Differentially-Private Sublinear-Time Clustering

Jeremiah Blocki*, Elena Grigorescu*, Tamalika Mukherjee* *
Department of Computer Science, Purdue University.
{jblocki, elena-g, tmukherj}@purdue.edu

***Abstract*—Clustering is an essential primitive in unsupervised machine learning. We bring forth the problem of sublinear-time differentially-private clustering as a natural and well-motivated direction of research. We combine the k-means and k-median sublinear-time results of Mishra et al. (SODA, 2001) and of Czumaj and Sohler (Rand. Struct. and Algorithms, 2007) with recent results on private clustering of Balcan et al. (ICML 2017), Gupta et al. (SODA, 2010) and Ghazi et al. (NeurIPS, 2020) to obtain sublinear-time private k-means and k-median algorithms via subsampling. We also investigate the privacy benefits of subsampling for group privacy.**

I. Introduction

Preserving privacy in data collection and distribution have long been a concern for industrial and governmental agencies, who are now rapidly adopting privacy standards and policies [22], [11], [6], [13]. Differential privacy [12] is the gold standard of privacy protection. A randomized function computed on a database is *differentially private* if the distribution of the function's output does not change by much with the presence or absence of an individual record. While existing research mostly focuses on computing efficient polynomial-time differentially-private algorithms, in dealing with a large amount of data, even linear-time algorithms may be prohibitive in costs. Hence, algorithms that can quickly output approximately accurate solutions while preserving privacy are of great interest in real-world computations on large datasets (e.g., billions of Facebook or Google, or Microsoft users). However, despite the fact that the literature on differentially private algorithms has grown rapidly in recent years, sublinear-time private algorithms for many natural problems are still lacking. In this work we focus on clustering problems and provide some basic sublinear-time private solutions derived from the existing efficient analogues.

Clustering is an essential primitive in unsupervised machine learning. Since many machine learning models deal with sensitive data, private clustering has been studied extensively in the polynomial-time setting [25], [14], [17], [30], [4], [20], [19], [28], [16], [29], [26]. Two of the most widely studied variants of clustering are the k-median and k-means problem. In the k-median problem, we are given n data points, and the goal is to find k centers that minimize the sum of distances from the data points to their nearest centers. The setup is the same for k-means, except the goal is to find k centers that minimize the sum of the squares of distances from the data points to their nearest centers. Both types of clustering are classical problems, and there is a rich field of research devoted to them in the non-private setting [2], [7], [18], [8], [9], [3], [23], [1], [27].

A. Contributions

We bring forth the problem of sublinear-time private clustering as a natural and well-motivated direction of research, and show some basic results derived from the non-private analogues on subsampled data. We expect that our results will entice further interest in understanding the best privacy guarantees in sublinear clustering settings.

Private sublinear clustering. We combine the techniques of sublinear-time clustering algorithms from Mishra et al. [24] and Czumaj et al. [10] with the private polynomial-time approximation clustering algorithms with a constant multiplicative factor of Balcan et al. [4], Gupta et al. [17] and Ghazi et al. [16] to obtain private sublinear-time clustering algorithms for k-median and

k-means clustering in metric spaces, as well as better approximation guarantees for the particular case of Euclidean space. To the best of our knowledge, these are the first *sublinear-time differentially-private clustering* algorithms formalized in the privacy literature.

We analyze the following sampling algorithm: pick a random sample from the input set; run a private k-median (or k-means) polynomial-time approximation algorithm on the random sample to obtain a k-median (or k-means) clustering of the sample; output this clustering. We show that for a small sample size, the average cost of the clustering induced by the random sample is not too far from the average cost of the optimum clustering of the input set. Our analysis closely follows the works of Mishra et al. [24] and Czumaj et al. [10], who gave sublinear time algorithms for clustering in the non-private setting using a constant α-approximation polynomial-time algorithm as a black-box. We extend their analysis to handle the case of using an (α, γ)-approximation polynomial time algorithm as black-box. We note that an additive approximation of $\gamma > 0$ is unavoidable for any *private* clustering algorithm, thus this extension was necessary. The approximation guarantee achieved by our algorithm is essentially the same as that of the black-box private algorithms (modulo an extra additive factor of ϵ).

For an arbitrary metric space (V, d) consisting of n points and input set $D \subseteq V$, assuming a private (α, γ)-factor approximation k-median (respectively k-means) algorithm, that runs in time $T(n)$, we can draw a sample $S \subseteq D$ of size $poly(\alpha, k \ln n)$ and obtain a k-median (respectively k-means) clustering $\hat{c}_S$ in time $T(s)$ such that with high probability $\mathsf{avg\text{-}cost}(\hat{c}_S) \leq \alpha \cdot \mathsf{avg\text{-}cost}(c_D) + \gamma + \epsilon$, where c_D is the optimum k-median (respectively k-means) clustering of D.

For the special case of clustering in d-dimensional Euclidean space, we achieve a sample complexity that is independent of the size of the input set $D \subseteq \mathbb{R}^d$ consisting of n points. Assuming a private (α, γ)-factor approximation k-median (respectively k-means) algorithm, that runs in time $T(n)$, we can draw a sample $S \subseteq D$ of size $poly(\alpha, d, k)$ and obtain a k-median (respectively k-means) clustering $\hat{c}_S$ in time $T(s)$ such that with high probability $\mathsf{avg\text{-}cost}(\hat{c}_S) \leq \alpha \cdot \mathsf{avg\text{-}cost}(c_D) + \gamma + \epsilon$, where c_D is the optimum k-median (respectively k-means) clustering of D.

Group privacy for sampling algorithms. Group privacy ensures that for pairs of inputs that differ on a small number of points, the privacy loss is still bounded. For example, in the setting of a health survey administered to families, a family may wish to preserve all its members' privacy. Any ξ-differentially private algorithm, ensures $(g\xi, 0)$-privacy for groups of size g. We show that our random sampling algorithm has better group privacy guarantees. In other words, an algorithm that runs an $(\xi, 0)$-differentially private mechanism on a subsample is $(T \cdot \xi, \delta_T)$-differentially private for groups of size g, for $0 \leq T \leq g$, where δ_T is the probability of the number of samples from the g elements is $> T$. We note that δ_T is often negligible even for $T \ll g$. In such cases, the guarantee of $(T \cdot \xi, \delta_T)$-differential privacy is arguably much stronger than the naive guarantee of $(g\xi, 0)$-group privacy.

B. Related Work

Sublinear-time approximate k-median clustering of a space in which the diameter of points is bounded was introduced by Mishra et al. [24]. They modeled clusterings as functions and studied the quality of k-median clusterings obtained by random sampling using computational learning theory techniques. Their sampling model was adapted by Czumaj et al. [10], who not only obtained better sample complexity bounds (independent of n) for the k-median problem, but they also extended the random sampling model and their analysis to give sublinear-time results for clustering variants such as k-means and min-sum clustering.

Private clustering was first studied by Gupta et al. [17], and Feldman et al. [14]. The k-median algorithm by [17] achieves superb approximation guarantees and runs in polynomial time in discrete spaces, however the algorithm is highly inefficient in Euclidean space (see [20] for a detailed exposition). A recent line of work has focused on producing an efficient polynomial time algorithm for clustering that achieves a constant (multiplicative) factor approximation in high-dimensional Euclidean space by adopting the techniques of Gupta et al. while maintaining efficiency [4], [20]. A different approach to private clustering was taken by [14]. They gave an efficient algorithm for k-median and k-means in Euclidean space by introducing the notion of private coresets. A recent line of work has adopted

their techniques to give clustering algorithms with better approximation guarantees and efficiency [15], [26], [16].

Privacy amplification by subsampling has been formally studied by Balle et al. [5]. Our result is a simple observation that tailors the privacy amplification achieved with respect to group privacy for a generic sampling algorithm that runs a private algorithm as a black-box in the sampling step.

II. Preliminaries

In the following discussion, let (V,d) be an arbitrary metric space.

The expected (average) value of a function f over uniform elements of a set X is denoted as $\mathbb{E}_X[f]$, i.e. $\mathbb{E}_X[f] = \sum_{x \in X} \Pr_{x \leftarrow X}[x] \cdot f(x)$, where x is picked uniformly from X.

Differential Privacy. Datasets D and D' are *neighboring* if $|D \Delta D'| = 1$. A randomized algorithm $\mathcal{M}$ taking as input a dataset D is (ξ,δ)*-differentially private* if for any two neighboring data sets D and D', and for any subset C of outputs of $\mathcal{M}$ it holds that $\Pr[\mathcal{M}(D) \in C] \leq e^{\xi} \cdot \Pr[\mathcal{M}(D') \in C] + \delta$. If $\delta = 0$, $\mathcal{M}$ is ξ*-differentially private.*

Clustering. Given an input set $D \subseteq V$, the goal of the k*-median* clustering problem is to find a set of centers (i.e. a clustering) $\{c_1,...,c_k\} \subseteq V$ such that the cost of clustering $\sum_{x \in D} \min_i d(x,c_i)$ is minimized. The goal of the k*-means* clustering problem is to find a clustering $\{c_1,...,c_k\} \subseteq V$ such that the cost $\sum_{x \in D} \min_i d^2(x,c_i)$ is minimized. Following the techniques of [24], we express clusterings as functions and refer to our full version for more details.

An (α, γ)-approximation algorithm that takes as input a set D and outputs clustering $\hat{c}_D$ guarantees that $\mathbb{E}_D[\hat{c}_D] \leq \alpha \cdot \mathbb{E}_D[c_D] + \gamma$, where c_D denotes the optimum clustering of D.

III. Private Sublinear time Approximate Clustering

In this section we describe the generic random sampling algorithm $\mathcal{A}'$ using a private (ξ,δ)-differentially private as a black-box, and in the sequel, we show that $\mathcal{A}'$ is (ξ',δ')-differentially private where ξ' and δ' are functions of ξ,δ (see Theorem 1). We present the accuracy results of $\mathcal{A}'$, i.e., the minimum sample size needed to guarantee that with high probability the approximate clustering cost of the sample S will be close to the true clustering cost of the input set D only for the case of k-median clustering (see Theorem 2, Theorem 3).

We remark that for the metric setting, both [10] and [24] consider clusterings where the centers are a subset of the set of input data points (also known as discrete clustering). By carefully conditioning on this requirement, [10] can make the sample complexity independent of n. Unfortunately, due to privacy concerns, we must consider the set of chosen centers to be *any* subset of the entire metric space, and not restricted to the input set (also known as continuous clustering). Thus we cannot hope to achieve a sample complexity independent of n in the metric setting, using their approach.

We present techniques used by [24] for our k-median clustering analysis in this section and describe the techniques used by [10] for our k-means clustering analysis in the full version.

A. Generic Algorithm $\mathcal{A}'$

We first present the basic sampling algorithm we employ, this model was first introduced in [24]. Note that the sampling probability ξ_1 should be chosen as $o(1)$.

Algorithm 1 General Sampling Scheme $\mathcal{A}'$

On input D,ξ_1
Sample each element of D independently w.p. ξ_1 and let S be the sample set.
Run (ξ,δ)-DP (α,γ)-approximation algorithm $\mathcal{A}$ on S to compute a set of private k-centers for S, denoted by C^*.
Output the clustering C^*.

B. Privacy of $\mathcal{A}'$

In this section we show that for an algorithm $\mathcal{A}'(D)$ which takes D as input and runs a (ξ,δ)-differentially private algorithm $\mathcal{A}$ on random sample $S \subseteq D$, it is the case that $\mathcal{A}'$ is (ξ',δ')-differentially private. Many works prove something similar to the following, e.g., [21], [5]. We refer to the full version for a proof.

Theorem 1. *If $\mathcal{A}$ is an (ξ, δ)-differentially private algorithm, and algorithm $\mathcal{A}'$ is the generic sampling*

algorithm defined above where each element is sampled independently with probability ξ_1, then $\mathcal{A}'$ is $(\xi',\ \delta')$-differentially private, where $\xi' = \ln\max\left\{\xi_1(e^{\xi}-1)+1,(\xi_1(e^{-\xi}-1)+1)^{-1}\right\}$*, and* $\delta' = \max\{\frac{e^{-\xi}\delta\xi_1}{(\xi_1(e^{-\xi}-1)+1)},\delta\xi\}$.

Observe that if $\mathcal{A}$ is (ξ,δ)-DP, then trivially, $\mathcal{A}'$ is also (ξ,δ)-DP. The privacy bounds achieved in the above theorem are significantly better than these naive bounds. For example, if we consider $\xi = 0.5, \xi_1 = 0.001$, for any $\delta \in [0,1)$, we achieve $\xi' < 0.00065$, and $\delta' = 0.001\delta$, which is orders of magnitude smaller than ξ and δ.

C. Private k-median in Metric Space

First, we state our general lemma which specifies the minimum sample complexity required to obtain a clustering in sublinear time using an (α,γ)-approximation clustering algorithm as a black-box, then we give a brief outline of the proof which closely follows [24].

Lemma 1. *Let* $\epsilon > 0, 0 < \delta < 1$ *and constant* $\alpha \geq 0.33$ *be approximation parameters. For an arbitrary metric space* (V,d) *of* n *points with diameter* M*, and a set of points* $D \subseteq V$ *assuming an* (α,γ)*-approximation* k*-median algorithm, that runs in time* $T(n)$*, we can draw a sample* $S \subseteq D$ *of size* s*,*

$$s = \Omega\left(\left(\frac{\alpha M}{\epsilon}\right)^2\left(k\ln n + \ln\frac{1}{\delta}\right)\right)$$

and obtain a k*-median clustering* $\hat{c}_S$ *in time* $T(|S|)$ *such that with probability at least* $1-\delta$*,* $\mathbb{E}_D[\hat{c}_S] \leq \alpha\mathbb{E}_D[c_D] + \gamma + \epsilon$*, where* c_D *represents the optimum clustering of* D*.*

Proof Outline. We first show that for the sample size specified in the above lemma, with high probability, the average cost of the optimum clustering c_D over the sample S and the entire input set D is close, i.e. $|\mathbb{E}_S[c_D] - \mathbb{E}_D[c_D]| < \frac{\epsilon}{4\alpha}$. To do this, one can model the k-median clusterings as k-median cost functions and show via Chernoff and union bounds that the empirical average value over the sample estimates the true average value for all functions, where the total number of possible clusterings, hence functions, is $O(n^k)$. Then we relate the average cost of the optimum clustering of S, say c_S, and the average cost of c_D over the sample S. Since c_S is the optimum clustering of S, its average cost will be at most that of the cost of c_D over S, in other words, $\mathbb{E}_S[c_S] \leq \mathbb{E}_S[c_D]$. By virtue of running $\mathcal{A}$ on sample S, we also have the guarantee that $\mathbb{E}_S[\hat{c}_S] \leq \alpha\mathbb{E}_S[c_S] + \gamma$. Lastly, in the same fashion, we show that for the sample size specified in the above lemma, with high probability, the average cost of the approximate clustering $\hat{c}_S$ over the sample S and the entire input set D is close, i.e., $|\mathbb{E}_D[\hat{c}_S] - \mathbb{E}_S[\hat{c}_S]| < \frac{\epsilon}{4\alpha}$. By combining all these relations we complete the proof.

We will use the algorithm in [17] as our black-box algorithm $\mathcal{A}$. By plugging in the approximation guarantees for $\mathcal{A}$ into our Lemma 1, we get the following accuracy guarantee for our algorithm $\mathcal{A}'$.

Theorem 2 (Accuracy of $\mathcal{A}'$)**.** *Let* $\epsilon > 0, 0 < \hat{\delta} < 1$ *be approximation parameters. For an arbitrary metric space* (V,d) *of* n *points with diameter* M*, and a private set of points* $D \subseteq V$*, given the* ξ*-DP* $(6, O(Mk^2\log^2(n/\xi)))$*-approximation* k*-median algorithm (from [17]), we have a* ξ'*-DP algorithm* $\mathcal{A}'$ *(as defined in Theorem 1) that can draw a sample* $S \subseteq D$ *of size* s*,*

$$s = \Omega\left(\left(\frac{M}{\epsilon}\right)^2\left(k\ln n + \ln\frac{1}{\hat{\delta}}\right)\right)$$

and obtain a k*-median clustering* $\hat{c}_S$ *such that with probability at least* $1-\hat{\delta}$*,* $\mathbb{E}_D[\hat{c}_S] \leq 6\mathbb{E}_D[c_D] + O(Mk^2\log^2(n/\xi)) + \epsilon$.

D. Private k-median in Euclidean Space

In this setting, we consider input set $D \subseteq \mathbb{R}^d$ and $|D| = n$. Note that the number of possible clusterings of D is now uncountably infinite. We apply techniques similar to the metric space setting, but in order to estimate the size of the set of clusterings, we adapt the approach of [24] and use ϵ-nets. Our general lemma is stated below.

Lemma 2. *Let* $\epsilon > 0, 0 < \delta < 1$*, and constant* $\alpha \geq 0.75$ *be approximation parameters. For* $D \subseteq \mathbb{R}^d$*, and diameter* M*, assuming an* (α,γ)*-approximation* k*-median algorithm that runs in time* $T(n)$*, we can draw a sample* S *of size* s *where*

$$s = \Omega\left(\left(\frac{M\alpha}{\epsilon}\right)^2\left(dk\ln\frac{\alpha dM}{\epsilon} + \ln\frac{1}{\delta}\right)\right),$$

and obtain a k*-median clustering* $\hat{c}_S$ *in time* $T(|S|)$ *such that with probability at least* $1-\delta$*,* $\mathbb{E}_D(\hat{c}_S) \leq \alpha\mathbb{E}_D(c_D) + \gamma + \epsilon$*, where* c_D *is the optimum* k*-median clustering of* D *in* $\mathbb{R}^d$*.*

We use the private clustering algorithm from [16] as our black-box algorithm $\mathcal{A}$, and obtain our sublinear-time result for $\mathcal{A}'$ by plugging in the approximation guarantees for [16] into Lemma 2 as follows.

Theorem 3 (Accuracy of $\mathcal{A}'$ for pure and approximate DP)**.** *Let $\epsilon > 0, 0 < \hat{\delta} < 1$, and constant α be approximation parameters. For private set $D \subseteq \mathbb{R}^d$ with diameter M, and an ξ-DP $\Big(w(1+\alpha),\ O\Big(\Big(\frac{kd+k^{O(1)}}{\xi}\Big)\text{poly log } n\Big)\Big)$-approximation k-median algorithm (from [16]), that runs in time $k^{O(1)}\text{poly}(nd)$, for $w(1+\alpha) \geq 0.75$, we have a ξ'-DP algorithm $\mathcal{A}'$ that can draw a sample $S \subseteq D$ of size s,*

$$s = \Omega\Big(\Big(\frac{Mw(1+\alpha)}{\epsilon}\Big)^2\Big(dk\ln\frac{w(1+\alpha)dM}{\epsilon} + \ln\frac{1}{\hat{\delta}}\Big)\Big)$$

and obtain a k-median clustering $\hat{c}_S$ in time $k^{O(1)}\text{poly}(sd)$ such that with probability at least $1-\hat{\delta}$, $\mathbb{E}_D[\hat{c}_S] \leq w(1+\alpha)\,\mathbb{E}_D[c_D] + O\Big(\Big(\frac{kd+k^{O(1)}}{\xi}\Big)\text{poly log } n\Big) + \epsilon$.

Moreover, by using the (ξ,δ)-DP algorithm from [16] with the same runtime and approximation ratio but with additive error $\gamma' := O\Big(\Big(\frac{k\sqrt{d}}{\xi}\cdot\text{poly log}\Big(\frac{k}{\delta}\Big)\Big) + \Big(\frac{k^{O(1)}}{\xi}\cdot\text{poly log } n\Big)\Big)$, we obtain a (ξ',δ')-DP algorithm $\mathcal{A}'$ that draws a sample of the same size, and obtains a k-median clustering such that with probability at least $1-\hat{\delta}$, $\mathbb{E}_D[\hat{c}_S] \leq w(1+\alpha)\mathbb{E}_D[c_D] + \gamma' + \epsilon$. Privacy parameters ξ',δ' are as defined in Theorem 1.

Note that the state-of-the-art non-private algorithm for k-median achieves an approximation ratio of $w = 2.633$ [1].

IV. Group Privacy in Sublinear setting

In this section, we give a group privacy result that holds for *any* sampling algorithm $\mathcal{A}'(D)$ that samples a set S from the input set D by independently sampling with probability ξ_1 and runs an ξ-DP algorithm $\mathcal{A}$ on S. Let D' be a set that differs on g elements with respect to D, and $0 \leq T \leq g$ be a threshold. Define $\delta_{T,\xi_1,g} := 1 - \sum_{j=0}^{T}\binom{g}{j}\xi_1^j(1-\xi_1)^{g-j}$, in other words, $\delta_{T,\xi_1,g}$ is the probability of choosing more than T elements that differ from elements in D' in the sample S.

Given that $\mathcal{A}$ is ξ-DP, we have already shown that $\mathcal{A}'$ is ξ'-DP (see Theorem 1). In the following theorem, we show that $\mathcal{A}'$ also gives us better group privacy guarantees.

Theorem 4. *If $\mathcal{A}'$ is an ξ'-DP sampling algorithm (as described above) then it gives $(T\cdot\xi', \delta_{T,\xi_1,g})$-privacy for groups of size g, where $\delta_{T,\xi_1,g} := 1 - \sum_{j=0}^{T}\binom{g}{j}\xi_1^j(1-\xi_1)^{g-j}$.*

We demonstrate how in many instances, our sampling algorithm $\mathcal{A}'$ achieves better group privacy guarantees for chosen ξ_1 and T such that $T \ll g$. (1) If we sample each element of the input set with probability $\xi_1 = 1/\sqrt{g}$, and set threshold $T = 2\sqrt{g}$, then $\mathcal{A}'$ is $(2\sqrt{g}\xi', \delta_{T,\xi_1,g})$ for $\delta_{T,\xi_1,g}$ negligible in g. (2) If we sample each element of the input set with probability $\xi_1 = 1/\log g$, and set threshold $T = 2g/\log g$, then $\mathcal{A}'$ is $((2g/\log g)\xi', \delta_{T,\xi_1,g})$ for $\delta_{T,\xi_1,g}$ negligible in g.

V. Acknowledgements

Elena would like to thank Marek Elias, Michael Kapralov and Aida Mousavifar for initial discussions on this topic while she was visiting EPFL. She also thanks her EPFL hosts for their hospitality.

REFERENCES

[1] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. *SIAM Journal on Computing*, 49, 2020.

[2] Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristic for k-median and facility location problems. STOC, 2001.

[3] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a ptas for k-median and k-means clustering. FOCS, 2010.

[4] Maria-Florina Balcan, Travis Dick, Yingyu Liang, Wenlong Mou, and Hongyang Zhang. Differentially private clustering in high-dimensional Euclidean spaces. ICML, 2017.

[5] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. NeurIPS, 2018.

[6] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. SOSP, 2017.

[7] Moses Charikar, Sudipto Guha, Éva Tardos, and David B Shmoys. A constant-factor approximation algorithm for the k-median problem. *Journal of Computer and System Sciences*, 65, 2002.

[8] Ke Chen. On k-median clustering in high dimensions. SODA, 2006.

[9] Ke Chen. A constant factor approximation algorithm for k-median clustering with outliers. SODA, 2008.

[10] Artur Czumaj and Christian Sohler. Sublinear-time approximation algorithms for clustering via random sampling. volume 30, 2007.

[11] Apple Differential Privacy Team. Learning with privacy at scale, 2017.

[12] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. volume 7, 2016.

[13] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. CCS, 2014.

[14] Dan Feldman, Amos Fiat, Haim Kaplan, and Kobbi Nissim. Private coresets. STOC, 2009.

[15] Dan Feldman, Chongyuan Xiang, Ruihao Zhu, and Daniela Rus. Coresets for differentially private k-means clustering and applications to privacy in mobile sensor networks. IPSN, 2017.

[16] Badih Ghazi, R. Kumar, and Pasin Manurangsi. Differentially private clustering: Tight approximation ratios. NeurIPS, 2020.

[17] Anupam Gupta, Katrina Ligett, Frank McSherry, Aaron Roth, and Kunal Talwar. Differentially private combinatorial optimization. 2010.

[18] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 291–300, 2004.

[19] Zhiyi Huang and Jinyan Liu. Optimal differentially private algorithms for k-means clustering. PODS, 2018.

[20] Haim Kaplan and Uri Stemmer. Differentially private k-means with constant multiplicative error. NeurIPS, 2018.

[21] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? *SIAM Journal on Computing*, 40, 2011.

[22] Daniel Kifer, Solomon Messing, Aaron Roth, Abhradeep Thakurta, and Danfeng Zhang. Guidelines for implementing and auditing differentially private systems. *CoRR*, abs/2002.04049, 2020.

[23] Shi Li and Ola Svensson. Approximating k-median via pseudo-approximation. *SIAM Journal on Computing*, 45, 2016.

[24] Nina Mishra, Dan Oblinger, and Leonard Pitt. Sublinear time approximate clustering. SODA, 2001.

[25] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. STOC, 2007.

[26] Kobbi Nissim and Uri Stemmer. Clustering algorithms for the centralized and local models. ALT, 2018.

[27] Rafail Ostrovsky, Yuval Rabani, Leonard J Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)*, 59, 2012.

[28] Moshe Shechner, Or Sheffet, and Uri Stemmer. Private k-means clustering with stability assumptions. AISTATS, 2020.

[29] Uri Stemmer. Locally private k-means clustering. SODA, 2020.

[30] Dong Su, Jianneng Cao, Ninghui Li, Elisa Bertino, and Hongxia Jin. Differentially private k-means clustering. CODASPY, 2016.