

Additive Noise Mechanisms for Making Randomized Approximation Algorithms Differentially Private

Jakub Tětek   
INSAIT, Sofia, Bulgaria

Abstract

The exponential increase in the amount of available data makes taking advantage of them without violating users' privacy one of the fundamental problems of computer science. This question has been investigated thoroughly under the framework of differential privacy. However, most of the literature has not focused on settings where the amount of data is so large that we are not even able to compute the exact answer in the non-private setting (such as in the streaming setting, sublinear-time setting, etc.). This can often make the use of differential privacy unfeasible in practice.

In this paper, we show a general approach for making Monte-Carlo randomized approximation algorithms differentially private. We only need to assume the error R of the approximation algorithm is sufficiently concentrated around 0 (e.g. $\mathbb{E}[|R|]$ is bounded) and that the function being approximated has a small global sensitivity Δ . Specifically, if we have a randomized approximation algorithm with sufficiently concentrated error which has time/space/query complexity $T(n, \rho)$ with ρ being an accuracy parameter, we can generally speaking get an algorithm with the same accuracy and complexity $T(n, \Theta(\epsilon\rho))$ that is ϵ -differentially private.

Our technical results are as follows. First, we show that if the error is subexponential, then the Laplace mechanism with error magnitude proportional to the sum of the global sensitivity Δ and the *subexponential diameter* of the error of the algorithm makes the algorithm differentially private. This is true even if the worst-case global sensitivity of the algorithm is large or infinite. We then introduce a new additive noise mechanism, which we call the zero-symmetric Pareto mechanism. We show that using this mechanism, we can make an algorithm differentially private even if we only assume a bound on the first absolute moment of the error $\mathbb{E}[|R|]$.

Finally, we use our results to give either the first known or improved sublinear-complexity differentially private algorithms for various problems. This includes results for frequency moments, estimating the average degree of a graph in sublinear time, rank queries, or estimating the size of the maximum matching. Our results raise many new questions and we state multiple open problems.

2012 ACM Subject Classification Security and privacy; Theory of computation \rightarrow Streaming, sublinear and near linear time algorithms

Keywords and phrases Differential privacy, Randomized approximation algorithms

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2024.73

Category RANDOM

Related Version *Full Version*: <https://arxiv.org/abs/2211.03695>

Funding This work was done while the author was at BARC, University of Copenhagen. He was supported by the VILLUM Foundation grant 54451. This research was partially funded from the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure).

Acknowledgements The author would like to thank Rasmus Pagh and anonymous reviewers for helping to improve this paper.



© Jakub Tětek;

licensed under Creative Commons License CC-BY 4.0

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2024).

Editors: Amit Kumar and Noga Ron-Zewi; Article No. 73; pp. 73:1–73:20



Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

With the increase in the amount of available data, the problem of analyzing it in a privacy-preserving manner has become a central problem in computer science. One commonly used tool for this task is differential privacy, which is a well-established notion of privacy that is commonly used in data analysis and machine learning. However, with some notable exceptions, the literature on differential privacy has focused on the setting where the amount of data is small enough that we would be able to practically solve a given problem exactly in the non-private setting. However, in practice, this assumption is often not realistic – this is after all the reason for the existence of (among others) streaming and sublinear-time algorithms.

Our goal is thus to get very efficient (sublinear) algorithms that at the same time guarantee differential privacy. In the streaming or sublinear-time setting, the error coming from the algorithm not being exact will generally speaking be much bigger than the amount of noise that the given problem necessitates for ensuring differential privacy. The main objective in this setting is thus not to simply minimize the amount of noise we add, but rather to achieve a given level of accuracy while minimizing the complexity (e.g. space, time, or query complexity) of the algorithm. Of course, some amount of noise inherently has to be added to achieve privacy, but this is usually so small, that one would need linear complexity to get such a level of accuracy even without privacy. In the sublinear regime, we thus usually do not have to worry whether a given level of accuracy is achievable and we instead focus as our central objective on the complexity needed to achieve it.

One of the main difficulties in making sublinear algorithms private is that most sublinear-time and streaming algorithms are randomized and give only probabilistic guarantees on the quality of the output. This makes adding noise based on global sensitivity¹ – which is commonly used to get differentially private algorithms – unsuitable for this situation, as in the worst case the global sensitivity of the approximation algorithm² can be very large even if the global sensitivity of the function being approximated is small. In this paper, we propose a way to get around this issue by showing additive noise mechanisms that only need that (1) the function being approximated has low global sensitivity and (2) the answer of the algorithm is sufficiently concentrated around the true value.

We give two main results. The first one states that if the error has subexponential tails³, adding the Laplace distribution suffices to achieve pure differential privacy – this can be seen as a generalization of the standard Laplace mechanism. The second result shows a similar result for a different distribution, and it only assumes bounded mean deviation (or higher moments) of the error instead of being subexponential. While the first result has much stronger assumption about the error distribution, it is also stronger in that it also works for multiple adaptive queries that are *not answered independently*. This is useful for example for streaming algorithms, where multiple queries can be answered using a single sketch and are thus not answered independently. Note that the standard composition theorem would allow us to perform multiple queries only if they were answered independently.

We use our results to give new differentially private algorithms for various problems: for maximum matching under node-level privacy, frequency moments, counting connected components under edge-level privacy, and rank queries. We also show how a common

¹ Global sensitivity of a function g with respect to a relation \sim is defined as $\sup_{x \sim x'} |g(x) - g(x')|$.

² Here, we see the approximation algorithm as a deterministic function of the input and a string of random bits.

³ A distribution is subexponential if its tails are dominated by the tail of an exponential distribution.

technique for designing relative-approximation sublinear-time algorithms – *advice removal by geometric search* – can be made differentially private. This implies an edge-differentially-private algorithm for estimating the average degree of a graph, improving upon the state of the art [5], but we think it could also be useful for many other problems. Our algorithm for maximum matching also answers an open problem from [5]. For the other mentioned problems, we give the most efficient differentially private algorithm known.

Notably, a concurrent work [7] gave a similar result. In the context of global sensitivity, their result makes weaker assumptions and results in algorithms that are less efficient since they rely on postprocessing to turn approximate differential privacy into pure differential privacy. We discuss this in detail in Section 1.3.

1.1 Our results

We now informally state our main technical results and then we state the results for specific results that we obtained using the technical results. We also show how to use our technical results to give algorithms for specific problems in Section 1.2. We start with a result for algorithms with subexponential error tails, which appears in Section 3:

► **Theorem 9, simplified version.** Assume we have an algorithm $A(D)$ for D being a dataset. Assume there exists a function g with global sensitivity $\leq \Delta_1$ w.r.t. D such that $A(D) - g(D)$ has *subexponential diameter*⁴ $\leq \Delta_2$, i.e. $\mathbb{P}[|A(D) - g(D)| \geq t] \leq 2e^{-t/\Delta_2}$ for $t \geq 0$. Then releasing $A(D) + \text{Laplace}(O((\Delta_1 + \Delta_2)/\epsilon))$ is ϵ -differentially private for $\epsilon \leq O(1)$.

Moreover, with noise $\text{Laplace}(O(k(\Delta_1 + \Delta_2)/\epsilon))$, this also holds if we make k such releases with different algorithms A_1, \dots, A_k chosen adaptively that are executed with the same randomness.

Note that this generalizes the claim that the Laplace mechanism with noise magnitude proportional to the global sensitivity gives differential privacy (this can be seen by setting $A(D) = g(D)$). Note also that in the second half, the algorithms use the same randomness, and we thus cannot get this part of the result by standard composition.

We then prove in Section 4 that in the case of a single query, it is sufficient to assume a bound on some deviation moments (multiple independently answered queries can be handled using the standard composition theorem).

► **Theorem 13, simplified version.** *Let us have an algorithm A such that there exists a function g with global sensitivity $\leq \Delta$ and such that $\mathbb{E}[|A(D) - g(D)|^3] \leq \Delta^3$ for any dataset D . Then for $\epsilon \leq O(1)$ there exists a random variable Y with $\mathbb{E}[|Y|] \leq O(\Delta/\epsilon)$, such that $A(\cdot) + Y$ is ϵ -differentially private.*

We now state results which, as we show in Section 5, follow from the above technical results. For all the following results, our algorithm is the most efficient known, except for the streaming algorithm for rank queries, which is incomparable with the algorithm from [19]. With the result for maximum matching, we (together with the concurrent [7]) answer positively the open question of Blocki, Grigorescu, and Mukherjee [5], and the algorithm for estimating the number of edges improves upon the algorithm from [5]. For more detailed comparison with [7], see Section 1.3.

⁴ The subexponential diameter of a distribution roughly corresponds to the smallest parameter σ , such that the tail of $\text{Laplace}(1/\sigma)$ dominates the tails of the distribution.

► **Corollaries 16–20, simplified versions.** *There exists an ϵ -differentially private*

1. *streaming algorithm for the frequency F_2 moment problem with space complexity $O(\frac{1}{\rho^2 \epsilon^2})$ (Corollary 16),*
2. *sublinear-time algorithm for estimating the number of connected components with time complexity $O(\frac{1}{\rho^2 \epsilon^2} \log \frac{1}{\rho \epsilon})$ (Corollary 17),*
3. *sublinear-time algorithm for estimating the size of the maximum matching with time complexity $d^{O(1/(\rho^2 \epsilon^2))} / (\rho \epsilon)^{O(1/(\rho \epsilon))}$ (Corollary 18),*
4. *streaming algorithm for answering k (adaptive) rank queries in space $O(\frac{k \log^2 k}{\rho \epsilon})$ (Corollary 19),*
5. *sublinear-time algorithm for approximating the number of edges in a graph in time $O(\frac{n}{\epsilon^2 \rho^2 \sqrt{m}})$ (Corollary 20).*

1.2 Our techniques

We now sketch the techniques that we use in our paper. In each part of this section, we also give a reference to the corresponding section of this paper.

1.2.1 Subexponential error, one query (Section 3)

Suppose we have a randomized algorithm $A(D)$ for D being a dataset that approximates a function $g(D)$ with global sensitivity $\leq \Delta_1$ for some parameter Δ_1 . Define the error R as the random variable $R = A(D) - g(D)$ and assume that it is tightly concentrated around 0, namely $\mathbb{P}[|R| > t] \leq 2e^{-t/\Delta_2}$ for some value Δ_2 (Δ_2 is an upper bound on the “subexponential diameter” of R). Intuitively speaking, Δ_2 determines the “scale” of R , and Δ_2 is in fact up to a constant factor an upper bound on $\mathbb{E}[|R|]$. Note that the tails of R decrease at least at the same rate as those of the Laplace distribution. This suggests Laplacian noise with large enough magnitude could “hide R ”. Indeed, we prove that Laplacian noise will guarantee privacy. We now sketch the proof.

High-level view. Assume for simplicity that both $\Delta_1, \Delta_2 \leq 1$. The same approach works for general values of Δ_1, Δ_2 by simple re-scaling. Let $Y \sim \text{Laplace}(c/\epsilon)$ for appropriately chosen value of c . We will prove that for any random variable X with subexponential diameter ≤ 3 ⁵ (that is $\mathbb{P}[|X| \geq t] \leq 2e^{-t/3}$), the probability density functions satisfy $f_{X+Y}(y)/f_Y(y) = e^{\pm\Theta(\epsilon)}$ for any y . We can use this to prove privacy, as we now show. For two neighboring datasets D_1, D_2 , we set $R_1 = A(D_1) - g(D_1)$ and $R_2 = A(D_2) - g(D_1)$ (note the asymmetry in the definitions). It then holds that R_1 has subexponential diameter $\Delta_2 \leq 1$. One can also show that the subexponential diameter of $R_2 = (A(D_2) - g(D_2)) + (g(D_2) - g(D_1))$ is ≤ 3 (Lemma 5). Let $y' = y - g(D_1)$. It then holds for any y that

$$\frac{f_{A(D_1)+Y}(y)}{f_{A(D_2)+Y}(y)} = \frac{f_{g(D_1)+R_1+Y}(y)}{f_{g(D_1)+R_2+Y}(y)} = \frac{f_{R_1+Y}(y')}{f_{R_2+Y}(y')} = \frac{f_{R_1+Y}(y')}{f_Y(y')} \cdot \frac{f_Y(y')}{f_{R_2+Y}(y')} = e^{\pm\Theta(\epsilon)}$$

where the last equality uses the claim $f_{X+Y}(y')/f_Y(y') = e^{\pm\Theta(\epsilon)}$ for $X = R_1$ and for $X = R_2$. This implies differential privacy.

⁵ We choose value 3 as this is the value we will need below. The claim holds also for larger constants with c chosen appropriately.

Bounding ratios of density functions. We now sketch why $f_{X+Y}(y)/f_Y(y) = e^{\pm\Theta(\epsilon)}$. Since Y is continuous, we may re-write

$$f_{X+Y}(y) = \mathbb{E}[f_Y(y - X)] = \frac{\epsilon}{2c} \mathbb{E}[e^{-\epsilon|X-y|/c}] = (*)$$

where the first equality is a standard identity [16]. We now use the inequalities $|X - y| \leq |X| + |y|$ and $|X - y| \geq |y| - |X|$. This allows to bound

$$\begin{aligned} (*) &\leq \frac{\epsilon}{2c} \mathbb{E}[e^{-\epsilon(|y|-|X|)/c}] = \frac{\epsilon}{2c} e^{-\epsilon|y|/c} \mathbb{E}[e^{\epsilon|X|/c}] \\ (*) &\geq \frac{\epsilon}{2c} \mathbb{E}[e^{-\epsilon(|X|+|y|)/c}] = \frac{\epsilon}{2c} e^{-\epsilon|y|/c} \mathbb{E}[e^{-\epsilon|X|/c}] \end{aligned}$$

while it holds that $f_Y(y) = \frac{\epsilon}{2c} e^{-\epsilon|y|/c}$. It is thus sufficient to prove that $\mathbb{E}[e^{\epsilon|X|/c}] \leq e^{O(\epsilon)}$ and $\mathbb{E}[e^{-\epsilon|X|/c}] \geq e^{-O(\epsilon)}$. While the first inequality is standard, the second is not. We will now sketch a proof for both.

Bounding the expectation of exponentials of a subexponential random variable. If we knew the density function of X , we could easily express the expectations as integrals. However, not only we do not have a bound on the density, but the density may even not exist. We thus use the following trick. We use the fact that for any real random variable Z , it holds that Z has the same distribution as $F_Z^{-1}(u)$ for $u \sim \text{Unif}(0, 1)$ where F_Z is the cumulative distribution function (CDF) of Z . This allows us to write $\mathbb{E}[e^{-\epsilon|X|/c}] = E_u[e^{-F_{\epsilon|X|/c}^{-1}(u)}]$ and similarly for $\mathbb{E}[e^{\epsilon|X|/c}]$.

Unlike the density function, we do have a bound on the cumulative distribution function. Specifically, we are assuming $\mathbb{P}[|X| > t] \leq 2e^{-t/3}$ which implies that $F_{\epsilon|X|/c}^{-1}(u) \leq -\frac{3\epsilon}{c} \log(\frac{1-u}{2})$. Upper-bounding the CDF like this reduces the problem to computing the expectation of a function of the uniform random variable, which can be done straightforwardly. This proves the desired bounds.

1.2.2 Subexponential error, multiple queries (Section 3)

We would like to be able to release answers to multiple queries which are not answered independently (such as if they are answered based on the same sketch). Since the answers are not independent, we cannot use the composition theorem. We now sketch an alternative approach.

For fixed queries, the above proof goes through with minor modifications even in the multivariate case. Instead of using the inequalities $|y - X| \leq |y| + |X|$ and $|y - X| \geq |y| - |X|$ in the case of $y, X \in \mathbb{R}$, we use the analogous version for ℓ_1 norms in the case of $y, X \in \mathbb{R}^k$: $\|y - X\|_1 \leq \|y\|_1 + \|X\|_1$ and $\|y - X\|_1 \geq \|y\|_1 - \|X\|_1$. We then use that if X is a vector of k subexponential random variables with diameter $\leq \Delta$, then $\|X\|_1$ has subexponential diameter $\leq 3k\Delta$ (Lemma 5).

This however gives the result only in the nonadaptive case, when the queries do not depend on the released values: the identity $f_{X+Y}(y) = \mathbb{E}[f_Y(X - y)]$ relies crucially on $X = (X_1, \dots, X_k)$ and $Y = (Y_1, \dots, Y_k)$ being independent. In the case of adaptive queries, X_i could depend on Y_1, \dots, Y_{i-1} (the query that we perform – and thus also the answer to it – can be influenced by the noise we add to the previous answers). We instead use our non-adaptive version of the claim and make it adaptive in a black box fashion, by proving a claim that may be of independent interest: If we have a countable number of mechanisms such that releasing the answers of any *fixed* subset of size k is ϵ -differentially private, then we may also pick this subset *adaptively* and it will still be ϵ -differentially private.

1.2.3 Error with polynomial tails (Section 4)

In the case that the error has polynomial tails, we only consider the case of a single query. Our techniques do not seem to generalize to the multivariate case, and we conjecture that this is impossible (see Section 6). The case when multiple queries are answered independently may be still handled by the standard composition theorem.

An approach similar to the one described above can be made to work, with the difference that we use the inequality $|x - X| \geq \max(0, |x| - |X|)$ instead of the weaker $|x - X| \geq |x| - |X|$. This approach, however, requires proving the following inequality for all $y, s \geq 0, \alpha > 1, 0 \leq \epsilon \leq 1$:

$$\int_0^1 \min \left((1 + |y|/s)^\alpha, \left| 1 - \frac{(1 - 2^{-1/\alpha})\epsilon}{(1 + |y|/s)(1 - u)^{1/\alpha}} \right|^{-\alpha} \right) du \leq 1 + \frac{2\alpha - 1}{\alpha - 1} \epsilon.$$

This is the technically most challenging part of this paper. The trick is to bound the inside of the integral for $u \in [0, 1 - \epsilon(1 + |y|/s)^{-\alpha}]$ by a simpler expression that can be successfully integrated. The rest of the interval $[0, 1]$ contributes at most ϵ , as its length is $\epsilon(1 + |y|/s)^{-\alpha}$ and the maximum value of the function being integrated is $\leq (1 + |y|/s)^\alpha$.

1.2.4 How to use our technical results (Section 5)

1.2.4.1 Error with subexponential tails

We now sketch how one can use Theorem 9 to get differentially private mechanisms for specific problems. Recall that the theorem states that if our algorithm approximates a function with global sensitivity $\leq \Delta_1$ and the subexponential diameter of the error is $\leq \Delta_2$, then adding noise from $Laplace(O(\Delta_1 + \Delta_2)/\epsilon)$ makes the algorithm ϵ -differentially private.

We start with a randomized approximation algorithm whose error is subexponentially concentrated around zero (often, this is either known or easy to prove) that approximates a parameter with a small global sensitivity Δ_1 . This is the case for example for the YYI maximum matching algorithm [34] under node-level privacy or the KLL sketch [20] for rank queries. Suppose the complexity (such as time/space/query/sample or other complexity) of the algorithm is $T(n, \rho)$ and has additive error of scale (i.e. with subexponential diameter) ρn . If we want the final error with privacy to be $O(\rho n)$, then we run the algorithm with error parameter $\epsilon \rho$, making the error's subexponential diameter be $O(\epsilon \rho n)$. We then get from Theorem 9 that adding noise of magnitude $O(\rho n)$ is sufficient to get ϵ -differential privacy, assuming Δ_1 is sufficiently small, giving us the desired result. The complexity of the private algorithm will thus be $T(n, \Theta(\epsilon \rho))$. This allows us to achieve a given level of accuracy⁶ under pure differential privacy, while not significantly worsening the algorithm's complexity. We describe this approach in greater detail and with general failure probabilities (not just constant) in Section 5.

To illustrate the second part of the theorem, consider for example a rank queries sketch (see Section 5.5 for details). The algorithms A_1, \dots, A_k correspond to making k adaptive queries to the sketch of a dataset (assume we are given k queries we need to perform) and the fact that the algorithms use the same randomness corresponds to us querying the same sketch (as compared to k independent sketches). Specifically, the algorithm A_i here builds

⁶ This is true unless ρ is very small, as otherwise the global sensitivity will necessitate some level of noise. As we noted, this usually happens only when $T(n, \epsilon \rho) \geq \Omega(n)$, making this uninteresting for our sublinear setting.

the sketch (using the shared randomness), and then performs the i -th query. Note that the fact that we only use one sketch *prevents us from using the composition theorem* to get this from just the first part of the theorem.

1.2.4.2 Error with polynomial tails

If the error has polynomial tails, we can make it private in a way essentially the same as in the subexponential case using Theorem 13. Moreover, it holds that if we have $\mathbb{E}[|A(D) - g(D)|] \leq \Delta$, then by taking a median of 5 independent executions of A , we get an algorithm A' whose error's third moment is also bounded: $\mathbb{E}[|A'(D) - g(D)|^3] \leq O(\Delta^3)$ [21]. This means that we can make an algorithm private even if we only have a bound on $\mathbb{E}[|A(D) - g(D)|]$, for g being a low-global-sensitivity function, or the mean squared error $\mathbb{E}[(A(D) - g(D))^2]$.

1.3 Related work

To the best of our knowledge, the work on differentially private approximation algorithms started with private sketches. Mir, Muthukrishnan, Nikolov, and Wright [25] gave pan-private⁷ sketches for heavy hitters. An improved sketch has been recently given by Pagh and Thorup [27]. A private version of the deterministic Misra-Gries sketch [26] for heavy hitters has been recently given by Tětek and Lebeda [22]. Heavy hitters were also investigated in the multi-party computation setting [18], in the local differential privacy setting [3], and using cryptographic assumptions [24, 17, 18].

A sketch for fractional frequency moments F_p for $0 \leq p \leq 1$ has been given by Wang, Pinelis, and Song [33]. After releasing this paper, Epasto, Mao, Andres, Mirrokni, Vassilvitskii, and Zhong [15] have given an algorithm general value of p in the continual release setting. A sketch for differentially private quantiles has been given by Alabi, Ben-Eliezer, and Chaturvedi [1]. A technique for stream sanitization has been given by Kaplan and Stemmer [19]; this work resulted in improved differentially private sketches for approximate quantiles. An approach for differentially privately estimating distances in euclidean spaces using private sketches has been given by Stausholm [31]. A general approach to making linear sketches differentially private was given by Zhao, Qiao, Redberg, Agrawal, Abbadi, and Wang [35].

A recent line of work has shown that many sketches already provide privacy *by themselves* or with only small modifications, without adding any noise. Blocki, Blum, Datta, and Sheffet [6] have shown that the Johnson-Lindenstrauss transform by itself ensures differential privacy. Smith, Song, and Guha Thakurta [30] have shown that this is also the case for the Flajolet-Martin sketch for counting distinct elements and a similar result is known for the LogLog algorithm [9, 11]. This was recently generalized [10] to show that this is not only the case for the two above-mentioned sketches, but in fact for a large class of sketches for counting distinct elements.

As far as sublinear-time algorithms are concerned, Sivasubramaniam, Li, and He [29] have shown a differentially private algorithm that returns a $2 + \rho$ approximation of the number of edges in a graph in time $\tilde{O}_{\rho, \epsilon}(\sqrt{n})$. This has been later improved by Blocki, Grigorescu, and Mukherjee [5] to $1 + \rho$ approximation in the same complexity. In that paper, the authors also give differentially private sublinear algorithms for approximate maximum matching and vertex cover. A sublinear time algorithm for estimating the median was recently discovered [8].

⁷ An algorithm on an input stream is said to be pan-private if releasing the internal state of the algorithm at any point in the computation is differentially private. It is a strictly stronger notation than differential privacy of the output.

Concurrent work

The problem of differentially private randomized approximation algorithms has been explored independently of this work by Blocki, Grigorescu, Mukherjee, and Zhou [7]. The techniques used in [7] differ significantly from those used in this paper. Specifically, in [7], the authors set the failure probability of the algorithm to be $\leq \delta$ (for example by probability amplification), thus limiting the global sensitivity of the algorithm up to an event of probability $\leq \delta$. This allows them to rely on the standard result for getting differential privacy based on global sensitivity⁸. Specifically, if the algorithm has complexity $T(n, \rho)$ and one uses probability amplification, then the approach of [7] gives an (ϵ, δ) -differential privacy in complexity $O(T(n, \epsilon\rho) \log \delta^{-1})$. They then show that one can post-process the output of the algorithm to achieve pure differential privacy.

In this paper, instead of relying just on global sensitivity, we instead prove privacy from first principles. At the cost of assuming that the error is sufficiently concentrated, we show that the probability amplification step is not needed, allowing us to get ϵ -differential privacy in the better complexity of $T(n, \epsilon\rho)$. Our approach allows us to give more efficient algorithms than Blocki, Grigorescu, Mukherjee, and Zhou [7] for several problems: estimating the average degree of a graph, estimating the size of a maximum matching, and estimating the number of connected components.

2 Preliminaries

2.1 Differential privacy

Throughout the paper, we assume that we have a symmetric “neighbor” relation \sim on the set of all possible datasets. Intuitively speaking, in the case when we have a database of users, this should correspond to two datasets being the same except for the data of one user whose privacy we are trying to protect.

► **Definition 1** ([12]). *A randomized algorithm M with range S is ϵ -differentially private if for any measurable $T \subseteq S$, it holds for any $x \sim x'$ for a symmetric “neighbor” relation \sim , that*

$$e^{-\epsilon} \leq \frac{\mathbb{P}[M(x) \in T]}{\mathbb{P}[M(x') \in T]} \leq e^{\epsilon}$$

This definition is commonly relaxed to a notion called approximate differential privacy, with the above notion then being called pure privacy. In this paper, we will focus only on pure differential privacy.

If the output of M is a continuous random variable, then it is sufficient to prove that for any y and $x \sim x'$ it holds $e^{-\epsilon} \leq f_{M(x)}(y)/f_{M(x')}(y) \leq e^{\epsilon}$, where f_X for X being a continuous random variable is the probability density function of X .

The global sensitivity [13] of a function g is defined as

$$\sup_{x \sim x'} |g(x) - g(x')|.$$

The authors have shown that if g has global sensitivity Δ , then adding $\text{Laplace}(\Delta/\epsilon)$ provides ϵ -differential privacy.

⁸ They also consider functions with low smoothed sensitivity instead of just low global sensitivity; we do not consider that in this paper.

In the context of graph problems, one often speaks of a mechanism being edge-differentially private, or node-differentially private. These terms refer to the relation \sim that is used. In the case of node-differential privacy, we have $G \sim G'$ iff one can get G from G' by deleting one vertex and the incident edges, or the other way around. In the case of edge-differential privacy, we have $G \sim G'$ iff one can get G from G' by deleting one edge, or the other way around.

2.2 Probability theory

If D is a distribution, we use $D^{\otimes k}$ for k being a natural number, to denote the k -fold product distribution of D . For a random variable Z , we denote by F_Z its cumulative distribution function. We denote by $F_Z^{-1}(p) = \inf\{x \in \mathbb{R} : F_Z(x) \geq p\}$ its generalized inverse. It holds that $F_Z^{-1}(u)$ has the same distribution as Z for $u \sim \text{Unif}(0, 1)$ [23]. We will need the following claim.

► **Fact 2** ([16]). *Let X, Y be independent random variables in \mathbb{R}^k , and assume Y has a probability density function (pdf) $f_Y(z)$. Then the pdf of $X + Y$ is $f_{X+Y}(z) = \mathbb{E}[f_Y(z - X)]$.*

Concentration of measure

The notions of *subexponential* random variables and *subexponential diameter* $\sigma_{se}[X]$ are central to concentration of measure. There are several different definitions for $\sigma_{se}[X]$, that differ by constant factors. See for example [32, Chapter 2] for an exposition of the various definitions. In this paper, one of the definitions is especially suitable for the way we use it in our proofs, and that is the definition that we use.

► **Definition 3.** *Let X be a real random variable. We define the subexponential diameter of X , denoted by $\sigma_{se}[X]$ as the smallest values for which for any $t > 0$ holds*

$$\mathbb{P}[|X| \geq t] \leq 2 \exp(-t/\sigma_{se}[X])$$

A random variable X is said to be subexponential if $\sigma_{se}[X] < \infty$.

It holds that $\sigma_{se}[cX] = c\sigma_{se}[X]$. It also holds that

► **Fact 4.** *For X being a random variable and $c \geq 0$, it holds that $\sigma_{se}[X+c] \leq \sigma_{se}[X] + c/\log 2$.*

Proof. We can bound $\mathbb{P}[X + c \geq t] \leq \min(1, 2e^{-(t-c)/\sigma_{se}[X]}) \leq \min(1, 2e^{-t/(\sigma_{se}[X]+c/\log 2)})$, thus implying the claim. ◀

Finally, the following is a standard claim, but we need the constant factor, which is specific to the definition of σ_{se} that we are using. We thus give a proof in the full version of this paper, based on the standard proof of triangle inequality for Orlicz norms.

► **Lemma 5.** *Let us have real random variables X_1, \dots, X_k . It holds*

$$\sigma_{se}\left[\sum_{i=1}^k X_i\right] \leq 3 \sum_{i=1}^k \sigma_{se}[X_i]$$

3 Algorithms with subexponential error

In this section, we show how algorithms, whose error has a small subexponential diameter, can be made differentially private. We start by proving a technical lemma. We will later bound the ratios between probability densities of our mechanism's answers by the exponential expectations that we now bound and, finally, we will use that to prove privacy in Theorem 9, which is the main theorem of this section.

Note that while the second of the two inequalities is standard, the first one is not. Our proof does not follow the strategy of the standard proof of the second inequality, which is based on a Taylor expansion and does not seem to straightforwardly apply to the first inequality.

► **Lemma 6.** *Suppose X is a random variable with subexponential diameter $\Delta \leq 1/2$. It holds $\mathbb{E}[e^{-|X|}] \geq \frac{2^{-\Delta}}{1+\Delta} \geq e^{-(1+\log 2)\Delta}$ and $\mathbb{E}[e^{|X|}] \leq \frac{2^\Delta}{1-\Delta} \leq e^{3\log(2)\Delta}$.*

Proof. Since X is subexponential with diameter Δ , it holds that $\mathbb{P}[|X| \geq z] \leq 2e^{-z/\Delta}$. Therefore, $F_{|X|}^{-1}(u) \leq -\Delta \log(\frac{1-u}{2})$. We use the fact that for $u \sim \text{Unif}(0, 1)$, the random variable $F_{|X|}^{-1}(u)$ has the same distribution as $|X|$. We can now bound

$$\begin{aligned}
 \mathbb{E}[e^{-|X|}] &= E_u[e^{-F_{|X|}^{-1}(u)}] \\
 &\geq E_u[e^{\Delta \log(\frac{1-u}{2})}] \\
 &= 2^{-\Delta} E_u[(1-u)^\Delta] \\
 &= 2^{-\Delta} \int_0^1 (1-u)^\Delta du \\
 &= 2^{-\Delta} \left[-\frac{(1-u)^{\Delta+1}}{\Delta+1} \right]_{u=0}^1 \\
 &= \frac{2^{-\Delta}}{1+\Delta} \geq e^{-(1+\log 2)\Delta}
 \end{aligned} \tag{1}$$

where the last inequality holds because we can equivalently write $2^{-\Delta}/(\Delta+1) \geq (2e)^{-\Delta}$, which simplifies to $e^\Delta \geq 1+\Delta$, which is a standard inequality. Similarly, we can bound $\mathbb{E}[e^{|X|}]$ as follows.

$$\begin{aligned}
 \mathbb{E}[e^{|X|}] &= E_u[e^{F_{|X|}^{-1}(u)}] \\
 &\leq E_u[e^{-\Delta \log(\frac{1-u}{2})}] \\
 &= \frac{2^\Delta}{1-\Delta} \leq e^{3\log(2)\Delta}
 \end{aligned}$$

where the second equality is by substituting Δ with $-\Delta$ in (1) (since as we have shown, (1) is equal to $2^{-\Delta}/(1+\Delta)$). We have here used that $\Delta < 1$ (otherwise the final expression may not be defined). The last inequality can be shown as follows: we take the ratio of the two sides resulting in $h(\Delta) = 4^\Delta(1-\Delta)$ and we show $h(\Delta) \geq 1$ for $0 \leq \Delta \leq 1/2$. The function h is concave (the second derivative is $-2 \cdot 2^{2\Delta} \log 4 + 2^{2\Delta} \log^2 4$, which can be easily seen to be negative), meaning that it is sufficient to check that the inequality holds at the endpoints of the interval $[0, 1/2]$: that $h(0) \geq 1$ and $h(1/2) \geq 1$. One can easily check this holds. ◀

We are now ready to prove a lemma about the ratio of the density function of a Laplace and of Laplace shifted by a subexponential random variable. We will then use this to prove differential privacy. Note that the random variables in the lemma do not have to be independent.

► **Lemma 7.** *Let X_1, \dots, X_k be random variables with subexponential diameter at most Δ and let $X = (X_1, \dots, X_k)$. Let $Y \sim \text{Lap}^{\otimes k}(k\Delta/\epsilon)$ for $\epsilon \leq 1/6$. Consider $y \in \mathbb{R}^k$. It holds $e^{-3(1+\log 2)\epsilon} \leq f_{X+Y}(y)/f_Y(y) \leq e^{9\log(2)\epsilon}$. Moreover, if $k = 1$ and $\epsilon \leq 1/2$, it holds $e^{-(1+\log 2)\epsilon} \leq f_{X+Y}(y)/f_Y(x) \leq e^{3\log(2)\epsilon}$.*

Proof. By Fact 2, we have

$$f_{X+Y}(y) = \mathbb{E}[f_Y(y - X)] = \left(\frac{\epsilon}{2k\Delta}\right)^k \mathbb{E}[\exp(-\frac{\epsilon}{k\Delta}\|X - y\|_1)]$$

For the sake of brevity, we let $\gamma = \left(\frac{\epsilon}{2k\Delta}\right)^k$. We may bound $\|X - y\|_1 \leq \|X\|_1 + \|y\|_1$ and $\|X - y\|_1 \geq \|y\|_1 - \|X\|_1$. This allows us to bound

$$\begin{aligned} f_{X+Y}(y) &= \gamma E \left[\exp\left(\frac{-\epsilon\|X - y\|_1}{k\Delta}\right) \right] \\ &\geq \gamma E \left[\exp\left(\frac{-\epsilon(\|X\|_1 + \|y\|_1)}{k\Delta}\right) \right] \\ &= \gamma \exp\left(\frac{-\epsilon\|y\|_1}{k\Delta}\right) E \left[\exp\left(\frac{-\epsilon\|X\|_1}{k\Delta}\right) \right] \\ &\geq \gamma \exp\left(\frac{-\epsilon\|y\|_1}{k\Delta}\right) \exp(-3(1 + \log 2)\epsilon) \end{aligned} \tag{2}$$

where the last inequality is by Lemma 6; we used that $\frac{\epsilon\|X\|_1}{k\Delta}$ has subexponential diameter $\leq 3\epsilon \leq 1/2$, since we have by Lemma 5 that $\sigma_{se}[\|X\|] \leq 3k\Delta$. In the case $k = 1$, we simply have that $\sigma_{se}[\frac{\epsilon\|X\|_1}{k\Delta}] \leq \epsilon$, in which case the final bound on (2) is

$$\geq \gamma \exp\left(\frac{-\epsilon\|y\|_1}{k\Delta}\right) \exp(-(1 + \log 2)\epsilon)$$

At the same time, $f_Y(y) = \gamma \exp(-\frac{\epsilon\|y\|_1}{k\Delta})$ and thus

$$\begin{aligned} \frac{f_{X+Y}(y)}{f_Y(y)} &\geq e^{-3(1+\log 2)\epsilon} && \text{if } k > 1 \\ \frac{f_{X+Y}(y)}{f_Y(y)} &\geq e^{-(1+\log 2)\epsilon} && \text{if } k = 1 \end{aligned}$$

Similarly, we can bound

$$\begin{aligned} f_{X+Y}(y) &= \gamma E \left[\exp\left(\frac{-\epsilon\|X - y\|_1}{k\Delta}\right) \right] \\ &\leq \gamma E \left[\exp\left(\frac{-\epsilon(\|y\|_1 - \|X\|_1)}{k\Delta}\right) \right] \\ &= \gamma \exp\left(\frac{-\epsilon\|y\|_1}{k\Delta}\right) E \left[\exp\left(\frac{\epsilon\|X\|_1}{k\Delta}\right) \right] \\ &\leq \gamma \exp\left(\frac{-\epsilon\|y\|_1}{k\Delta}\right) \exp(9\log(2)\epsilon) \end{aligned}$$

in the case $k > 1$ and

$$f_{X+Y}(y) \leq \gamma \exp\left(\frac{-\epsilon\|y\|_1}{k\Delta}\right) \exp(3\log(2)\epsilon)$$

in the case $k = 1$. Thus, we have

$$\begin{aligned} \frac{f_{X+Y}(y)}{f_Y(y)} &\leq e^{9 \log(2)\epsilon} && \text{if } k > 1 \\ \frac{f_{X+Y}(y)}{f_Y(y)} &\leq e^{3 \log(2)\epsilon} && \text{if } k = 1 \end{aligned} \quad \blacktriangleleft$$

We now state a useful technical lemma; the proof appears in the full version. This lemma states that in general, if we have a countable number of mechanisms and releasing any fixed k of them is differentially private, then picking the mechanisms adaptively will not violate differential privacy. The proof roughly follows the outline of the proof of adaptive composition [28]. In what follows, we again use the subscript \cdot_r to denote a random bitstring used as the source of randomness of the mechanisms.

► **Lemma 8.** *Let us have a countable number of mechanisms $M_{1,r}, \dots$, such that releasing the value $(M_{i_1,r}(D), \dots, M_{i_k,r}(D))$ is ϵ -differentially private for any fixed $i_1, \dots, i_k \in [n]$. Then the mechanism $(M_{j_1,r}(D), \dots, M_{j_k,r}(D))$ is ϵ -differentially private for $j_1, \dots, j_k \in [n]$ such that j_ℓ for $\ell \in [k]$ is drawn from a distribution which is a function of $M_{j_1,r}(D), \dots, M_{j_{\ell-1},r}(D)$.*

We are now ready to prove the main theorem of this section. In what follows, we denote by $A_r(\cdot)$ the algorithm A executed with randomness r . We formalize the multiple-query setting as having one algorithm which takes as part of its input a query. This differs somewhat from the presentation in the introduction which assumed we have a sequence of algorithms, which we chose there as it required less notation. Note also that while we are not assuming that the algorithm does not know which phase it is (the value of i), we may without loss of generality assume this is passed as part of the query. Note that the condition on the queries x_1, \dots, x_k below simply states that the queries can be chosen adaptively based on the released values, and that they do not have to be deterministic. Note also that the randomness r must not be released, as the privacy also relies on that randomness.

► **Theorem 9.** *Let us have an algorithm $A(D, x)$ for a database D and a query $x \in U$, where U is a countable universe. Assume there exists a function $g(D, x)$ with its global sensitivity w.r.t. D being $\leq \Delta_1$ for any x , and such that $\sigma_{se}[A(D, x) - g(D, x)] \leq \Delta_2$ for any D, x .*

Pick at random $Y_i \sim \text{Laplace}(c(\Delta_1 + \Delta_2)k/\epsilon)$ for $c = 3 + 12 \log 2$ and for $\epsilon \leq c/6$ and pick r independently uniformly on $\{0, 1\}^\infty$. Then for queries $x_1, \dots, x_k \in U$ where the query x_i is drawn from a distribution that is a function of $A_r(D, x_1) + Y_1, \dots, A_r(D, x_{i-1}) + Y_{i-1}$, releasing $(A_r(D, x_1) + Y_1, \dots, A_r(D, x_k) + Y_k)$ is ϵ -differentially private, with the privacy also being over the randomness of r .

If $k = 1$, then $c = 1 + 4 \log 2$ and $\epsilon \leq c/2$ is sufficient.

Proof. Let us have two neighboring databases D_1, D_2 . Let $Y = (Y_1, \dots, Y_k)$, and for $x = (x_1, \dots, x_k)$, let $A'(D, x) = (A_r(D, x_1) + Y_1, \dots, A_r(D, x_k) + Y_k)$ for r uniform on $\{0, 1\}^\infty$. Similarly, let $g(D, x) = (g(D, x_1), \dots, g(D, x_k))$. We prove that for any fixed (non-adaptive) queries $x = (x_1, \dots, x_k)$ it holds $f_{A'(D_1, x)}(y)/f_{A'(D_2, x)}(y) \leq e^\epsilon$. If we prove this, the theorem follows: the inequality $f_{A'(D_1, x)}(y)/f_{A'(D_2, x)}(y) \geq e^{-\epsilon}$ holds by symmetry and these bounds together imply ϵ -differential privacy. Lemma 8 then imply that A' is differentially private even for adaptive queries.

Let $R_1 = A'(D_1, x) - g(D_1, x)$ and $R_2 = A'(D_2, x) - g(D_1, x)$ (note the asymmetry in the definitions). We are assuming R_1 has subexponential diameter $\leq \Delta_2$. By Fact 4, it holds that R_2 has subexponential diameter $\leq \Delta_1/\log(2) + \Delta_2$. We now have from Lemma 7 the following bounds

$$\begin{aligned}\frac{f_{g(D_1,x)+R_1+Y}(y)}{f_{g(D_1,x)+Y}(y)} &= \frac{f_{R_1+Y}(y-g(D_1,x))}{f_Y(y-g(D_1,x))} \leq e^{9\log(2)\epsilon/c} \\ \frac{f_{g(D_2,x)+R_2+Y}(y)}{f_{g(D_1,x)+Y}(y)} &= \frac{f_{R_2+Y}(y-g(D_2,x))}{f_Y(y-g(D_1,x))} \geq e^{-3(1+\log 2)\epsilon/c}\end{aligned}$$

which in turn allows us to bound

$$f_{A'(D_1,x)}(y)/f_{A'(D_2,x)}(y) = \frac{f_{g(D_1,x)+R_1+Y}(x)}{f_{g(D_1,x)+Y}(x)} \cdot \frac{f_{g(D_1)+Y}}{f_{g(D_2)+R_2+Y}(x)} \leq e^{(3+12\log 2)\epsilon/c} = e^\epsilon$$

If $k = 1$, the same computation gives the desired bound for $c = 1 + 4\log 2$, since Lemma 7 in that case gives

$$\begin{aligned}\frac{f_{g(D_1,x)+R_1+Y}(y)}{f_{g(D_1,x)+Y}(y)} &\leq e^{3\log(2)\epsilon/c} \\ \frac{f_{g(D_2,x)+R_2+Y}(y)}{f_{g(D_1,x)+Y}(y)} &\geq e^{-(1+\log 2)\epsilon/c}\end{aligned}$$

which again results in the bound $f_{A'(D_1,x)}(y)/f_{A'(D_2,x)}(y) \leq e^\epsilon$. ◀

4 Algorithms with bounded mean error

In this section, we show a weaker version of Theorem 9 that only requires that the error has some number of bounded moments, instead of requiring that it is subexponential. We start by defining the distribution that we will use in our additive noise mechanism.

► **Definition 10.** *Zero-symmetric Pareto distribution with shape parameter $\alpha > 1$ and scale parameter $s > 0$, denoted $ZSPareto_\alpha(s)$, is defined by the PDF*

$$\frac{1}{2s}(\alpha - 1)(|x|/s + 1)^{-\alpha}$$

Before we can prove the main theorem of this section, we need the following technical lemma. The proof is rather technical and it appears in the full version of this paper.

► **Lemma 11.** *Let us have any $0 \leq \epsilon \leq 1$, $\alpha > 1$, and $x \geq 0$. It holds*

$$\int_0^1 \min\left((1+x)^\alpha, \left|1 - \frac{(1-2^{-1/\alpha})\epsilon}{(1+x)(1-u)^{1/\alpha}}\right|^{-\alpha}\right) du \leq 1 + \frac{2\alpha-1}{\alpha-1}\epsilon \quad (3)$$

We are now ready to prove a lemma which bound the privacy loss of our mechanism. In what follows, we use the notation $\|X\|_p$ for a random variable X to denote the L_p norm $\sqrt[p]{\mathbb{E}[|X|^p]}$.

► **Lemma 12.** *Let X be a real random variable such that $\|X\|_\alpha \leq \Delta$ and let $Y \sim ZSPareto_\alpha(s)$ for $s = (1 - 2^{-1/\alpha})^{-1}\Delta/\epsilon$ for $0 \leq \epsilon \leq 1$ and $\alpha > 1$. Consider $y \in \mathbb{R}$. It holds $e^{-(1-2^{-1/\alpha})\alpha\epsilon} \leq f_{X+Y}(y)/f_Y(y) \leq e^{\frac{2\alpha-1}{\alpha-1}\epsilon}$.*

Proof. Since $\|X\|_\alpha \leq \Delta$, it holds by the higher-order Chebyshev inequality⁹ that $\mathbb{P}[|X| \geq z] \leq \Delta^\alpha/z^\alpha$. Therefore, $F_{|X|}^{-1}(u) \leq \Delta/\sqrt[\alpha]{1-u}$. We will use that $|y - X| \leq |y| + |X|$ and later below also that $|y - X| \geq \max(0, |y| - |X|)$. We start with the simpler case of proving a lower bound.

$$\begin{aligned}
\frac{f_{X+Y}(y)}{f_Y(y)} &= \frac{\mathbb{E}[f_Y(y - X)]}{(|y|/s + 1)^{-\alpha}} & (4) \\
&= \frac{\mathbb{E}[(|y - X|/s + 1)^{-\alpha}]}{(|y|/s + 1)^{-\alpha}} \\
&\geq E \left[\left(\frac{|y|/s + |X|/s + 1}{|y|/s + 1} \right)^{-\alpha} \right] \\
&= E \left[\left(1 + \frac{|X|/s}{|y|/s + 1} \right)^{-\alpha} \right] \\
&\geq \mathbb{E}[(1 + |X|/s)^{-\alpha}] \\
&\geq \mathbb{E}[1 - \alpha|X|/s] \\
&= 1 - \alpha\mathbb{E}[|X|]/s \\
&\geq 1 - \frac{\alpha(1 - 2^{-1/\alpha})\Delta}{\Delta} \epsilon \geq e^{-(1-2^{-1/\alpha})\alpha\epsilon} & (5)
\end{aligned}$$

where (4) holds by Fact 2, and (5) uses that $\mathbb{E}[|X|] = \|X\|_1 \leq \|X\|_\alpha \leq \Delta$. We now show the upper bound part. We prove an upper bound in terms of the integral which we have bounded in Lemma 11.

$$\begin{aligned}
\frac{f_{X+Y}(y)}{f_Y(y)} &= \frac{\mathbb{E}[(|y - X|/s + 1)^{-\alpha}]}{(|y|/s + 1)^{-\alpha}} \\
&\leq E \left[\min \left((1 + |y|/s)^\alpha, \left| \frac{|y|/s - |X|/s + 1}{|y|/s + 1} \right|^{-\alpha} \right) \right] \\
&= E \left[\min \left((1 + |y|/s)^\alpha, \left| 1 - \frac{|X|/s}{|y|/s + 1} \right|^{-\alpha} \right) \right] \\
&= E_u \left[\min \left((1 + |y|/s)^\alpha, \left| 1 - \frac{F_{|X|}^{-1}(u)/s}{|y|/s + 1} \right|^{-\alpha} \right) \right] \\
&\leq E_u \left[\min \left((1 + |y|/s)^\alpha, \left| 1 - \frac{(1 - 2^{-1/\alpha})^{-1}\epsilon}{(|y|/s + 1)(1 - u)^{1/\alpha}} \right|^{-\alpha} \right) \right] \\
&= \int_0^1 \min \left((1 + |y|/s)^\alpha, \left| 1 - \frac{(1 - 2^{-1/\alpha})^{-1}\epsilon}{(|y|/s + 1)(1 - u)^{1/\alpha}} \right|^{-\alpha} \right) du \\
&\leq 1 + \frac{2\alpha - 1}{\alpha - 1} \epsilon \leq e^{\frac{2\alpha - 1}{\alpha - 1} \epsilon}
\end{aligned}$$

where we have proven the inequality second to last in Lemma 11. ◀

We are now ready to prove the main theorem of this section.

⁹ The higher-order Chebyshev inequality states that for X being a real random variable, it holds $\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq t^\alpha/\mathbb{E}[|X - \mathbb{E}[X]|^\alpha]$ for any $\alpha \geq 0$.

► **Theorem 13.** *Let us have an algorithm $A(D)$ such that there exists a function $g(D)$ with global sensitivity $\leq \Delta_1$ w.r.t. D for which for any input D , it holds $\mathbb{E}[|A(D) - g(D)|^\alpha] \leq \Delta_2^\alpha$ for some $\alpha > 1$. Let $Y \sim ZSPareto_\alpha(c(\Delta_1 + \Delta_2)/\epsilon)$ for $c = \alpha + 2 + 1/(\alpha - 1)$ and $\epsilon \leq c$, independent of the randomness of A ; then $A(D) + Y$ is ϵ -differentially private with respect to D .*

Proof. This proof follows the strategy of the proof of Theorem 9. Let us have two neighboring databases D_1, D_2 . We again prove that for any y , it holds $f_{A(D_1)}(y)/f_{A(D_2)}(y) \leq e^\epsilon$; this implies the theorem. We also again set $R_1 = A(D_1) - g(D_1)$ and $R_2 = A(D_2) - g(D_1)$. We are assuming $\|R_1\|_\alpha \leq \Delta_2$ and by the triangle inequality, we have that $\|R_2\|_\alpha \leq \Delta_1 + \Delta_2$. Therefore, we have

$$\begin{aligned} \frac{f_{g(D_1)+R_1+Y}(y)}{f_{g(D_1)+Y}(y)} &= \frac{f_{R_1+Y}(y - g(D_1))}{f_Y(y - g(D_1))} \leq \exp\left(\frac{2\alpha - 1}{(\alpha - 1)c}\epsilon\right) \\ \frac{f_{g(D_2)+R_1+Y}(y)}{f_{g(D_1)+Y}(y)} &= \frac{f_{R_1+Y}(y - g(D_2))}{f_Y(y - g(D_1))} \geq \exp\left(-(1 - 2^{-1/\alpha})\alpha\epsilon/c\right) \end{aligned}$$

which in turn allows us to bound

$$\begin{aligned} f_{A(D_1)+Y}(y)/f_{A(D_2)+Y}(y) &= \frac{f_{g(D_1)+R_1+Y}(y)}{f_{g(D_1)+Y}(y)} \cdot \frac{f_{g(D_1)+Y}(y)}{f_{g(D_2)+R_1+Y}(y)} \\ &\leq \exp\left(\left(\frac{2\alpha - 1}{\alpha - 1} + (1 - 2^{-1/\alpha})\alpha\right)\epsilon/c\right) \leq e^\epsilon \end{aligned}$$

where we now argue the last inequality; that will conclude the proof. If we set $c = 2 + 1/(\alpha - 1) + \alpha - 2^{-1/\alpha}\alpha$, the last inequality would be an equality. By monotonicity, it is thus sufficient to prove that $2 + 1/(\alpha - 1) + \alpha - 2^{-1/\alpha}\alpha \leq 2 + \log 2 + 1/(\alpha - 1)$. This is equivalent to $2^{-1/\alpha}\alpha \geq \alpha - \log 2$, which in turn can be re-written as $2^{-1/\alpha} \geq 1 - \log(2)/\alpha$. This follows from the inequality $e^x \geq 1 + x$ for $x = -\log(2)/\alpha$. ◀

5 Implications of our results

In this section, we give several implications of Theorem 9 and Theorem 13. This list is by no means meant to be exhaustive. We start with the more straightforward applications and focus on the more involved ones later, with one part being deferred to the full version of this paper.

Recall that, as we discussed, the goal in the sublinear setting is not to simply add small amount of noise, but rather to achieve a given level of error as efficiently as possible while guaranteeing differential privacy. This is so because in this setting, the amount of error coming from the algorithm not being exact tends to be much greater than the amount of noise needed to achieve privacy when not subject to having limited resources.

5.1 The general approach

In all applications, we take a known algorithm for a given problem, and use either Theorem 9 or Theorem 13 to argue that adding noise to the algorithm's answer ensures privacy.

Assume the original algorithm had complexity $T(n, \rho)$ and assume for example that the error is of magnitude ρn , namely that for error R , it holds $E[R^2]^{1/2} \leq O(\rho n)$ (this can be generalized to $\leq O(\rho f(x))$ for x being the input and f being any function). We run the algorithm with parameter $\rho' = \epsilon\rho$ and add noise of magnitude $O(\rho n)$ (more generally

$O(\rho f(x))$). By Theorem 13 with $\alpha = 2$, as long as the approximated function's sensitivity is $\Delta \leq \epsilon \rho n$, this is ϵ -differentially private¹⁰. At the same time, the error is $\leq O(\rho n)$ with arbitrarily high constant probability. The time complexity is $T(n, \epsilon \rho)$.

If we want to achieve a failure probability of β , we run this algorithm $\Theta(\log \beta^{-1})$ times and take the median. By a standard probability amplification argument, the success probability will be as desired. To achieve ϵ -differential privacy by composition, we have to divide the privacy budget between the runs, resulting in complexity $O(T(n, \epsilon \rho / \log \beta^{-1}) \log \beta^{-1})$. This can be summarized (and generalized with the general function $f(x)$) as follows:

► **Lemma 14.** *Suppose there is an algorithm approximating a function g with global sensitivity Δ such that $E[(A(x) - g(x))^2]^{1/2} \leq \rho f(x)$ for some function f with time/space/query complexity $T(n, \rho)$. Then for $\epsilon \leq O(1)$ there exists an ϵ -differentially private algorithm A' such that when $\epsilon \rho \geq \Omega(\Delta/f(x))$, it holds $P[|A'(x) - g(x)| > \rho f(x)] \leq \beta$ and that has complexity $O(T(n, \epsilon \rho / \log \beta^{-1}) \log \beta^{-1})$.*

A more efficient approach for decreasing failure probability exists in the case of subexponential error. Assume the same setting as above, except that the error's subexponential diameter is ρn instead having only moment bounds (like above, we can generalize to $\rho f(x)$ instead of ρn). We run the algorithm with parameter $\rho' = \epsilon \rho / \log \beta^{-1}$ and add noise of magnitude $\Theta(\rho n / \log \beta^{-1})$. This algorithm is ϵ -differentially private by Theorem 9, as long as $\Delta \leq \epsilon \rho n$. The noise has subexponential diameter $O(\rho' n)$ and by Lemma 5, the total error will up to a constant have the same subexponential diameter. By the definition of subexponential diameter, the probability that the error is $\geq \Theta(\rho n)$ is $\leq \beta$ as desired. This results in complexity $O(T(n, \epsilon \rho / \log \beta^{-1}))$ saving us one $\log \beta^{-1}$ factor. We can summarize this as

► **Lemma 15.** *Suppose there is an algorithm approximating a function g with global sensitivity Δ such that $\sigma_{se}[A(x) - g(x)] \leq \rho f(x)$ for some function f with time/space/query complexity $T(n, \rho)$. Then for $\epsilon \leq O(1)$, there exists an ϵ -differentially private algorithm A' such that when $\epsilon \rho \geq \Omega(\Delta/f(x))$, it holds $P[|A'(x) - g(x)| > \rho f(x)] \leq \beta$ and that has complexity $O(T(n, \epsilon \rho / \log \beta^{-1}))$.*

We are now ready to give private algorithms for specific problems.

5.2 Frequency moment F_2

In their seminal paper, Alon, Matias, and Szegedy [2] show a sketch that allows one to estimate the F_2 frequency moment, defined as $F_2(x_1, \dots, x_n) = \sum_{i=1}^n x_i^2$. In the streaming setting, the vector x_1, \dots, x_n is given through a stream of updates y_1, \dots, y_k of the form $y_j = (\ell_j, D_j)$ where D can be negative and we define $x_i = \sum_{j=1}^k I[\ell_j = i] D_j$. Two inputs are then adjacent if they differ in one value y_j for some j . The algorithm from [2] uses space $O(\frac{1}{\rho^2})$ and has mean squared error of $\leq \rho^2 F_2^2 \leq \rho^2 n^4$. The sensitivity of the F_2 moment is n . This implies the following

► **Corollary 16.** *For $\epsilon \leq O(1)$ and $\rho \leq 1/(\epsilon n)$, there is an ϵ -differentially private algorithm that returns an additive $\pm \rho n^2$ approximation of the frequency moment F_2 with probability $1 - \beta$, and has space complexity $O(\frac{\log^3 \beta^{-1}}{\rho^2 \epsilon^2})$.*

¹⁰This upper bound on the sensitivity ensures that Δ_2 in Theorem 13 dominates.

This improves upon the concurrent work [7] which gives an algorithm with complexity $O(\frac{\log^4 n}{\rho^2 \epsilon^2})$. Shortly after releasing this paper, an approach also appeared with the incomparable complexity of $O(\log(n) \log^3 \beta^{-1} / \rho^2)$ was shown (for ϵ being not too small) that also has multiplicative approximation guarantees and works (with some loss in the complexity) in the continual release setting [15]. The setting of F_p for $p \in [0, 1]$ has been considered in [33].

5.3 Connected components

An algorithm is known [4] that returns an estimate \hat{c} of the number of connected components c of a simple graph in time $O(\frac{1}{\rho^2} \log \frac{1}{\rho})$ and has mean squared error $\mathbb{E}[(\hat{c} - c)^2] \leq \rho^2 n^2$. At the same time, the number of connected components has global sensitivity 1 with respect to edge additions/deletions. This gives us the following

► **Corollary 17.** *For $\epsilon \leq O(1)$ and $\rho \leq 1/(en)$, there is an ϵ -edge-differentially private algorithm that returns an additive $\pm \rho n$ approximation of the number connected components with probability $1 - \beta$, and has complexity $O(\frac{\log^3 \beta^{-1}}{\rho^2 \epsilon^2} \log \frac{\log \beta^{-1}}{\rho \epsilon})$.*

No private sublinear-complexity algorithm for estimating the number of connected components was previously known.

5.4 Maximum matching

Yoshida, Yamamoto, and Ito [34] show an algorithm that can approximate the size of the maximum matching to within multiplicative $1 + \rho$ in time $d^{O(1/\rho^2)} (1/\rho)^{O(1/\rho)}$ for d being the maximum degree of the input graph. It works by implementing an oracle for a matching of size within factor $1 + \rho/2$ of the maximum matching; for a specified vertex, this oracle answers whether the vertex is matched in the oracle's matching. The algorithm then samples $\Theta(1/\rho^2)$ vertices and checks the fraction that is matched in the oracle's matching. The error coming from the oracle is $\leq \rho n/2$ in the worst case and thus has subexponential diameter $O(\rho n)$. The error coming from the sampling has subexponential diameter $O(\rho n)$ by the Hoeffding inequality. By Lemma 5, the subexponential diameter of the error is thus $O(\rho n)$. At the same time, the global sensitivity of the maximum matching size is ≤ 1 with respect to the removal of one vertex. This gives us the following

► **Corollary 18.** *For $\epsilon \leq O(1)$ and $\rho \leq 1/(en)$, there is an ϵ -node-differentially-private algorithm that returns an additive $\pm \rho n$ approximation of the maximum matching size with probability $1 - \beta$ in time $d^{O(\log^2(\beta^{-1})/(\rho^2 \epsilon^2))} / (\rho \epsilon)^{O(\log(\beta^{-1})/(\rho \epsilon))}$.*

Together with the concurrent [7], this solves the open problem posed in [5] where the authors show a hybrid $(2, \rho n)$ approximation, while we give a purely additive $\pm \rho n$ approximation.

5.5 Rank queries

Karnin, Lang, and Liberty [20] develop a sketch of size $O(\frac{1}{\rho})$ that allows one to answer rank queries with error with subexponential diameter ρn . We show how to use their sketch to answer range queries over an ordered universe. For small number of queries, this improves upon the work of [19] which has a logarithmic dependency on the universe size. This gives us the following corollary.

► **Corollary 19.** *There is a sketch that allows ϵ -differentially private algorithm that returns k rank queries (potentially adaptive) for $\epsilon \leq O(1)$ with an additive $\pm \rho n$ error with probability $1 - \beta$, and has complexity $O(\frac{k \log^2(k/\beta)}{\rho \epsilon})$.*

Proof. We use the KLL sketch with error parameter $\rho' = \rho\epsilon/(k \log(k/\beta))$. This means that the error has a subexponential diameter of $\leq \rho\epsilon n/(k \log(k/\beta))$. Therefore, by Theorem 9, it holds that using for each query a Laplace mechanism with error magnitude $\Theta(\rho n/\log(k/\beta))$ will result in ϵ -differential privacy.

By Lemma 5, the overall subexponential diameter of the error of each answer is $O(\rho n/\log(k/\beta))$ and therefore the probability of error $O(\rho n)$ is $1 - \beta/k$. By the union bound, the overall success probability is $\geq 1 - \beta$. \blacktriangleleft

This is in comparison to the approach of Kaplan and Stemmer [19] which results in space complexity $O(\frac{\log |U| \log(k/\beta)}{\rho\epsilon})$, improving by a factor of $\log |U|$ for constant k, β , where U is the universe.

5.6 Relative approximation sublinear-time algorithms

In the full version of this paper, we show a general theorem that implies that many sublinear-time algorithms that have relative error guarantees can be made differentially private. Specifically, it applies to many algorithms that rely on a commonly used “advice removal” technique [14]. Among other things, this theorem implies the following algorithm. This improves upon the work of Blocki, Grigorescu, and Mukherjee [5] who give an algorithm with complexity $\tilde{O}_{\epsilon,\rho}(\sqrt{n})$ under the assumption $m \geq \Omega(n)$.

► **Corollary 20.** *For $\epsilon \leq O(1)$ and $\rho \leq 1/(\epsilon n)$, there exists an ϵ -edge-differentially private algorithm that returns a $1 + \rho$ -approximation of the average degree of a graph with probability $1 - \beta$ and has complexity $O(\frac{n \log^3 \beta^{-1}}{\epsilon^2 \rho^2 \sqrt{m}})$.*

6 Open problems and conjectures

We are convinced that our results are only the beginning of the story and that there are many interesting related open problems. These include using normal noise with approximate/zero-concentrated differential privacy, generalizing our method to input-dependent error magnitudes, improving the constants, lower bounds for answering multiple queries when we have polynomial tails. For a more detailed discussion, see the full version of this paper.

References

- 1 Daniel Alabi, Omri Ben-Eliezer, and Anamay Chaturvedi. Bounded space differentially private quantiles. *arXiv preprint*, 2022. [arXiv:2201.03380](#).
- 2 Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29, 1996.
- 3 Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Guha Thakurta. Practical locally private heavy hitters. *Advances in Neural Information Processing Systems*, 30, 2017.
- 4 Petra Berenbrink, Bruce Krayenhoff, and Frederik Mallmann-Trenn. Estimating the number of connected components in sublinear time. *Information Processing Letters*, 114(11):639–642, 2014.
- 5 J Blocki, E Grigorescu, and T Mukherjee. Privately estimating graph parameters in sublinear time. In *49th International Colloquium on Automata, Languages, and Programming (ICALP 2022)*, 2022.
- 6 Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The johnson-lindenstrauss transform itself preserves differential privacy. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 410–419. IEEE, 2012.

- 7 Jeremiah Blocki, Elena Grigorescu, Tamalika Mukherjee, and Samson Zhou. How to Make Your Approximation Algorithm Private: A Black-Box Differentially-Private Transformation for Tunable Approximation Algorithms of Functions with Low Sensitivity. In Nicole Megow and Adam Smith, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2023)*, volume 275 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 59:1–59:24, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.APPROX/RANDOM.2023.59.
- 8 Jonas Boehler and Florian Kerschbaum. Secure sublinear time differentially private median computation, February 1 2022. US Patent 11,238,167.
- 9 Seung Geol Choi, Dana Dachman-Soled, Mukul Kulkarni, and Arkady Yerukhimovich. Differentially-private multi-party sketching for large-scale statistics. *Cryptology ePrint Archive*, 2020.
- 10 Charlie Dickens, Justin Thaler, and Daniel Ting. (nearly) all cardinality estimators are differentially private. *arXiv preprint*, 2022. arXiv:2203.15400.
- 11 Marianne Durand and Philippe Flajolet. Loglog counting of large cardinalities. In *Algorithms-ESA 2003: 11th Annual European Symposium, Budapest, Hungary, September 16-19, 2003. Proceedings 11*, pages 605–617. Springer, 2003.
- 12 Cynthia Dwork. Differential Privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
- 13 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.
- 14 Talya Eden, Dana Ron, and C Seshadhri. On approximating the number of k-cliques in sublinear time. In *Proceedings of the 50th annual ACM SIGACT symposium on theory of computing*, pages 722–734, 2018.
- 15 Alessandro Epasto, Jieming Mao, Andres Munoz Medina, Vahab Mirrokni, Sergei Vassilvitskii, and Peilin Zhong. Differentially private continual releases of streaming frequency moment estimations. *arXiv preprint*, 2023. arXiv:2301.05605.
- 16 geetha290krm (<https://math.stackexchange.com/users/1064504/geetha290krm>). Does $f_{X+Y}(z) = e[f_Y(z-x)]$ hold? Mathematics Stack Exchange, 2022. URL:<https://math.stackexchange.com/q/4544852> (version: 2022-10-04). arXiv:<https://math.stackexchange.com/q/4544852>.
- 17 Badih Ghazi, Noah Golowich, Ravi Kumar, Rasmus Pagh, and Ameya Velingker. On the power of multiple anonymous messages: Frequency estimation and selection in the shuffle model of differential privacy. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 463–488. Springer, 2021.
- 18 Ziyue Huang, Yuan Qiu, Ke Yi, and Graham Cormode. Frequency estimation under multiparty differential privacy: One-shot and streaming. *arXiv preprint*, 2021. arXiv:2104.01808.
- 19 Haim Kaplan and Uri Stemmer. A note on sanitizing streams with differential privacy. *arXiv preprint*, 2021. arXiv:2111.13762.
- 20 Zohar Karnin, Kevin Lang, and Edo Liberty. Optimal quantile approximation in streams. In *2016 IEEE 57th annual symposium on foundations of computer science (focs)*, pages 71–78. IEEE, 2016.
- 21 Kasper Green Larsen, Rasmus Pagh, and Jakub Tětek. Countsketches, feature hashing and the median of three. In *International Conference on Machine Learning*, pages 6011–6020. PMLR, 2021.
- 22 Christian Janos Lebeda and Jakub Tětek. Better differentially private approximate histograms and heavy hitters using the misra-gries sketch. *arXiv preprint*, 2023. arXiv:2301.02457.
- 23 Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press, 2015.

- 24 Luca Melis, George Danezis, and Emiliano De Cristofaro. Efficient private statistics with succinct sketches. *arXiv preprint*, 2015. [arXiv:1508.06110](#).
- 25 Darakhshan Mir, Shan Muthukrishnan, Aleksandar Nikolov, and Rebecca N Wright. Pan-private algorithms via statistics on sketches. In *Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 37–48, 2011.
- 26 Jayadev Misra and David Gries. Finding repeated elements. *Science of computer programming*, 2(2):143–152, 1982.
- 27 Rasmus Pagh and Mikkel Thorup. Improved utility analysis of private counts sketch. *arXiv preprint*, 2022. [arXiv:2205.08397](#).
- 28 Ryan M Rogers, Aaron Roth, Jonathan Ullman, and Salil Vadhan. Privacy odometers and filters: Pay-as-you-go composition. *Advances in Neural Information Processing Systems*, 29, 2016.
- 29 Harry Sivasubramaniam, Haonan Li, and Xi He. Differentially private sublinear average degree approximation, 2020.
- 30 Adam Smith, Shuang Song, and Abhradeep Guha Thakurta. The flajolet-martin sketch itself preserves differential privacy: Private counting with minimal space. *Advances in Neural Information Processing Systems*, 33:19561–19572, 2020.
- 31 Nina Mesing Stausholm. Improved differentially private euclidean distance approximation. In *Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 42–56, 2021.
- 32 Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- 33 Lun Wang, Iosif Pinelis, and Dawn Song. Differentially private fractional frequency moments estimation with polylogarithmic space. *arXiv preprint*, 2021. [arXiv:2105.12363](#).
- 34 Yuichi Yoshida, Masaki Yamamoto, and Hiro Ito. An improved constant-time approximation algorithm for maximum $\tilde{}$ matchings. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 225–234, 2009.
- 35 Fuheng Zhao, Dan Qiao, Rachel Redberg, Divyakant Agrawal, Amr El Abbadi, and Yu-Xiang Wang. Differentially private linear sketches: Efficient implementations and applications. *arXiv preprint*, 2022. [arXiv:2205.09873](#).