteration i. We say that the point $x \in P_i$ is *good* if $\text{dist}_\infty(x, E_i) \leq 2 \cdot \text{dist}_\infty(x, \text{OPT}(P,k))$ otherwise we say that x is *bad*. (Note that the definition of good and bad depends on E_i, and not E_i^*). We claim the following.

THEOREM 5.2. *For $P \subset \Re^d$, $|P| = n$, the set E^* is an $\alpha \log n$-private bi-criteria approximation to opt(P,k). More specifically, with probability $1 - \delta \log n$,*

$$\sum_{p \in P} \textit{dist}(p, E^*) \leq 4\sqrt{d} \sum_{p \in P} \textit{dist}(p, \textit{OPT}(P,k)) + \sqrt{d} \cdot \log n \cdot \gamma,$$

where $\gamma = \gamma(k,d,\delta,\alpha) = \ln(d \cdot f(k,d)/\delta) f(k,d)^2/\alpha$.

PROOF. Fix some iteration i, with probability at least $1-\delta$ the following holds: Let B_i be the set of bad points of iteration i discarded at iteration i, and let G_i be the set of good points of iteration i discarded at iteration $i+1$. Let $b \in B_i$ and $g \in G_i$. Since b is discarded and g is not discarded at iteration i then,

$$\text{dist}(b, E_i^*) \leq \text{dist}(g, E_i^*) \,. \tag{14}$$

Since g is good we also have that

$$\begin{aligned} \text{dist}(g, E_i^*) &\leq \text{dist}(g, E_i) + \sqrt{d}\gamma(k,d,\delta,\alpha)/n_i \\ &\leq 2\sqrt{d} \cdot \text{dist}(g, \text{OPT}(P,k)) + \sqrt{d}\gamma(k,d,\delta,\alpha)/n_i, \end{aligned} \tag{15}$$

and since $E_i^* \subseteq E^*$

$$\text{dist}(g, E^*) \leq \text{dist}(g, E_i^*) \;\text{ and }\; \text{dist}(b, E^*) \leq \text{dist}(b, E_i^*). \tag{16}$$

Combining (14), (15), and (16) we obtain that

$$\begin{aligned} \text{dist}(b, E^*) &\leq \text{dist}(g, E_i^*) \\ &\leq 2\sqrt{d} \cdot \text{dist}(g, \text{OPT}(P,k)) + \sqrt{d}\gamma(k,d,\delta,\alpha)/n_i. \end{aligned} \tag{17}$$

For a point $b \in B_i$ we do not know the relation between $\text{dist}(b, E^*)$ and $\text{dist}(b, P)$ and the former may be much larger. It follows from Eq. (17) that we can "charge" the distance $\text{dist}(b, E^*)$ to the distance $\text{dist}(g, \text{OPT}(P, k))$ for some point $g \in G_i$.

We argue that $|B_i| \leq |G_i|$, which means that we have enough points in G_i to avoid duplicate charges. Let $g(b) \in G_i$ be the point charged for $b \in B_i$. Along with (17) this gives that

$$\sum_{b \in B_i} \text{dist}(b, E^*) \leq 2\sqrt{d} \sum_{g(b) \in G_i} \text{dist}(g(b), \text{OPT}(P, k)) + \sqrt{d}(\gamma(k, d, \delta, \alpha)/n_i)|B_i|.$$

In addition to throwing away the bad elements in B_i, we also throw away good elements (with respect to E_i) in the i'th iteration, call this set F_i ($|F_i| = n_i/2 - |B_i|$). Using Eq. (16) and the fact that $g \in F_i$ is good we obtain that for every point $g \in F_i$, $\text{dist}(g, E^*) \leq 2\sqrt{d} \cdot \text{dist}(g, \text{OPT}(P, k)) + \sqrt{d}(\gamma(k, d, \delta, \alpha)/n_i)$. Thus, in total, for the points discarded during iteration i, $(B_i \cup F_i)$,

$$\sum_{b \in B_i} \text{dist}(b, E^*) + \sum_{g \in F_i} \text{dist}(g, E^*) \leq 2\sqrt{d} \sum_{g(b) \in G_i} \text{dist}(g(b), \text{OPT}(P, k)) + 2\sqrt{d} \sum_{g \in F_i} \text{dist}(g, \text{OPT}(P, k)) + \sqrt{d}\gamma(k, d, \delta, \alpha).$$

All points are eventually discarded, and discarded only once. Taking this sum over all iterations gives us that

$$\sum_{p \in P} \text{dist}(p, E^*) \leq 4\sqrt{d} \sum_{p \in P} \text{dist}(p, \text{OPT}(P, k)) + \sqrt{d} \log n \gamma(k, d, \delta, \alpha).$$

We still have to argue why $|B_i| \leq |G_i|$. Let z be a point which is not amongst the $n/8k$ closest points in L_∞ to a point of $\text{OPT}(P, k)$. For every point $v \in \text{OPT}(P, k)$ there exists a point $e \in E_i$ such that $\text{dist}_\infty(e, v) \leq \text{dist}_\infty(z, v)$. By the triangle inequality,

$$\text{dist}_\infty(z, e) \leq \text{dist}_\infty(z, v) + \text{dist}_\infty(v, e) \leq 2\text{dist}_\infty(z, v),$$

and so z must be good. It follows that $|B_i| \leq n_i/8$. At iteration $i + 1$ we discard $n_{i+1}/2 = n_i/4$ points so at least $n_i/8$ of those discarded in iteration $i+1$ must be good for iteration i, *i.e.*, $|G_i| \geq n_i/8$. □

6. A GENERIC LOWER BOUND ON THE ADDITIVE ERROR OF PRIVATE CORESET SCHEMES

We now prove a stronger bound than given in Claim 2.4 on the additive error β in private coreset schemes for k-median (an analogous bound holds for k-mean). We will consider points that reside in a subset of constant diameter of a metric space $\mathcal{M}$, and will assume the existence of sufficiently many points, the distance between every two of which is at least 1. We will show that if such a collection of points exists, then β should be large. More formally:

THEOREM 6.1. *Let $\mathcal{A}$ be an $(\alpha, \epsilon, \beta, \delta)$-private coreset scheme for k-median for sets of points P residing in the unit ball of a metric space. Assume that there exist $D + 1$ points $\mathcal{D} = \{\mathbf{0}, p_1, \ldots, p_D\}$ in this unit ball, such that the distance between every two points in $\mathcal{D}$ is at least 1. If $D > 2k$ then, $\beta = \Omega((\ln D + \ln 1/\delta)/\alpha)$.*

PROOF. We will construct a collection of point sets and show that unless β is large $\mathcal{A}$ fails to produce a good coreset, on at least one of the point sets, with probability greater than δ.

We call a subset σ containing exatly $k - 1$ distinct points in $\{p_1, \ldots, p_D\}$ *admissible*. For an admissible σ, define the set P_σ to include n points as follows: $n - (k-1) \cdot \ell$ points are located at $\mathbf{0}$, and ℓ points are located at each of the $k - 1$ locations $p \in \sigma$ (the value of ℓ will be set below, and we assume $n \geq k\ell$). Observe that $\text{cost}(P_\sigma, \{\mathbf{0}\} \cup \sigma) = 0$ and $\text{SD}(P_\sigma, P_{\sigma'}) = \ell \cdot \text{SD}(\sigma, \sigma')$.

If a set C satisfies $(1 - \epsilon) \cdot \text{cost}(P_\sigma, \cdot) - \beta \leq \text{cost}(C, \cdot) \leq (1 + \epsilon) \cdot \text{cost}(P_\sigma, \cdot) + \beta$, then we say that $C \in \text{good}(\sigma)$. We begin by showing that if σ and σ' differ on too many points then $\text{good}(\sigma) \cap \text{good}(\sigma') = \emptyset$. In the following we set $T = \lceil \frac{9\beta}{(1-\epsilon)\ell} \rceil$.

CLAIM 6.2. *Let σ and σ' be admissible such that $SD(\sigma, \sigma') \geq 2T$, and let $C \in \text{good}(\sigma)$ be a set of weighted points. Then, $C \notin \text{good}(\sigma')$.*

PROOF. Let O be a set of k points minimizing $\text{cost}(C, \cdot)$. We first show that $\left|\{p \in \sigma : \text{dist}(p, O) \geq \frac{1}{3}\}\right| < T$.

Since $C \in \text{good}(\sigma)$ we have that $(1 - \epsilon) \cdot \text{cost}(P_\sigma, O) - \beta \leq \text{cost}(C, O) \leq \text{cost}(C, \{\mathbf{0}\} \cup \sigma) \leq (1+\epsilon) \cdot \text{cost}(P_\sigma, \{\mathbf{0}\} \cup \sigma) + \beta = \beta$, and hence, $\text{cost}(P_\sigma, O) \leq 2\beta/(1 - \epsilon)$. On the other hand, if $\text{dist}(p, O) \geq \frac{1}{3}$ for some $p \in \sigma$ then every one of the ℓ points at location p contributes at least $\frac{1}{3}$ to $\text{cost}(P_\sigma, O)$. If that happens for at least T points $p \in \sigma$ we get that $\text{cost}(P_\sigma, O) \geq \frac{1}{3}T\ell \geq 3\beta/(1-\epsilon)$, contradicting our bound on $\text{cost}(P_\sigma, O)$. We hence get that $|\{p \in \sigma : \text{dist}(p, O) \geq \frac{1}{3}\}| < T$.

Since every pair of points in D (and therefore every pair of points in σ) are of distance at least one from each other then each point of O is within distance less than $\frac{1}{3}$ from at most a single point of σ. Since $|\sigma| = k-1$ and $|\{p \in \sigma : \text{dist}(p, O) \geq \frac{1}{3}\}| < T$ we get that more than $k - 1 - T$ of the points of O have a point of σ within distance less than $\frac{1}{3}$. Hence, for less than $|O| - (k - 1 - T) = k - (k - 1 - T) = T + 1$ points of O there may be a point in $\sigma' \setminus \sigma$ that is within distance less than $\frac{1}{3}$. Hence, $|\{p \in \sigma' \setminus \sigma : \text{dist}(p, O) < \frac{1}{3}\}| \leq T$.

To prove the claim, we need to show that $C \notin \text{good}(\sigma')$. Note that $\text{SD}(\sigma, \sigma') \geq 2T$ hence, $|\sigma \cap \sigma'| \leq |\sigma'| - 2T$ (recall that $\text{SD}(\sigma, \sigma') = |\sigma'| - |\sigma \cap \sigma'|$). We get that $\left|\left\{p \in \sigma' : \text{dist}(p, O) < \frac{1}{3}\right\}\right| \leq \left|\left\{p \in \sigma' \setminus \sigma : \text{dist}(p, O) < \frac{1}{3}\right\}\right| + |\sigma \cap \sigma'| \leq T + |\sigma'| - 2T = |\sigma'| - T$. Equivalently, $|\{p \in \sigma' : \text{dist}(p, O) \geq \frac{1}{3}\}| \geq |\sigma'| - (|\sigma'| - T) = T$ and hence $C \notin \text{good}(\sigma')$. □

Let S be a collection of admissible sets such that $2T \leq \text{SD}(\sigma, \sigma') \leq 24T$ for all $\sigma, \sigma' \in S$. Pick $\sigma \in S$. It follows from Claim 6.2 that $\text{good}(\sigma) \cap \text{good}(\sigma') = \emptyset$ for all $\sigma' \in S \setminus \{\sigma\}$.

Consider $\sigma' \in S \setminus \{\sigma\}$. We have that $\Pr[\mathcal{A}(P_{\sigma'}) \in \text{good}(\sigma')] \geq 1 - \delta$ and, since $\mathcal{A}$ is α-differentially private, the ratio $\Pr[\mathcal{A}(P_\sigma) \in \text{good}(\sigma')] / \Pr[\mathcal{A}(P_{\sigma'}) \in \text{good}(\sigma')]$ is at least $e^{-\alpha \cdot \text{SD}(P_\sigma, P_{\sigma'})}$ and hence $\Pr[\mathcal{A}(P_\sigma) \in \text{good}(\sigma')] \geq e^{-\alpha 24T\ell} \cdot (1 - \delta)$. We get that

$$\begin{aligned} &\Pr[\mathcal{A}(P_\sigma) \notin \text{good}(\sigma)] \\ &\geq e^{-\alpha 24T\ell} \cdot \Pr\left[\mathcal{A}(P_\sigma) \in \bigcup_{\sigma' \in S \setminus \{\sigma\}} \text{good}(\sigma')\right] \\ &= e^{-\alpha 24T\ell} \cdot \sum_{\sigma' \in S \setminus \{\sigma\}} \Pr\left[\mathcal{A}(P_\sigma) \in \text{good}(\sigma')\right] \quad (18) \\ &\geq (|S| - 1) \cdot e^{-\alpha \cdot 24T \cdot \ell}(1 - \delta), \end{aligned}$$

where the equality in (18) holds since the sets $\text{good}(\sigma')$ are disjoint. On the other hand, $\Pr[\mathcal{A}(P_\sigma) \notin C_\sigma] \leq \delta$ and hence we get

$$(|S| - 1) \cdot e^{-\alpha \cdot 24T \cdot \ell} \leq \frac{\delta}{1 - \delta}. \quad (19)$$

CLAIM 6.3. *There exists a set S of size $|S| \geq k^{8T}(D-k)^{6T}/2^{18T}$ such that $2T \leq SD(\sigma,\sigma') \leq 24T$ for all $\sigma,\sigma' \in S$.*

PROOF. We will use the following theorem on constant weight codes:

THEOREM 6.4. *[28] Let $A(n,2\delta,w)$ denote the maximum number of codewords in any binary code of length n, constant weight w, and Hamming distance 2δ. Let q be a prime power such that $q \geq n$, then $A(n,2\delta,w) \geq \binom{n}{w}/q^{\delta-1}$.*

For an admissible σ we will use the notation $[\sigma]$ for the D-bit indicator vector of σ wrt $\{p_1,\ldots,p_D\}$. Note that $\mathrm{dist}_{\mathrm{H}}([\sigma],[\sigma']) = 2 \cdot \mathrm{SD}(\sigma,\sigma')$. Hence, to construct S it suffices to construct a collection of D-bit vectors of weight $k-1$ satisfying $4T \leq \mathrm{dist}_{\mathrm{H}}(v,v') \leq 48T$.

Each vector in our collection is a concatenation of a $(k-1)$-bit vector of weight $k-1-16T$ with a vector taken from a binary code of length $D-k+1$, constant weight $16T$ and distance $4T$. The resulting vectors are all of length D and of weight $k-1$, and the Hamming distance between any two vectors is at most $32T + 16T = 48T$ and at least $4T$. The corresponding sets are all of size $k-1$ (and hence admissible), and the difference between any two sets is between $2T$ and $24T$. We get that exists a set S as required with

$$\begin{aligned} |S| &\geq \binom{k-1}{16T} \cdot A(D-k+1,4T,16T) \\ &\geq \binom{k-1}{16T} \cdot \frac{\binom{D-k+1}{16T}}{(2(D-k+1))^{2T}} \\ &\geq (k/2)^{8T} \cdot \frac{((D-k)/2)^{8T}}{(2(D-k))^{2T}} = \frac{k^{8T}(D-k)^{6T}}{2^{18T}}. \end{aligned}$$

□

As we assume $D > 2k$, we get that $|S| \geq \frac{k^{8T}(D/2)^{6T}}{2^{18T}} \geq \frac{D^{6T}}{2^{24T}}$. Plugging this bound for $|S|$ in Eq. (19) (and neglecting the "-1") we get $e^{6T\ln D - 24\ln 2 - 24\alpha T\ell} \leq e^{-\ln\frac{1-\delta}{\delta}}$, or, equivalently, $24\alpha T\ell > 6T\ln D - 24\ln 2 + \ln\frac{1-\delta}{\delta}$. Setting $\ell = \beta/2$ we get that $T = O(1)$ (assuming $\epsilon < 1/2$), and hence we get that $\beta = \Omega((\ln 1/\delta + \ln D)/\alpha)$ as required.

For private coresets in the Euclidean space we get:

COROLLARY 6.5. *Let $\mathcal{A}$ be an $(\alpha,\epsilon,\beta,\delta)$-private coreset scheme for k-median where the points P reside in the d-dimensional unit ball. Then, $\beta = \Omega((\ln 1/\delta + \sqrt{d})/\alpha)$.*

PROOF. Applying Theorem 6.1 the set $\mathcal{D}$ of 2^d vertices of the unit cube we get a lowerbound of $\Omega(\ln 2^d/\alpha) = \Omega(d/\alpha)$ for the case where the points P are taken from a ball of radius $\sqrt{d}$. Deflating this bound by a factor of $\sqrt{d}$ we get that $\beta = \Omega(\sqrt{d}/\alpha)$, and as by Claim 2.4 $\beta = \Omega(\ln(1/\delta)/\alpha)$ we get that $\beta = \Omega((\ln 1/\delta + \sqrt{d})/\alpha)$. □

Acknowledgments

Research partly supported by the Israel Science Foundation grants No. 860/06 and No. 975/06, by the United states - Israel Binational Science Foundation, grant No. 2006204 and the by the Frankel Center for Computer Science at Ben-Gurion University.

7. REFERENCES

[1] P. Agarwal, S. Har-Peled, and K.R. Varadarajan. Geometric approximations via coresets. *Combinatorial and Computational Geometry - MSRI Publications*, 52:1–30, 2005.

[2] P. Agarwal, S. Har-Peled, and K.R. Varadarajan. Approximating extent measures of points. *JACM: Journal of the ACM*, 51, 2004.

[3] P.K. Agarwal, C.M. Procopiuc, and K.R. Varadarajan. Approximation algorithms for k-line center. *ESA*, pages 54–63, 2002.

[4] M. Badoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. *STOC*, pages 250–257, 2002.

[5] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. *PODS*, pages 273–282, 2007.

[6] A. Beimel, K. Nissim, and E. Omri. Distributed private data analysis: Simultaneously solving how and what. *CRYPTO*, pages 451–468, 2008.

[7] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. *PODS*, pages 128–138, 2005.

[8] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. *STOC*, pages 609–618, 2008.

[9] L.S. Buriol, G. Frahling, S. Leonardi, and C. Sohler. Estimating clustering indexes in data streams. *ESA*, pages 618–632, 2007.

[10] T.M. Chan. Faster core-set constructions and data stream algorithms in fixed dimensions. *SCG*, pages 152–159, 2004.

[11] K. Chen. On k-median clusteing in high dimensions. *SODA*, pages 1177-1185, 2006.

[12] K.L. Clarkson. Subgradient and sampling algorithms for l1 regression. *SODA*, pages 257–266, 2005.

[13] A. Czumaj, F. Ergun, L. Fortnow, A. Magen, I. Newman, R. Rubinfeld, and C. Sohler. Approximating the weight of the euclidean minimum spanning tree in sublinear time. *SIAM J. Computing* vol. 35(1), pages 91–109, 2005.

[14] A. Czumaj and C. Sohler. Sublinear-time approximation algorithms for clustering via random sampling. *RSA: Random Structures & Algorithms*, 30, 2007.

[15] I. Dinur and K. Nissim. Revealing information while preserving privacy. *PODS*, pages 202–210. ACM, 2003.

[16] C. Dwork. Differential privacy. *ICALP*, LNCS, pages 1–12, 2006.

[17] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. *EUROCRYPT*, LNCS, pages 486–503. Springer, 2006.

[18] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. *TCC*, pages 265–284, 2006.

[19] C. Dwork, F. McSherry, and K. Talwar. The price of privacy and the limits of lp decoding. *STOC*, pages 85–94, 2007.

[20] C. Dwork and K .Nissim. Privacy-preserving datamining on vertically partitioned databases. *CRYPTO*, pages 528–544, 2004.

[21] C. Dwork and S. Yekhanin. New efficient attacks on statistical disclosure control mechanisms. *CRYPTO*, pages 469–480, 2008.

[22] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy

breaches in privacy preserving data mining. *PODS*, pages 211–222, 2003.
[23] D. Feldman, A. Fiat, and M. Sharir. Coresets for weighted facilities and their applications. *FOCS*, pages 315–324, 2006.
[24] D. Feldman, A. Fiat, M. Sharir, and D. Segev. Bi-criteria linear-time approximations for generalized k-mean/median/center. *SoCG*, pages 19–26, 2007.
[25] D. Feldman, M. Monemizadeh, and C. Sohler. A PTAS for k-means clustering based on weak coresets. *Symposium on Computational Geometry*, pages 11–18, 2007.
[26] G. Frahling, P. Indyk, and C. Sohler. Sampling in dynamic data streams and applications. *Int. J. Comput. Geometry Appl*, 18(1/2):3–28, 2008.
[27] G. Frahling and C. Sohler. Coresets in dynamic geometric data streams. *STOC*, pages 209–217, 2005.
[28] R. L. Graham and N. J. A. Sloane. Lower bounds for constant weight codes. *IEEE Trans. Inform. Theory*, 26(1):37–43, 1980.
[29] S. Har-Peled. Clustering motion. *FOCS*, pages 84–93, 2001.
[30] S. Har-Peled. No coreset, no cry. *FSTTCS*, pages 324–335, 2004.
[31] S. Har-Peled and A. Kushal. Smaller coresets for k-median and k-means clustering. *SoCG*, pages 126–134, 2005.
[32] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. *STOC '04*, pages 291–300, 2004.
[33] S. Har-Peled and K.R. Varadarajan. High-dimensional shape fitting in linear time. *GEOMETRY: Discrete & Computational Geometry*, 32, 2004.
[34] S. Har-Peled and K.R. Varadarajan. Projective clustering in high dimensions using core-sets. *SoCG*, pages 312–318, 2002.
[35] S.P. Kasiviswanathan, H.K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? *FOCS*, pages 531–540, 2008.
[36] C. Lammersen and C. Sohler. Facility location in dynamic geometric data streams. *ESA*, pages 660–671, 2008.
[37] F. McSherry and K. Talwar. Mechanism design via differential privacy. *FOCS*, pages 94–103, 2007.
[38] N. Mishra and M. Sandler. Privacy via pseudorandom sketches. *PODS*, pages 143–152, 2006.
[39] K. Nissim. Private data analysis via output perturbation. In Charu Aggarwal and Philip S. Yu, editors, *Privacy-Preserving Data Mining, Models and Algorithms*, pages 383–414.
[40] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. *STOC*, pages 75–84, 2007.
[41] V. Rastogi, S. Hong, and D. Suciu. The boundary between privacy and utility in data publishing. *VLDB*, pages 531–542, 2007.

APPENDIX

A. TOOLS FOR DIFFERENTIAL PRIVACY

We review some of the basic tools for constructing differentially private algorithms. For more detail, the reader is referred to [7, 18, 40, 37, 39].

Composition.

The following lemma allows us to combine the outcome of several differentially private analyses in creating a differentially private algorithm:

LEMMA A.1. *Let $\mathcal{A}(\cdot)$ be any randomized computation, let $\mathcal{A}_0(\cdot)$ be α_0-differentially private, and let $\mathcal{A}_1(\cdot,\cdot)$ be a (parameterized) computation such that $\mathcal{A}_1(C_0,\cdot)$ is α_1-differentially private for all C_0. Then,*

1. *Algorithm $\mathcal{B}$ that on input P computes $C_0 \leftarrow \mathcal{A}_0(P)$, then $C \leftarrow \mathcal{A}(C_0)$ and outputs C is α_0-differentially private.*
2. *Algorithm $\mathcal{B}'$ that on input P computes $C_0 \leftarrow \mathcal{A}_0(P)$, then $C_1 \leftarrow \mathcal{A}_1(C_0, P)$ and outputs (C_0, C_1) is $(\alpha_0+\alpha_1)$-differentially private.*

The Framework of Global Sensitivity.

DEFINITION A.2 ([18]). *Let f be a deterministic function that given a set P of n points returns a value in $\Re^d$. The* global sensitivity *of f is the maximal change that can occur by a change in a single point, i.e.,*

$$GS_f(n) = \max \|f(P) - f(P')\|_1,$$

where the maximum is taken over all neighboring point sets P, P'.

DEFINITION A.3. *The Laplace distribution, denoted Lap(λ) has the probability density function $h(x) = \frac{1}{2\lambda}\exp(-\frac{|x|}{\lambda})$.*

The following theorem from [18] gives a simple recipe for constructing differentially private analyses:

THEOREM A.4 ([18]). *Let f be a function with global sensitivity $GS_f(n)$. The algorithm that given a point set P outputs $f(P) + Y$ where $Y \sim Lap^d(\frac{GS_f(n)}{\alpha})$ preserves α-differential privacy.*

We use the following simple fact about the Laplace distribution and simple corollaries of Lemma A.4:

FACT A.5. *Let $Y \sim Lap(\frac{1}{\alpha})$ for $\alpha > 0$. Then, $\mathbb{E}[Y] = 0$ and $\Pr[|Y| \geq z] = \exp(-\alpha z)$ for $z > 0$. In particular, $\Pr[|Y| \geq \frac{\ln(1/\delta)}{\alpha}] = \delta$.*

COROLLARY A.6. *The algorithm that given a set P of n points in the interval $[0,1]$ outputs $Z = \frac{1}{n}\sum_{p\in P} p + Y$ where $Y \sim Lap(\frac{1}{\alpha n})$ preserves α-differential privacy.*

COROLLARY A.7. *Let $I_1, \ldots, I_t$ be disjoint sub-intervals of $[0,1]$. The algorithm that given a set P of n points in the interval $[0,1]$ outputs $(Z_1, \ldots, Z_t)$ where $Z_i = |I_i \cap P| + Y_i$ and $Y_i \sim Lap(\frac{2}{\alpha})$ preserves α-differential privacy.*

COROLLARY A.8. *Let $q \in \Re$. The algorithm that given a set P of n points in the interval $[0,1]$ outputs $Z = \sum_{p\in P} dist(p,q) + Y$ where $Y \sim Lap(\frac{1}{\alpha})$ preserves α-differential privacy.*