Differential Privacy in Distributed Settings

Nina Mesing Stausholm Nielsen

IT University of Copenhagen, Computer Science Dept.
Principal supervisor: Rasmus Pagh

Submitted on September 30th, 2021
Defended on November 22nd, 2021

Resumé

Som et resultat af de mange forskellige digitale enheder der er til radighed, samt hurtige og billige datalagringsmuligheder, indsamles og deles enorme m ngder data konstant med forskellige formal. Alene omfanget af denne dataindsamling giver udfordringer indenfor diverse forskningsomrader, sasom adgangskontrol, kryptogra og databeskyttelse. Mens sikkerhed er et komplementerende forskningsfelt, fokuserer vi i denne afhandling pa databeskyttelse (*data privacy*).

Vi betragter *di erential privacy*[1], der med sin stringente de nition samt det, at vi ikke beh ver antagelser om analytikerens baggrundsviden eller beregningskraft, er blevet state-of-the-art indenfor databeskyttelse. Di erential privacy lader os eksplicit kvanti cere tabet af privacy og giver derved transparens omkring de givne privacy garantier i en applikation. Speci kt vil vi prim rt fokusere pa en distribueret model, hvor data er fordelt mellem mange kuratorer, hvilket ofte er tilf ldet i praksis, som for eksempel data indsamlet fra mobile enheder eller landsspeci kke data angaende, for eksempel, en pandemi.

Hovedbidragene i denne afhandling er som f lger:

- Vi introducerer en kompakt opsummering af et datas t, en sakaldt *sketch*, til e ektivt og pr cist at estimere kardinaliteten af et datas t, mens di erential privacy opretholdes. En vigtig anvendelse er at to sadanne sketches over m ngder A og B kan kombineres til en sketch for den symmetriske di erens, $A \vartriangle B$, og dermed giver mulighed for at estimere kardinaliteten af den symmetriske di erens mellem inputm ngder der holdes af forskellige kuratorer, og derfor ikke kan udveksles.
- Vi introducerer en di erentially private sketch til at estimere Euklidisk afstand mellem reelle inputvektorer holdt af forskellige kuratorer og beviser at vores sketch opnar bedre privacy, e ektivitets og pr cisionsgarantier end tidligere arbejde.
- Vi introducerer en ny st jfordeling, Aretefordelingen, der giver en di erentially private mekanisme, Arete mekanismen. Denne mekanisme giver mulighed for at udf re statistisk analyse over et datas t med en fejl, der er eksponentielt aftagende i privacy parameteren ε, og derved forbedrer Laplace mekanismen nar kravene om privacy er lave (for store v rdier af ε). Derudover har Aretest jfordelingen en kontinuert t thedsfunktion samt er uendeligt delelig[2] hvilket betyder, at vi kan fordele den n dvendige st j for at opna di erential privacy mellem mange kuratorer og derved tillade privat, distribueret analyse med h j pr cision.

Vi de nerer formelt di erential privacy samt de to centrale problemer *v gtet kardinalitetsestimering* og *approksimation af Euklidisk afstand*, og giver stringente beviser for hvert af de n vnte resultater. Derudover n vner og diskuterer vi et antal abne problemer i forl ngelse af bidragene fra denne afhandling.

[1] Der findes endnu ikke et veletableret dansk udtryk for differential privacy og vi benytter derfor det engelske begreb.

[2] *In nitely divisible*

Abstract

As a result of the variety of digital devices available and fast and cheap storage, huge amounts of data are constantly collected and shared for various purposes. The mere extent of the data creation and collection introduces challenges within various research fields such as access control, cryptography and privacy protection. While security is a complementary field of research, we will in this thesis focus on privacy protection.

We study *differential privacy*, the state-of-the-art privacy technique, due to the stringent definition and the fact that we make no assumptions about the background knowledge of the analyst or their computational power. Differential privacy lets us explicitly quantify the privacy loss and establishes transparency in an application's privacy guarantees. Specifically, we will primarily focus on a distributed setting, where data is split among many curators as is often the case in practice, such as, for example, data collected from mobile devices or country-specific data about, say, a pandemic.

The main contributions of this dissertation are as follows:

- We introduce a compact summary of a dataset, a *sketch*, for efficiently and accurately estimating the cardinality of the dataset while preserving differential privacy. An important application is that two such sketches over sets A and B can be combined into a sketch for the symmetric difference, $A \triangle B$, allowing for privately estimating the cardinality of the symmetric difference between input sets held by different curators, and so cannot be exchanged.
- We introduce a differentially private sketch for estimating the Euclidean distance between real input vectors held by different curators and prove that our sketch achieves better privacy, efficiency and accuracy guarantees than previous work.
- We introduce a new noise distribution, the Arete distribution, which permits a differentially private mechanism, the Arete mechanism. This mechanism incurs error exponentially decreasing in the privacy parameter ε, improving over the Laplace mechanism in the low privacy regime (for large ε). Furthermore, the noise distribution has a continuous density function and is infinitely divisible, meaning that we can distribute the noise necessary to ensure differential privacy among several data curators to allow for private, distributed analysis with high accuracy.

We formally define differential privacy along with the two central problems of *weighted cardinality estimation* and *Euclidean distance approximation* and give stringent proofs for each of the results mentioned. Furthermore, we present and discuss several open problems extending the contributions of this thesis.

Dedication

I wish to dedicate this work to my dear brother, who is just now embarking on the adventures of the PhD student. Dan, I wish you the best of luck. Enjoy this opportunity and remember that:

> *Experience [...] was merely the name men gave to their mistakes.*[3]
> *{ Oscar Wilde, The Picture of Dorian Grey.*

Acknowledgements

First and foremost, I wish to thank my principal supervisor Rasmus Pagh for his excellent support and guidance. Rasmus has, throughout my time as a PhD student, encouraged and helped me to develop myself as an independent researcher by continuously challenging me while always having a helping hand, a word of advice or a comforting remark ready at hand in trying times { not to forget a generous supply of praise, motivating me to keep going. I could not have wished for a better supervisor.

I also wish to thank my co-supervisors, Mikkel Thorup and Troels Lund, and all of my colleagues at ITU and BARC for three wonderful (and occasionally less wonderful) years, great discussions, your patience when I needed to blow o some steam (Christian Lebeda, Mille Nielsen and Mads Lassen, that one was for you) and a steep, but incredibly rewarding, learning slope. A huge thanks to Graham Cormode and Edith Cohen for collaborating with me on exciting research questions and a particular thanks to Graham Cormode for hosting my stay abroad at Warwick University and including me in his research group from the very start. To the anonymous reviewers having peer-reviewed the works presented in this thesis (and helped to improve said work through helpful comments) and to anyone who has given me valuable feedback and advice on everything ranging from talks and written work to where the best ice cream place is: Thank you!

I want to thank my family and friends for their in nite, loving support and always being there for me in good times and bad. I can always count on you for a comforting hug or to celebrate with me over a glass of bubbly. Thanks for being patient with me and always, always taking the time to listen when I needed to get something o my chest { regardless of whether it was real, valid concerns or just the occasional rant because life isn't fair.

One nal (major) thanks to the VILLUM Foundation, whose generous grant has supported me throughout my PhD studies[4].

[3] Experience was of no ethical value. It was merely the name men gave to their mistakes.

[4] Investigator Grant 16582, Basic Algorithms Research Copenhagen (BARC)

Contents

Chapter 1

Introduction

> All the world is made of faith, and trust, and pixie dust.
>
> J. M. Barrie
> Peter Pan

Recent developments in technology enable powerful and continuous data collection and curation. The statistical properties of this data are valuable for developments in other elds { a simple example is to understand serious diseases better, allowing us to diagnose patients sooner and treat them better. Data analysis o ers the opportunity to learn about the world and ourselves.

However, data analysis also raises several concerns, such as storage, speed { and privacy. While privacy is commonly considered a human right, it is not so easy to argue *why* we do not want others to go through our trash or read our mail. And similarly, why we want to keep our Google searches, location or health diaries to ourselves. Nevertheless, despite being well aware of the data collection, we keep using the ashlight app, which requires access to the microphone for no apparent reason because it had a nicer interface. Or click the "accept all cookies" button because we are in a hurry and somehow expect data collectors to "play nice". And who really cares if I watched that cat video... again? So what is privacy, and why is it so important?

> *Arguing that you don't care about the right to privacy because you have nothing to hide is no di erent than saying you don't care about free speech because you have nothing to say.*
> *{ Edward Snowden, [126]*

> *Privacy matters; privacy is what allows us to determine who we are and who we want to be.*
> *- Edward Snowden, [131]*

Over the recent years, data protection laws and regulations (such as the General Data Protection Regulation (GDPR) in the EU [66], the California Consumer Privacy Act [123] and India's Personal Data Protection Bill [85]) have been introduced in an attempt to hinder unauthorized data collection and to ensure transparency in the handling of user data.

A widely used method for preserving data privacy is via anonymization, where we remove obvious identi ers such as name, IP address or social security number from the data. Many examples have proven that anonymization does not provide su ciently strong privacy guarantees [15, 26, 44, 83, 115, 133], since anonymized data is vulnerable to *linkage attacks* (also known as re-identi cation attacks). In such attacks, anonymized data records are re-identi ed by linking the anonymized dataset with non-anonymized auxiliary information. In a prominent example, Sweeney showed that upwards 87% of the US population are uniquely identi able given only their 5-digit zip code, gender and date of birth [132]. Sweeney later re-identi ed anonymized hospital records of (at the time) Massachusetts Governor William Weld by comparing hospital records with public voter registration records using these three simple demographics [133]. Other examples

include when Calandrino et al. uncovered Amazon customers' purchase history via collaborative filtering (the "Customers who bought this item also bought..." feature) [26] or when Narayanan and Shmatikov re-identified Netflix users by comparing the anonymized user ratings dataset with the publicly available user ratings dataset from IMDB [115], assuming that users are likely to give the same movie similar ratings at approximately the same time on the two platforms. In conclusion:

Anonymized data isn't { *Cynthia Dwork.*

Even aggregate data and statistics about data may leak sensitive information: The Fundamental Law of Information Recovery says that *overly accurate answers to too many questions will destroy privacy in a spectacular way* [49]. Intuitively, every time a statistical result about a dataset is released, a bit of information about each data record is leaked { death by a thousand cuts. Dinur and Nissim [49] exploited this principle with a reconstruction attack, proving that one can reconstruct large parts of a dataset via several aggregate query results if we answer each query too accurately. This work was the first in a line of research about reconstruction and re-identification of data from auxiliary information, eventually leading to the definition of *differential privacy*[1], which by design protects against such attacks. This definition enables us to make privacy guarantees that hold regardless of the computational power and auxiliary information available to the adversary. Another equally important property of differential privacy is that we can quantify the privacy loss even when the same data is subjected to multiple queries, allowing us to fix a privacy budget for an application. Intuitively, differential privacy permits statistical analysis of a dataset while protecting the privacy of each individual record in the dataset by adding noise to query results. We define differential privacy in the next chapter.

Several examples have shown that a single unit curating large amounts of data is vulnerable to large-scale data leaks: in the Facebook-Cambridge Analytica scandal, data about millions of Facebook users, collected by the data analytics firm without the users' consent, was leaked and used to build voter profiles for political campaigns [28]. Other examples include when hackers gained unauthorized access to the guest reservation database of the Marriott International hotel chain [101], or when the social media Whisper (branded as the "safest place on the Internet") left age and location data tied to anonymous posts exposed in a database openly accessible { even without password-protection [51], thus being vulnerable to linkage attacks. We consider a more general setting, where data may be distributed among many data curators. In particular, we are interested in the case where data is too extensive or sensitive to share, and each curator releases only a summary of their data. Going even further, we may create summaries that can be combined (offline) to enable analysis over the entire dataset without each curator sharing more than the summary.

In this thesis, we will discuss private data analysis over distributed data. In Chapters 3 and 4, we will discuss how to create differentially private summaries supporting efficient and accurate statistical analysis for two fundamental problems in data management and data analysis, F_0 estimation and Euclidean distance approximation. In Chapter 5 we return to the basics of differential privacy and introduce a new noise distribution, the Arete distribution, which can be applied to ensure differential privacy of real-valued queries. The Arete distribution combines the best properties of previous noise distributions to give a differentially private mechanism that can be distributed among several parties while ensuring low error. Before formally presenting our results, we give the background and introduce and motivate these results in Chapter 2.

The papers making up this thesis are listed below with references to the chapters treating the results in-depth:

1. *Efficient Differentially Private F_0 Linear Sketching* (ICDT 2021) by Rasmus Pagh and Nina Mesing Stausholm[2] [119]. The results are presented in Chapter 3.

2. *Improved Differentially Private Euclidean Distance Approximation* (PODS 2021) by Nina Mesing Stausholm [130]. The results are presented in Chapter 4.

[1] The name Differential Privacy was suggested by Michael Schroeder [63]

[2] Full name: Nina Mesing Stausholm Nielsen

3. *The Arete Distribution for Di erentially Private Noise Addition* (in submission) by Rasmus Pagh and Nina Mesing Stausholm. The results are presented in Chapter 5.

During my time as a PhD student, I co-authored an additional publication, *Hardness of Bichromatic Closest Pair with Jaccard Similarity* (ESA 2019) alongside Rasmus Pagh and Mikkel Thorup [120]. This work extends a result from my Master's thesis and is not included in this dissertation, as it deals with the problem of similarity search rather than data privacy.

Chapter 2

Background

In this chapter, we give the necessary background and the intuitive introductions to the results presented in Chapters 3, 4 and 5.

2.1 Differential Privacy

2.1.1 The Intuitive Explanation

Consider a dataset containing sensitive information about a group of people. A statistical query $q : X \to \mathrm{R}^k$ for $k \geq 1$ is a function that can be applied to the dataset to learn statistical properties of the data, such as the mean of their ages or the number of individuals satisfying a certain predicate. Such analysis over sensitive data may lead to privacy issues, as exhibited in the following example:

Example: *Suppose that a hospital generates a dataset over the daily positive (anonymous) test results for diabetes for statistical purposes. A nosy nurse knows that patient X is to be tested for diabetes and keeps an eye on the number of positive tests reported in the dataset to determine whether or not patient X has diabetes.*

Releasing query results in a differentially private manner intuitively means that for similar datasets, any query output is essentially equally likely. Informally, differential privacy guarantees that an analyst cannot determine whether any specific data record was present in or absent from the dataset from an observed query result. The goal is to protect the data of every individual while permitting statistical analysis on the dataset as a whole.

An intuitive requirement for a privacy definition for data analysis is that analysts know no more about any individual in the dataset after the analysis than before. However, consider a study showing that smoking leads to lung cancer. The analyst learns that the smoker next door is more likely to get lung cancer, thereby violating the privacy guarantee. This requirement is too strict, as it allows us to learn nothing from the analysis, and the example is not considered a privacy breach according to differential privacy since it is the *conclusions* of the study rather than the presence in or absence from the dataset that affect the smoker. Differential privacy essentially says that nothing can be learned from a dataset that could not be learned from the same dataset with any individual's data removed.

2.1.2 The Formal Definition

As we are interested only in hiding whether a single individual's data is in the dataset or not, we define *neighboring datasets* and *sensitivity* of a query. Let $D \subseteq X$ be a dataset. Datasets are neighbors if they are identical except for a single record, and the sensitivity of a query intuitively measures how much the result of the query can differ when applied to neighboring datasets. That is, the sensitivity of a query helps to

understand how much noise is necessary { or rather, how big a di erence the noise must hide. The following de nition is useful for de ning neighboring inputs when X is numeric:

Definition 2.1 (ℓ_p-norm)**.** *For real $p \geq 1$, the ℓ_p-norm of $x \in \mathbb{R}^u$ is*

$$\|x\|_p = \left(\sum_{j=1}^{u} |x_j|^p \right)^{1/p}.$$

We de ne neighboring inputs and sensitivity of a query formally:

Definition 2.2 (Neighboring inputs)**.** *Inputs $x, y \in X$ are* neighbors*, sometimes also called* adjacent*, if they di er in at most one data record. We denote neighboring inputs by $x \doteq y$. If $X \subseteq \mathbb{R}^u$, $u \geq 1$, x and y are neighbors if $\|x - y\|_1 \leq 1$.*

Definition 2.3 (ℓ_p-sensitivity)**.** *Let $q : X \to \mathbb{R}^k$ be a query. The ℓ_p-sensitivity of q is*

$$\Delta := \max_{x,y \in X:\ x \doteq y} \|q(x) - q(y)\|_p.$$

Unless otherwise speci ed, whenever we simply write the *sensitivity*, we refer to the ℓ_1-sensitivity. Occasionally (in Chapter 4), we also consider the ℓ_2-sensitivity but will specify the distinction.

We now give the formal de nition of di erential privacy:

Definition 2.4 (Di erential Privacy [55, 60])**.** *A randomized mechanism $M : X \to \mathrm{Range}(M)$ is (ε, δ)-di erentially private for $\varepsilon, \delta \geq 0$ if for any neighboring inputs $x, y \in X$ and for all $S \subseteq \mathrm{Range}(M)$*

$$\Pr[M(x) \in S] \leq e^{\varepsilon} \Pr[M(y) \in S] + \delta$$

where the probability is over the random coin ips performed by M. If $\delta = 0$, we say that M is ε-di erentially private.

Note that one can think of $M(x)$ as a probability distribution over the possible values of M applied to x. We often simply write $M(x)$ to denote the output of M applied to x instead of the more accurate $M(x)$.

We will mainly concern ourselves with ε-di erential privacy, sometimes referred to as *pure* di erential privacy, but will touch upon (ε, δ)-di erential privacy, also called *approximate* di erential privacy, in Chapter 4. While not quite accurate [107], a common, intuitive interpretation of approximate di erential privacy is that we get pure di erential privacy except with probability δ [112]. While approximate di erential privacy is *theoretically* weaker than pure di erential privacy, in practice, the guarantees are essentially the same for su ciently small δ. A useful de nition is that of *privacy loss*:

Definition 2.5 (Privacy loss incurred by observing ξ)**.** *Let $x, y \in X$ be neighbors. For observed output $M(x)$, the privacy loss incurred by observing ξ is de ned as*

$$\ln \left(\frac{\Pr[M(x) = \xi]}{\Pr[M(y) = \xi]} \right), \qquad \Pr[M(y) = \xi] \neq 0.$$

For an ε-di erentially private mechanism, the privacy loss is *always* bounded by ε, while for an (ε, δ)-di erentially private mechanism, the privacy loss is bounded by ε with probability at least $1 - \delta$. Di erential privacy allows us to explicitly quantify the greatest possible privacy loss in terms of the privacy parameter ε and so allows us to compare algorithms: for a certain accuracy guarantee, we may ask if a technique provides better privacy, or for a certain privacy guarantee, which technique has better accuracy. For more details about di erential privacy, we refer the reader to [56, 63, 137].

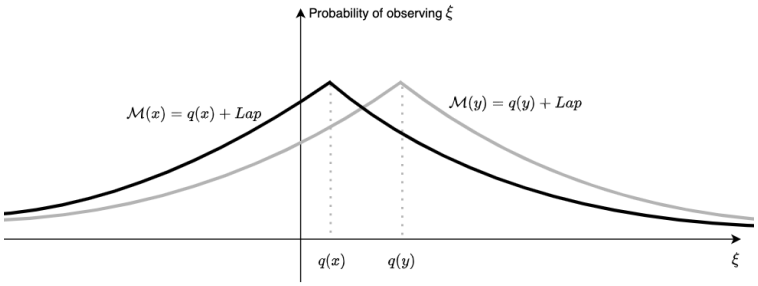


Figure 2.1: Illustration of densities for the output of mechanism M on inputs x (black) and y (gray), where M adds Laplace noise to the query output $q(x)$ and $q(y)$. If M is ε-differentially private and x and y are neighbors, then the two distributions will differ by at most a factor e^{ε} in any point.

2.1.3 The Power of Differential Privacy

Di erential privacy is a mathematical worst-case guarantee about a randomized mechanism. It is important to note that it is a property of the *mechanism* rather than the perturbed data: Given a query output, there is no way to determine whether it satis es di erential privacy, only whether it was released via a provably di erentially private mechanism. One of the primary strengths of di erential privacy is this mathematical guarantee, as we can formally prove the privacy level ensured by a mechanism in terms of ε and .

Moreover, there are strong results about the privacy guarantees under post-processing and composition: The post-processing property states that one cannot decrease the privacy level by performing additional computations to a query result released by a di erentially private mechanism:

Lemma 2.1 (Post-processing [60]). *For (ε;)-di erentially private mechanism M_1, the composition M_2 M_1 satis es (ε;)-di erential privacy for any mechanism M_2 (which need not itself be di erentially private).*

The following result about di erential privacy under composition allows us to determine how the privacy level degrades when releasing multiple statistics about the same data:

Lemma 2.2 (Composition [57, 59]). *Let D be a dataset and suppose that mechanisms $M_1; :::; M_n$ satisfy (ε_i; $_i$)-di erential privacy. The mechanism computing $(M_1(D); :::; M_n(D))$ is (ε;)-di erential privacy, where $\varepsilon := \sum_{i=1}^{n} \varepsilon_i$ and $:= \sum_{i=1}^{n}$ $_i$.*

2.1.4 Privacy via Noise Addition

Di erential privacy requires randomness and is typically ensured by injecting random noise into the analysis { usually simply by adding random noise to the query result before releasing the perturbed result. Adding well-chosen noise calibrated to the sensitivity of the query ensures di erential privacy by making it hard to determine whether a speci c dataset was the input resulting in the observed output or whether the true input was a neighboring dataset.

Example: *Returning to the example from before, suppose that the number of positive results is reported with an added random value from f 1; 0; 1g so we cannot con dently determine whether patient X's test result was positive. While this noise addition does not actually ensure di erential privacy*[1]*, it gives a good idea of how we may hide patient X's test result in the statistics by adding noise.*

A fundamental question concerns the tradeo s between accuracy and privacy, and so we ask how much noise is necessary to ensure di erential privacy. Noise addition extends naturally to vector queries, where di erential privacy can be ensured by adding noise to each coordinate of the query output. Furthermore, we must select a suitable noise distribution (i.e., real, discrete, binary, etc.) for the query in question to

[1]Suppose that neighboring datasets D and D' have true number of positive test results z and z 1, resp. Observing noisy output $z + 1$ happens with non-zero probability on input D and probability 0 on input D'.

ensure di erential privacy. A common example of noise addition for real-valued queries is to use the Laplace mechanism, which adds noise from the Laplace distribution and is "-di erentially private for appropriate parameters:

Lemma 2.3 (Laplace Mechanism [60]). *For query $q : R^u ! R^k$, k 1 and input x 2 R^u, the Laplace mechanism outputs $q(x) + '$ where $'$ $Lap( )^k$. If is the sensitivity of q, the Laplace mechanism with parameter = =" is "-di erentially private.*

2.1.5 Local Differential Privacy

So far, we have mainly considered the *central* model of di erential privacy, where data is held by a single, trusted curator. In this case, the curator computes the query output, adds appropriate noise, and releases the perturbed, private query result. Consider a setting where there is no trusted, central curator, but sensitive data is distributed among multiple curators who cannot share the data for privacy concerns. As we still want to analyze the collection of (distributed) data, the privatization process must be transferred such that each curator locally applies a di erentially private mechanism before releasing a private version of their data. Statistical analysis can then be performed by an untrusted third party on the collection of private data contributions. This model is often referred to as the *local* model of di erential privacy [95, 54].

Example: *Suppose that the hospital has a new test, the patients can perform in their own homes. As the hospital still wants statistics over the number of positive results, patients are asked to register positive test results via a website. Before submitting the positive result to the hospital records, a script on the website locally adds random noise to ensure privacy: each positive result is reported with probability p. If patient X's test is positive, the result will show up in the statistics with probability p. While the nurse cannot con dently determine if the test is negative, it may be possible to determine if the result is positive. So suppose that patients also register negative tests, and these are (falsely) reported as positive with some (small) probability q, thus hiding the true positive tests among a few false positive ones. This technique is known as* randomized response *[141].*

As noise is added to each data contribution, the total amount of noise depends on the number of participants, so the accuracy is likely to su er compared to the accuracy obtainable in the central model. On the other hand, the local model does not have the trust assumptions of the central model, and the data collector is not responsible for protecting the privacy of the collected data. For these reasons, several large organizations have deployed systems using local di erential privacy, such as Google's RAPPOR [65], Apple's iOS [8, 46, 78] and Microsoft's Windows 10 [48]. In Section 2.2 we introduce two fundamental problems in data analysis and brie y describe how to solve these problems in a locally di erentially private way. In Section 2.3.3 we consider a middle ground between central and local di erential privacy, making use of cryptographic techniques to simulate the trusted central curator. For more details on local di erential privacy, we refer to the surveys [16, 41, 100, 140, 146, 148].

2.2 Statistics Over Distributed Data

In line with the distributed setting mentioned in Section 2.1.5, there are many real-world settings where the amount of data is huge and possibly distributed among multiple participants. Therefore, a query cannot immediately be answered without data sharing. Examples include data collected by search engines or apps for mobile devices collecting and analyzing data submitted by users. We introduce the results formally presented in Chapters 3 and 4 in this section and the results presented in Chapter 5 in Section 2.3.

We now discuss data structures, *sketches*, permitting accurate statistical analysis without the need to see the whole dataset. We introduce two fundamental problems, F_0 estimation and Euclidean Distance approximation, and speci c sketches that can be used to solve these problems. We are especially interested in sketch de nitions that support *combining* sketches over distributed data to obtain a sketch for the whole data, as this allows us to perform analysis over the distributed data without sharing the actual data. While

we begin this section by considering non-private sketches, we discuss how to construct di erentially private versions of in Section 2.2.4.

2.2.1 Sketches

A sketch of a dataset is a compact data structure representing (or "sketching") said dataset. Sketches are usually de ned with a particular set of statistical queries in mind, and we may (approximately) answer statistical queries about the data without seeing the entire dataset by applying a speci ed procedure to the sketch. We want sketch techniques that are simple to implement, require only little space compared to the data it represents, and can be computed e ciently. In particular, we care about the tradeo between the accuracy of the query result versus space usage and computation/update time. In order to answer queries about distributed data, we study sketches that can be combined to represent the *collection* of the data. Suppose that each curator creates a sketch of their own data, and the collection of sketches can be analyzed to learn statistical properties about the whole data. We will limit ourselves to numeric sketches, such that they can be thought of as binary or real-valued vectors. There are many examples of sketches, where the sum of two sketches is a sketch for the union of the two datasets, but of particular interest are the so-called *linear* sketches, which for input vector x is a linear mapping of x. That is, for *sketch matrix* S, Sx is the sketch of x. We remark that sketch matrix S is usually chosen at random from a suitable family of matrices to avoid consistently mapping di erent items to the same output (although, for a *xed* sketch matrix, this may still be the case). Linear sketches have the powerful property that for input vectors $x; y$ and sketch matrix S, we have $Sx \quad Sy = S(x \quad y)$ { that is, the di erence between two such linear sketches is a sketch for the *di erence* $x \quad y$.

Example: *Let us return to the example from before, although without the added noise { that is, we have a dataset over the positive tests. Suppose that the number of tests performed is enormous, and we wish to combine daily reports to get a weekly report. Instead of storing a data record for every positive test, simply store the number of positive tests each day. Adding up daily numbers gives a number for the entire week. Now, suppose that the hospital has two di erent tests: the regular one performed at the hospital and the new test taken in the patients' homes. The hospital recommends that you get both in the same week to be sure of the result. In addition to knowing the daily number of positive results, the hospital also wants to know how many people got only one positive test result to see how many of the weekly positives are not quite sure { either because one of the tests were negative or because the patient took only one test. Solving this problem requires a more elaborate sketch than simply the number of positive results, as the individuals who took both tests must be subtracted when combining sketches.*

As mentioned, sketches are often randomized and so instead of explicitly materializing the sketch matrix, which could be very large, a common approach is to de ne the transformation using appropriate, randomly chosen hash functions. We refer to [40] for a survey on sketch techniques and results.

We are now ready to discuss two fundamental problems in statistical analysis: F_0 estimation and Euclidean Distance approximation in a bit more detail. We also discuss linear sketches for solving these two problems, and while we limit ourselves to the intuition here, the sketch matrices (including di erentially private versions) are formally de ned in Chapters 3 and 4. We refer to Section 2.2.4 for the intuition behind the di erentially private sketches for these problems.

2.2.2 Problem: F_0 Estimation

A classical problem in data management and database query processing is that of computing the number of distinct items in a multiset (the *cardinality*). We formally de ne the problem as follows: let $U = [u]$ be the universe and S an input multiset over items from U. It is often convenient to represent S by a vector $x \in \mathbb{Z}^u$ where x_j counts the occurrences of each item $j \in [u]$ in S, and so the task is to compute $\|x\|_0 = \sum_{j \in U} 1[x_j \neq 0]$ (the problem is therefore also referred to as F_0: the 0^{th} frequency moment). As we are interested in estimating the size of the symmetric di erence between *sets*, we consider input set S with

the characteristic vector $x \in \{0,1\}^u$ where $x_j = 1$ if and only if $j \in S$ for each item $j \in [u]$. The observant reader will notice that the example given in Section 2.2.1 concerns this problem.

For large datasets, we wish to solve the problem without explicitly enumerating the set. Therefore, we represent data by a sketch that allows us to *estimate* the cardinality with a tradeoff in accuracy versus space. Well-known sketches for cardinality estimation include HyperLogLog, FM-sketches and bottom-k sketches, but although these sketches can all be merged to give a sketch for the union of the input data, they are not linear, and so cannot be subtracted to give a sketch for the difference between the two input datasets, which is the main application of interest. We now introduce a linear sketch (over the field of two elements, GF(2)) which can be used to accurately estimate the cardinality of the input set while ensuring that the difference of two such sketches is a sketch for the symmetric difference of the input sets.

The KOR Sketch

Consider our main application, where we want to estimate the size of the symmetric difference of datasets held by two different curators. There is no way of knowing the size of the symmetric difference before releasing the individual sketches, so one difficulty is to decide on a suitable sketch size. While we want the sketch to be as small as possible, a sketch that is too small compared to the input set (the symmetric difference) will be overfilled, leading to inaccurate estimates due to hash collisions (too much information is lost). On the other hand, recalling that we will add noise to each coordinate of the sketch to preserve privacy, the signal of the input set will drown in noise if the sketch is too large. In Chapter 3 we define a linear sketch building on a technique by Kushilevitz, Ostrovsky and Rabani [99] (STOC 1998) – therefore named *the KOR sketch* – which allows us to fix an appropriate sketch size, without knowing the input size. Intuitively, the idea is to fix a sketch size and compute $\log(u)$ identical (partial) sketches for samples of decreasing size of the input: for $i = 0, \ldots, \log(u) - 1$, one can think of the input to the i^{th} partial sketch as a sample whose size is approximately a $1/2^{i+1}$ fraction of the size of the (whole) input set. As the size of the (sampled) input sets is halved at each step, the fixed sketch size is suitable for at least one of the samples and thereby permits an accurate estimate of the size of this particular sample. Subsequently, one can correct this estimate for the sampling step to get an accurate estimate for the whole (unsampled) input set (with high probability). The KOR sketch is the collection of the $\log(u)$ partial sketches.

The $\log(u)$ samples are created using a standard hashing-based subsampling technique (see for example [144]) to ensure that each invocation of the sketch samples identically. This is important for the application of estimating the size of the symmetric difference, as identical items in two input sets must either be sampled from both input sets or not at all: suppose $x_s, y_s \in \{0,1\}^u$ are samples of input sets $x, y \in \{0,1\}^u$. The sum $x_s + y_s$ over GF(2) must represent (a sample of) the symmetric difference of x and y. We refer to Section 3.4.1 for the details and formal description of the KOR sketch.

2.2.3 Problem: Euclidean Distance Approximation

A well-known problem in various data analysis applications is approximating the Euclidean distance between two real-valued vectors. Euclidean distance is a useful measure for similarity of two vectors and finds applications in fields such as nearest-neighbor search [3, 87], computational geometry [35], document comparison [134], data streams [86], clustering [22, 38], graph sparsification [128], low-rank approximation [37], numerical linear algebra [36, 52, 145] and many more. We concern ourselves with *squared* Euclidean distance, but as a solution to one problem implies a solution to the other, we do not explicitly make a distinction.

Example: *Suppose that the two tests for diabetes both give a score between 0 and 1, indicating the severity of the disease, instead of the positive/negative answer, as has been discussed so far. The hospital wishes to compare the results of the take-home test with the results of the test performed at the hospital, and so each patient takes both tests and receives a score from each. The hospital collects all of the test scores in two data vectors and computes their Euclidean distance to measure the similarity of the test results.*

For input vectors, $x; y \in \mathrm{R}^u$, the (squared) Euclidean distance between x and y is de ned as

$$kx \quad yk_2^2 = \sum_{j=1}^{u} (x_j \quad y_j)^2:$$

As the input dimension u may be very large, we are once again interested in techniques for estimating the Euclidean distance between two vectors without sharing the entire input { i.e., we want a sketch that (approximately) preserves Euclidean distance. The Johnson-Lindenstrauss transformations are a natural choice for such a sketch.

Johnson-Lindenstrauss Transforms

The Johnson-Lindenstrauss transformations are a family of linear transformations that (with high probability) preserve $`_2$-norm of real input vectors up to a small error while reducing the dimension of the input vectors. Speci cally, Johnson-Lindenstrauss matrices are linear maps satisfying the following lemma:

Lemma 2.4 ((Distributional) Johnson-Lindenstrauss Lemma [90])**.** *For any* $0 < \ ; p < 1=2$ *and any input dimension* $u > 0$*, there exists a random* $k \quad u$*-matrix* S *where* $k = O(\ ^2 \log(1=p))$ *such that for any input* $x \in \mathrm{R}^u$

$$(1 \quad)kxk_2^2 \quad kSxk_2^2 \quad (1 + \ )kxk_2^2$$

with probability at least $1 \quad p$*.*

As $kSx \quad Syk_2^2 = kS(x \quad y)k_2^2$, the lemma implies that for input vectors $x; y \in \mathrm{R}^u$ we have

$$(1 \quad)kx \quad yk_2^2 \quad kSx \quad Syk_2^2 \quad (1 + \ )kx \quad yk_2^2:$$

Thus, it su ces to release the sketches Sx and Sy to allow for Euclidean distance estimation.

Johnson-Lindenstrauss matrices are particularly interesting due to the remarkable result by Jayram & Woodru [89] and later Kane et al. [92] stating that the optimal output dimension is $k = (\ ^2 \log(1=p))$, and so depends only on the accuracy parameter and the error probability p { not on the input dimension u.

The literature suggests several families of linear transformations satisfying the Johnson-Lindenstrauss Lemma [1, 3, 43, 87, 93], but of particular interest are very sparse matrices as they allow us to compute sketches fast. In Chapter 4 we de ne and discuss the Sparser Johnson-Lindenstrauss transformation by Kane & Nelson [93] which has sparsity $s = O(\ ^1 \log(1=p))$ { i.e., each column has at most s non-zero entries. In Section 2.2.4 we discuss how to achieve a di erentially private version of the Sparser Johnson-Lindenstrauss matrix and give an introduction to the results presented in Chapter 4.

As Johnson-Lindenstrauss matrices are known from the dimensionality reduction literature, we follow convention and refer to them as *projections* and *transformations* rather than *sketches* in Chapter 4.

2.2.4 Differentially Private Sketches

As touched upon in Section 2.1.4, we can compute di erentially private versions of linear sketches by adding suitable noise to each coordinate of the sketch. Such private sketches can be released to allow for analysis, and the post-processing property of di erential privacy (Lemma 2.1) ensures that since the sketch is di erentially private, then any estimator based on that sketch is also di erentially private. Therefore, an untrusted analyst can use the privately released sketches to compute answers for statistical queries, and in particular, the analysis can be performed o ine { that is, without the curators necessarily being active at the time of analysis. This property is particularly valuable in a distributed setting where participants release private sketches of their own data, which can be merged (o ine) to permit analysis over the collection of data, or if data is created at di erent times as suggested in the following example:

Example: *Recall the example in Section 2.2.1 where we wanted (non-private) daily sketches which could be combined to a weekly sketch. Adding noise to the daily sketches lead to the questions:*

1. *Is the daily sketch robust against the noise, and so still allows for accurate approximations of the daily number of positive results after noise addition?*

2. *Can the noisy daily sketches be combined into a noisy weekly sketch?*

Similar to our interest in linear sketches that could be merged in Section 2.2.1, we now take the idea further and consider *private* sketches that may be combined to give a *private* sketch for the combined data. The idea is simple: for non-private linear sketch S, input vector x and random noise vector $'$ from a suitable distribution, we can build a di erentially private sketch $Sx + '$. For inputs $x; y$ and noise $'; \ ,$ we have

$$(Sx + ') + (Sy + \ ) = S(x + y) + (' + \ ) \quad \text{and} \quad (Sx + ') \quad (Sy + \ ) = S(x \quad y) + (' \quad):$$

Hence, adding or subtracting such noisy linear sketches gives a noisy sketch for $x \quad y$ with noise $' \quad .$ While this observation immediately leads to the question: *what accuracy guarantees can we get from the combination of noisy sketches, when the noise increases when combining sketches?* (this question is discussed further in Chapter 3), we will in this thesis only consider combining *two* sketches. Note that we can explicitly determine the privacy guarantees of the sketch resulting from combining two noisy sketches.

An important property of using the same sketch matrix for di erent inputs is that combining two such sketches gives a similar sketch for a function of the input sets. Therefore, to use two sketches to build a sketch for the di erence of the inputs, it is essential that the sketch matrix (or equivalently, the hash function de ning the sketch matrix) is public, and only the noise is kept secret.

We now give the intuition behind the private sketches as well as introduce our results presented formally in Chapters 3 and 4.

Differentially Private F_0 Estimation

In Chapter 3 we consider a weighted generalization of cardinality estimation, where we are also given a public weight vector $w \ 2 \ (0; 1]^u$, such that each item $j \ 2 \ [u]$ has a weight, w_j. Given input set $S \quad U$ we aim to estimate the weight of S, i.e., $\sum_{j \ 2 S} w_j$. In terms of the characteristic vector for S, $x \ 2 \ f0; 1g^u$, we must estimate $kx \quad wk_1$, where denotes the Hadamard (i.e., entrywise) product. Note that this problem reduces to standard cardinality estimation in the important case where the weights are all 1, $w = (1; : : : ; 1)$. We here give an intuitive description of how to construct sketches for weighted F_0 that are small, di erentially private, and computationally e cient and refer to Chapter 3 for a formal de nition as well as the analysis of the privacy, accuracy, size, and time guarantees.

As discussed above, we can make a linear sketch di erentially private by adding appropriate noise to each entry of a linear sketch. The KOR sketch from Section 2.2.2 can be represented by a binary vector, and we privatize it using the standard privacy technique *randomized response* [141], which has become one of the main techniques in local di erential privacy due to the high scalability. The technique was rst introduced by Warner in 1965 as a way of privately conducting surveys about embarrassing or illegal questions and is best explained via an example:

Example: *Suppose that we study the fraction of individuals with some property P by asking a group of people whether or not they have property P. As P may be sensitive, each participant ips a coin before answering the question. If the coin comes up heads, the participant answers the question truthfully, and if the coin comes up tails, the participant ips the coin once more. If on the second toss, the coin shows heads, the participant answers "yes" and on tails, they answer "no". From the collected, private answers, one can estimate the fraction of people with property P while not being able to determine whether any speci c individual has property P or not. Recall the example from Section 2.1.5, where we reported each positive test result with a probability p and each negative result with probability q. This noise addition was, in fact, randomized response.*

In our private sketch for F_0 estimation we use randomized response as follows: Let $Sx \ 2 \ f0; 1g$ be the (non-private) KOR sketch for input $x \ 2 \ f0; 1g^u$ and ip each bit in Sx independently with probability

$p(\varepsilon) < 1/2$ depending on the privacy parameter ε. Then, each bit in the sketch is reported falsely with probability $p(\varepsilon)$. These bit flips can be expressed in terms of a noise vector $\eta \in \{0,1\}^u$ where $\Pr[\eta_j = 1] = p(\varepsilon)$ independently for each entry $j \in [u]$ and so we define the *noisy KOR sketch* as $Sx + \eta$, where the noise addition happens over GF(2). We prove in Chapter 3 that the resulting sketch is ε-differentially private for an appropriate choice of p as a function of ε.

Moreover, for every choice of accuracy parameter $0 < \alpha < 1$ and privacy parameter $\varepsilon = O(1)$ there exists a distribution H of linear sketches of size $k = O(\log^2(u)\varepsilon^{-2}\alpha^{-2})$ such that $Sx + \eta$ is ε-differentially private for $S \sim H$. Given $Sx + \eta$ we can compute an estimate of $\|x - w\|_1$ that is accurate within a factor $1 \pm \alpha$, plus additive error $O(\log(u)\varepsilon^{-2}\alpha^{-2})$. In the unweighted case, recent work [4, 129, 65, 138] has shown how to privately and efficiently compute an estimate for the size of the symmetric difference between two sets using (non-linear) sketches such as FM-sketches and Bloom Filters. These methods have an error bound depending on an upper bound m on the size of the input sets and achieve error no better than $O(\sqrt{m})$. Hence, for $\alpha = o\left(1/\sqrt{m}\right)$ and $\log(u)/\varepsilon = m^{o(1)}$, the noisy KOR sketch improves over these results. Letting m denote the size of the symmetric difference, McGregor et al. [103] show that an additive error of $\tilde{\Omega}(\sqrt{m}/e^{\varepsilon})$ (where the $\tilde{\Omega}$ notation suppresses a polylogarithmic factor) is necessary to estimate m under the constraint of ε-differential privacy. In contrast, setting $\alpha = \sqrt[3]{\log(u)/(\varepsilon^2 m)}$ to balance the relative and additive error, our noisy KOR sketch achieves error $\tilde{O}(m^{2/3}/\varepsilon^{2/3})$.

In order to efficiently privatize our sketch, it is essential that we can compute the noise efficiently. While Mir et al. [110] and Kenthapadi et al. [96] also studied differentially private sketches for F_0 estimation, their algorithms for calibrating noise are not computationally efficient: Kenthapadi et al. [96] computes the sketch via a Johnson-Lindenstrauss matrix (see Section 2.2.3) where all entries are i.i.d. normally distributed and add Gaussian noise to the sketch. To calibrate the noise to the sketch, one must first compute the sensitivity of the sketch matrix, which requires time superlinear in the universe size u. Mir et al. [110] also compute a standard linear sketch but add noise via the exponential mechanism, which requires quasipolynomial time in the sketch size. In comparison, we can sample noise in time proportional to the size of the sketch. We remark that for a fixed ε, the size of the KOR sketch is polynomially related to the lower bound of $\Omega(\log(u)\alpha^{-2})$ bits by Jayram & Woodruff [89].

We formally prove these results in Chapter 3 where we also discuss how to use the noisy KOR sketch for a distributed streaming implementation for estimating the size of the union between two input streams. We conclude the chapter by mentioning a few open questions related to private F_0 estimation.

Differentially Private Euclidean Distance Approximation

This section introduces the techniques and results presented formally in Chapter 4. Euclidean distance approximation was introduced in Section 2.2.3, and we recall that the Johnson-Lindenstrauss matrices suggest (non-private) sketches allowing us to solve this problem accurately. In line with the rest of this chapter, we use the terms *matrix*, *sketch* and *transformation* interchangeably. In Chapter 4 we use projection and transformation to be consistent with the literature on Johnson-Lindenstrauss matrices. Our main focus will be on the Sparser Johnson-Lindenstrauss transform by Kane & Nelson [93], while we also discuss the Fast Johnson-Lindenstrauss transform by Ailon & Chazelle [3]. We prove that such a sketch is robust against noise and permits accurate distance approximation from sketches that have been privatized via noise addition. We recall the Laplace mechanism from Lemma 2.3 and define the Gaussian mechanism, which adds Normally distributed noise:

Lemma 2.5 (Gaussian Mechanism [57, 63])**.** *For query $q : \mathbb{R}^u \to \mathbb{R}^k$, $k \geq 1$ and input $x \in \mathbb{R}^u$, the Gaussian mechanism outputs $q(x) + \eta$ where $\eta \sim N(0, \sigma^2)^k$. For $\varepsilon \in (0,1)$, if $\sigma \geq \Delta_2 \sqrt{2\ln(1.25/\delta)}/\varepsilon$ is the ℓ_2-sensitivity of q, the Gaussian mechanism with parameter $\sigma^2 \geq 2\Delta_2^2 \ln(1.25/\delta)/\varepsilon^2$ is (ε, δ)-differentially private.*

Consider input vector $x \in \mathbb{R}^u$ and Johnson-Lindenstrauss matrix S. As the sketch Sx is real-valued, applying either the Laplace or the Gaussian mechanism ensures differential privacy of the noisy sketch, but one may ask which mechanism gives better error guarantees. Specifically, we suggest an ε-differentially

private sketch based on the Sparser Johnson-Lindenstrauss transform (SJLT) [93] with added Laplace noise: For any $0 < \ ; p < 1=2$ and $u > 0$ there is a distribution H over $k \quad u$-matrices with $k = \quad \left(\quad ^2 \log(1=p) \right)$ and sparsity $s = O\left(\quad ^1 \log(1=p) \right)$ and a distribution D over R such that for $S \quad H$, $x \ 2$ R^u and $' \quad D^k$, the sketch $(S; Sx + \ ')$ is $"$-di erentially private and can be computed in time $O(skxk_0 + k)$, including the time to sample the noise. For $y \ 2$ R^u and $\quad D^k$, $k(Sx + \ ') \quad (Sy + \ )k_2^2 \quad 2ks="^2$ is an unbiased estimator for $kx \quad yk_2^2$. Previously, Kenthapadi et al. [96] de ned a $(";\ )$-di erentially private sketch based on the Johnson-Lindenstrauss matrix by Indyk & Motwani [87], where all entries are i.i.d. normally distributed with added Gaussian noise and use this noisy sketch to obtain an unbiased estimator for Euclidean distance approximation with low variance. Besides having a better privacy guarantee, our sketch is also faster to compute due to the sparsity of the sketch matrix, and our estimator has lower variance for $< p^{O(1=\ )}$ than that of Kenthapadi et al. (we spare the reader the expression for the variance here, but refer to Chapter 4, where we also dive into the formal proofs). For larger values of , we may apply the Gaussian rather than the Laplace mechanism to compute a sketch with similar accuracy and privacy guarantees as that of Kenthapadi et al. while still being more e cient.

Another important property of our sketch is that it avoids a sizable initialization cost inherent to the results in [96], as a consequence of their choice of sketching matrix: in order to calibrate the noise to the sketch matrix, one must rst compute the sensitivity of the sketch, which requires superlinear time in the size of the universe. The Sparser Johnson-Lindenstrauss matrix has a xed sensitivity, so the noise can immediately be calibrated to this known sensitivity. We provide all of the proofs and arguments in Chapter 4, where we also describe a di erentially private version of the Fast Johnson-Lindenstrauss Transform (FJLT) by Ailon & Chazelle [3] with added Gaussian noise. This private sketch based on FJLT o ers a tradeo in speed for variance for certain settings of the parameters $u;\ $, and p compared to the private SJLT.

Differentially Private F_0 Estimation Revisited

We remark that our private sketch for Euclidean distance approximation discussed above (and formally presented in Chapter 4) can also be applied to estimate the cardinality of the symmetric di erence between two input sets: For binary vectors $x; y \ 2 \ f0; 1g^u$ we have that $Sx \quad Sy = S(x \quad y)$ is the Johnson-Lindenstrauss transformation of the symmetric di erence $x \quad y$. As S is a Johnson-Lindenstrauss transformation, $kSx \quad Syk_2^2$ accurately approximates $kx \quad yk_2^2$, which in turn is the same as $kx \quad yk_0$, since x and y are binary vectors. Thus, the results from Chapter 4 can be applied to estimate the size of the symmetric di erence, m, between inputs x and y with error (std. deviation) $O(m=\sqrt{k} + \sqrt{sm}=" + s\sqrt{k}="^2)$. We remark that for $= \sqrt[3]{\log(u)=("^2 m)}$ (as we chose to balance the relative and additive error of our noisy KOR sketch), this error is proportional (suppressing polylogarithmic factors) to $m^{2=3}="^{2=3}$, hence we get error guarantees comparable to those of the noisy KOR sketch.

2.3 Revisiting Privacy via Noise Addition

We now return to the question of how much noise is necessary to preserve di erential privacy as touched upon in Section 2.1.4. Although we are still interested in distributed data, we discuss a few di erentially private mechanisms in the central model of di erential privacy before moving on to the distributed setting. We consider mainly single real-valued queries. While the Laplace mechanism is asymptotically optimal for $"$-di erential privacy when $" \ ! \ 0$, we study what happens for larger values of $"$ { that is, for a low level of privacy. Geng, Kairouz, Oh & Visvanath [69, 70] introduced the *Staircase mechanism* and proved that this noise distribution is optimal for answering real-valued queries $"$-di erentially privately. In Section 2.3.2 we introduce a new $"$-di erentially private mechanism, the *Arete mechanism*, and compare it to the Staircase mechanism. The Arete mechanism adds noise from our new Arete noise distribution. We go into further detail with the Arete mechanism in Chapter 5.

In Section 2.3.3 we return to a distributed setting: In the previous section, we ensured di erential privacy by adding su cient noise to each data release (each sketch). That is, we considered the local model of di erential privacy. We now discuss how to answer aggregate queries privately without requiring *local*

di erential privacy. The idea is to aggregate the data contributions securely using cryptographic techniques such that the aggregator sees only the aggregate query result and not the individual data contributions. Then each participant needs only add noise *shares* to their data contribution. During aggregation, the noise shares add up to noise that masks the (aggregate) query output. This way, we can lower the noise level to that of the central model. We describe how noise from the Arete distribution in *in nitely divisible* and so can be divided among n participants such that the aggregate result has the same privacy and error guarantees as applying the Arete mechanism in the central model.

2.3.1 The Staircase Mechanism

Geng et al. [70, 69] presented the ε-di erentially private Staircase mechanism for real-valued queries, which adds noise from their new Staircase distribution. We limit ourselves to discussing single-dimensional real-valued queries as handled in [70] and refer the reader to [69] for multi-dimensional queries. The Staircase distribution with parameters ε, 2 $[0; 1]$ and query sensitivity > 0 is a mixture of uniform distributions, and so the density function of the Staircase distribution, f_{SC}, is a piece-wise constant function (See Figure 2.2(b)) de ned as follows:

$$f_{SC}(t) := \begin{cases} a( ); & t \ 2 \ [0; \quad) \\ e^{\ \varepsilon} a( ); & t \ 2 \ [\quad ; \quad) \\ e^{\ k\varepsilon} f_{SC}(t \quad k \quad); & t \ 2 \ [k \quad ; (k+1) \quad); \ k \ 2 \ \mathbb{N} \\ f_{SC}(\quad t); & t < 0 \end{cases}$$

where $a( ) = (1 \quad e^{\ \varepsilon})=(2 \ (\ + e^{\ \varepsilon}(1 \quad)))$ is a normalizing constant, ensuring that f_{SC} de nes a probability measure { i.e., that $\int_{\mathbb{R}} f_{SC}(t)dt = 1$.

The Staircase mechanism places the majority of the probability mass in a uniform distribution on an interval of length (for optimal) $2 \quad = \quad e^{\ (\varepsilon)}$ around zero and only probability mass

$$2\sum_{k=1}^{1} a( )e^{\ \varepsilon k} \quad = O(\quad {}^{1}e^{\ \varepsilon}) = e^{\ (\varepsilon)}$$

in the tails. Geng & Viswanath prove that this noise distribution is optimal for single real-valued queries. While the Laplace mechanism is by now the standard choice of mechanism to achieve ε-di erential privacy, the Staircase mechanism has error exponentially decreasing in ε, and so ensures better accuracy than the Laplace mechanism for large ε. For $\varepsilon \ ! \ 0$, the Staircase distribution approaches the Laplace distribution. Speci cally, the Staircase distribution has (absolute) noise magnitude $(\ e^{\ \varepsilon=2})$ (for the optimal choice of), while the Laplace distribution has noise magnitude $(\ =\varepsilon)$.

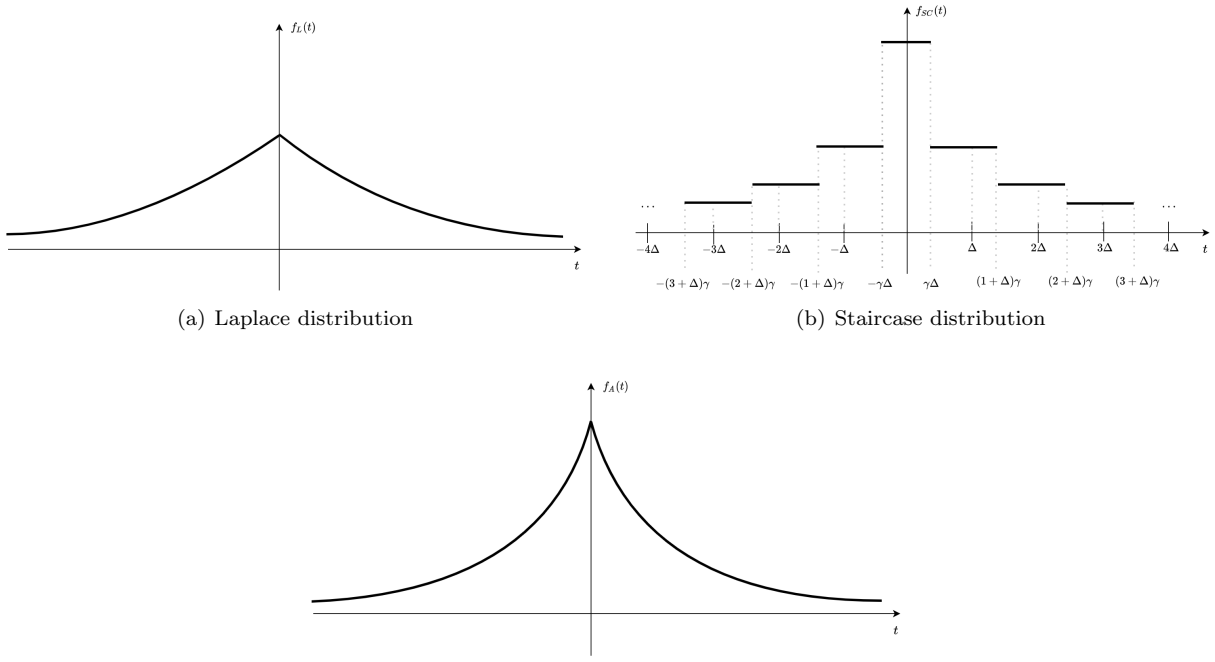
2.3.2 The Arete Mechanism

In Chapter 5 we present our new noise distribution, the Arete[1] distribution, and a mechanism for single real-valued queries which adds noise from this distribution, the *Arete mechanism*. We prove that for suitable parameters, the Arete mechanism is ε-di erentially private.

The Arete Distribution

Intuitively, the Arete distribution can be thought of as a symmetric -distribution, mirrored at zero, which is attened slightly to avoid the singularity at 0 (See Figure 2.2(c)). More precisely, a random variable with the Arete distribution for parameters $; \ ; \ > 0$ has the same distribution as $Z := X_1 \quad X_2 + Y$, for independent -distributed random variables $X_1; X_2 \quad (\ ; \ )$ and Laplace distributed random variable $Y \quad Lap(\ )$, where the latter is added to atten the distribution.

[1] The name *Arete* is inspired by the word arête (pronounced "ah-ray't"), which is both a sharp-crested mountain ridge, while also a concept from Greek mythology, Arete (pronounced "ah-reh-'tay") referring to moral virtue and excellence: the notion of the fulfillment of purpose or function and the act of living up to one's full potential [142].



(a) Laplace distribution

(b) Staircase distribution

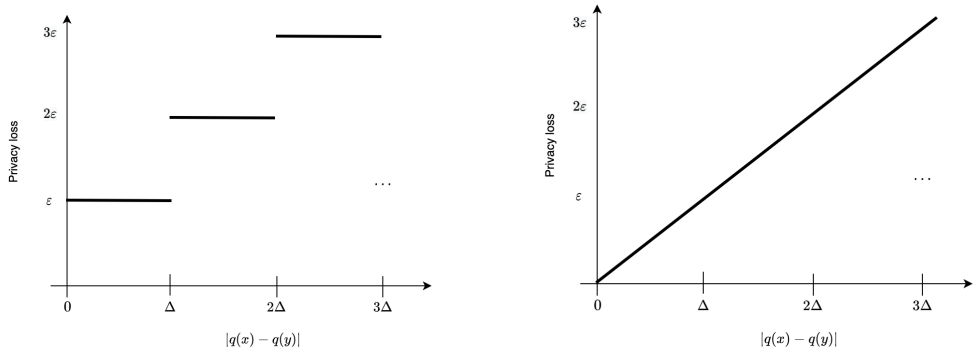
(c) Arete distribution

Figure 2.2: Illustration of the density functions for the Laplace, Staircase and Arete distributions. The purpose of this figure is to give an *intuitive* idea of the shapes. See Figure 5.2 for plots showing the densities of the Arete vs. Laplace distributions for certain parameter settings.

Properties

The Arete distribution has a continuous density function which is symmetrically and monotonely decreasing { that is, the density function is symmetric around 0, and monotonely decreases for t 0. The distribution also has low expected (absolute) value for suitable parameters, exponentially decreasing in $"$, and so can be considered a counterpart to the Staircase distribution with a continuous density function and comparable magnitude of the noise. Speci cally, the Staircase mechanism has expected error ($e^{"=2}$) for the optimal parameter setting, while there exists a parameter setting (not necessarily optimal) such that the Arete mechanism has expected error $O(e^{"=4})$, where is the sensitivity of the query. Furthermore, the Arete mechanism has variance $O({}^2e^{"=4})$ for the *same* parameter setting, while the Staircase mechanism achieves variance (${}^2e^{2=3"}$) for the parameter setting optimizing for variance. We remark that for there is not generally a parameter setting simultaneously optimizing for both error and variance for the Staircase mechanism.

Apart from having error and privacy guarantees comparable to those of the Staircase mechanism, the Arete mechanism has a few additional desirable properties. As brie y mentioned, the Staircase distribution has a piece-wise constant density function, and so the privacy guarantee worsens in a step-wise manner for query outputs that are slightly further apart than the sensitivity (i.e., for inputs that are almost but not quite neighboring). Although the di erential privacy guarantee does not need to be satis ed for such inputs, it is still relevant to study what privacy guarantees can be made for inputs that are not quite neighbors. For the Staircase mechanism, the privacy guarantee is immediately (at least) halved: suppose that $jx \quad yj = 1 +$ for an arbitrarily small . If $jq(x) \quad q(y)j >$, we may observe a noisy query output $= q(x) + ' = q(y) +$ such that $f_{SC}(') = a( )e^{"k}$ and $f_{SC}( ) = a( )e^{"(k+2)}$. The Arete distribution has a continuous density function, and so the privacy guarantee deteriorates more smoothly (see Figure 2.3). Furthermore, the Arete distribution was inspired by the search for an alternative to the Staircase mechanism, which could be applied



(a) Staircase mechanism (b) Arete mechanism

Figure 2.3: Illustrations of the worst-case privacy loss of the Staircase and Arete mechanisms depending on the difference between query outputs. As we have no closed form for the density of the Arete distribution, we cannot explicitly determine the privacy loss, so the graph given here is an approximation.

in a distributed setting while keeping the error low. We brie y touched upon in nite divisibility earlier in this section but will discuss this key property of the Arete distribution a bit further in the next section.

2.3.3 Differential Privacy and Cryptographic Primitives

Consider the simple but important problem of real summation in a setting where an untrusted aggregator can sum data contributions x_j 2 R from n participants without looking at each individual contribution, thus only seeing the result of the aggregation. If we can add random noise to the contributions, which adds up to appropriate noise for the *query output* during the aggregation, we can achieve the noise level from the central model with the trust assumptions of the local model. The lines of work on secure multiparty aggregation [77, 108] and the shu e model of di erential privacy [19] allow an untrusted unit to aggregate data contributions without seeing the individual contributions using cryptographic techniques. We describe the strategy using secure multiparty aggregation: Suppose that n participants add a bit of noise (a noise share) $'_j$ to their own data contribution x_j before submitting $x_j + '_j$ to the aggregator. The noise share $'_j$ does not need to ensure di erential privacy of $x_j + '_j$, as the data contributions are shared via a secure channel and the aggregator sees only the result $\sum_{j=1}^{n}(x_j + '_j) = \left(\sum_{j=1}^{n} x_j\right) + '$ for $' := \sum_{j=1}^{n} '_j$. We want noise shares $'_j$ such that the aggregated noise $'$ ensures di erential privacy of the query result. Such noise shares can be chosen for *in nitely divisible* distributions: Intuitively, a distribution D is in nitely divisible if, for any n 2 N, we can express random variable X D as a sum of n independent, identically distributed random variables. That is, if for any n 1 there exist i.i.d. random variables $X_1; : : : ; X_n$ such that $\sum_{j=1}^{n} X_j$ has the same distribution as X. Note that X_i need not have distribution D.

As the Laplace distribution is in nitely divisible (a Laplace distributed random variable can be described as the sum of di erences between -distributed random variables), secure multiparty aggregation [2, 77] and the shu e model [73] have previously been used together with Laplace noise to get essentially the same error guarantees as in the central model for real summation. Since the -distribution is also in nitely divisible (as a -distributed random variable can be expressed as the sum of -distributed random variables) and recalling that a random variable with the Arete distribution can be expressed as $Z = X_1$ $X_2 + Y$ for -distributed $X_1; X_2$ and Laplace distributed Y, the Arete distribution is also in nitely divisible { this observation is formalized in Chapter 5. That is, one can choose noise shares that sum to Arete distributed noise during aggregation, and so we can simulate an application of the Arete mechanism in the central model of di erential privacy by adding such noise shares to the data contribution of each participant. We note that the Staircase distribution is not in nitely divisible as it is a mixture of uniform distributions, and so we cannot simulate the Staircase mechanism using a secure channel.

We give a formal de nition of in nite divisibility as well as proofs that the Laplace, , and Arete distributions are in nitely divisible in Chapter 5. We then also argue how to combine secure multiparty aggregation and the shu e model with the Arete mechanism to solve two open questions from previous work: [76] studies private real summation with several di erent secure multiparty aggregation protocols, and [73] applies the shu e model to sum real numbers privately. Both of [73, 76] use noise shares adding up to Laplace noise. Exchanging these noise shares with shares adding up to a random variable with the Arete distribution, we can achieve error exponentially decreasing in " for queries over distributed data while avoiding a dependency on the number of participants, inherent in the local model of di erential privacy.

2.3.4 Privacy Amplification Techniques

The Staircase and Arete mechanisms signi cantly improve on the error for large values of " { that is, for low privacy guarantees. Suppose we want to *increase* the privacy level of such query results. In that case, we now consider two techniques for privacy ampli cation: *privacy ampli cation by sub-sampling* [10], where the privacy level is increased by considering only a sample of the input and *privacy ampli cation via shu ing* [64], where a trusted shu er ensures that the data contributions are anonymized before being forwarded to the analyzer. Balle et al. [10] give a general technique to amplify privacy, showing that applying a di erentially private mechanism to only a sample of a dataset rather than the whole dataset gives better privacy guarantees: Let M be an $(\varepsilon;)$-di erentially private mechanism and $M_S := M \quad S$ be the composition of M with sample technique S. Then M_S is $(\varepsilon^0; h( ))$-di erentially private for an $0 < \varepsilon^0 \quad \varepsilon$ and some function h of . We refer to [10] for the details and results concerning di erent variations and sampling techniques, but as an example, Balle et al. show that if S randomly (and without replacement) samples m items from the input set of size $n \quad m$, then, $\varepsilon^0 = \log(1 + (m=n)(e^{\varepsilon} \quad 1))$ and $h( ) = \frac{m}{n}$. Erlingsson et al. [64] show that a (permutation invariant) mechanism satisfying ε-di erential privacy (for $\varepsilon < 1$) in the local model also satis es $O(\varepsilon\sqrt{\log(1=)=n;})$-di erential privacy in the shu e model [19]. One advantage of this result is that it suggests a method to achieve privacy ampli cation while still analyzing the entire dataset instead of only a subsample, which was the case for the work of Balle et al.

Chapter 3

Differentially Private F_0 Estimation

This chapter is based on the paper *E cient Di erentially Private F_0 Linear Sketching* (ICDT 2021) by Rasmus Pagh & Nina Mesing Stausholm [119].

3.1 Introduction

Estimating the number of distinct values in a set (its *cardinality*) without explicitly enumerating the set is a classical and fundamental problem in data management. Sampling-based methods [81] can in many cases be improved by using algorithms designed with data streams in mind [94]. Streaming algorithms based on *linear sketches* can also be used to estimate changes as a dataset evolves [97] and for approximate query processing in distributed settings [6, 40]. As a motivating example, consider the following SQL query:

```
SELECT P.name
FROM PATIENTS P, DIABETES_RESULTS D
WHERE P.egg_allergy = true AND P.name = D.name AND D.result = positive
```

The size (in bytes) of the query result is a sum weighted by string length over the names that appear in subsets of two relations. That is, estimating the size of the join result is about estimating the *weighted* size of a set intersection.

In the example above, the information that a tuple with a particular person exists (and satis es a speci c predicate) can potentially be sensitive: as an allergy to egg is fairly uncommon, one may be able to identify an individual with a positive test result. If the database is distributed, with relations on di erent servers that are not allowed to expose sensitive information, it is not trivial how to estimate this join size.

Di erential privacy is often considered the gold standard for providing rigorous privacy guarantees, and while it is known to come with pitfalls [98], work in the database community has led to privacy-preserving database systems supporting (limited) SQL, see e.g. [106, 143] and their references. A challenge in such systems is that the set of queries is often not known ahead of time, so *budgeting* the disclosure of detailed information is highly non-trivial. An appealing approach to privacy-preservation, even when faced with unknown queries, is to release a private sketch of the dataset from which we can compute approximate query answers (as a side e ect, this also eliminates the need for interaction, allowing for o ine query results). In this chapter, we will consider private linear sketches for the problem of weighted cardinality estimation:

Consider two players with sets A and B from a universe $U = [u]$, resp. For every element $j \in U$ let $w_j \in (0;1]$ be a xed, public weight and for input set $A \subseteq U$ consider the corresponding weight vector w_A with $(w_A)_j = w_j \cdot \mathbf{1}[j \in A]$ for each $j \in U$. The goal is to estimate the weight of the symmetric di erence $\|w_{A \triangle B}\|_1$, in a di erentially private way. In this chapter, we show how to solve this problem using the following idea: a player computes a *linear sketch* of the input set, from which it is possible to estimate the weight of the set. Before releasing the sketch, the player adds noise to each entry of the sketch to ensure

that the sketch is differentially private. We prove that the sketch is robust against noise, i.e., that one may accurately estimate the weight of the input set from the private sketch, and argue that the sum of two such sketches is a sketch for the symmetric difference. We formalize the idea in Section 3.4 and remark that if $w_j = 1$ for all $j \in U$, then the problem reduces to estimating the set *size*, a problem often referred to as F_0, and so the weighted version can be considered a generalization of F_0.

If we, along with the estimate of the weight of the symmetric difference $kw_{A \triangle B}k_1$, have estimates of kw_Ak_1 and kw_Bk_1, then one can also estimate $kw_{A \cup B}k_1$, $kw_{A \setminus B}k_1$, $kw_{A \cap B}k_1$ and $kw_{B \setminus A}k_1$ as argued in Section 3.4.3. To facilitate such estimates, each player also outputs a differentially private version of their set weight. As it is not clear how to estimate $kw_{A \triangle B}k_1$ from kw_Ak_1 and kw_Bk_1, it seems insufficient to have each party simply compute and release private versions of kw_Ak_1 and kw_Bk_1.

Our main results in this chapter are that we define and construct a noisy linear sketch over GF(2), the field of size 2, with the following properties:

- ε-differentially private
- Computationally efficient
- Allows estimating the weight of the symmetric difference with small relative error
- Space usage is polynomially related to the lower bound (for fixed ε)

Previously known results satisfy at most 3 of these properties – see Figure 3.1 for an overview. We discuss previous work further in Section 3.2. Our sketch can be computed and stored for offline computations, so two players need not be active simultaneously but can release their sketches when ready. A self-contained description of our linear sketch can be found in Section 3.4. Readers familiar with the sketching literature will realize that our sketch combines a method of Kushilevitz, Ostrovsky, and Rabani [99] with a standard hashing-based subsampling technique (see, e.g., [144]), and so refer to our (non-private) sketch as the *KOR sketch*. We use a Randomized Response Technique [141] with noise parameter $p(\varepsilon) < 1/2$, and show in Section 3.5.1 how to choose $p(\varepsilon)$ to ensure ε-differential privacy for the sketch. This noisy counterpart is referred to as the *noisy KOR sketch*. We generally leave out ε in the noise parameter and write simply p. The bulk of our analysis is proving that the KOR sketch is sufficiently robust against noise to allow precise estimation from the differentially private sketch. We show in Section 3.5.2 that the noisy KOR sketch permits a $(1 + \delta)$-approximation for the weight of the input set with high probability and argue in Section 3.4.3 how to privately estimate the weight of the symmetric difference from two noisy KOR sketches. A related but non-linear and non-private sketch has previously been used for estimating the size of symmetric difference by Mitzenmacher et al. [114]

For convenience, we often represent a set $A \subseteq U$ by its characteristic vector $x_A \in \{0,1\}^U$, where $(x_A)_j = 1$ if and only if $j \in A$. We often leave out the subscript A and simply write x to represent the input. As the KOR sketch is linear and can be represented by a matrix, let S denote the sketch matrix and write Sx for the (non-private) KOR sketch for input x. We formally define the sketch matrix in Section 3.4.1. Let $x \circ w$ denote the Hadamard product. Our main theorem is:

Theorem 3.1 (Noisy KOR sketch)**.** *Let $w \in (0,1]^U$ be given. For every choice of $0 < \delta < 1$ and $\varepsilon = O(1)$ there exists a distribution H over GF(2)-linear sketches mapping a vector $x \in \{0,1\}^U$ to $\{0,1\}^\ell$, where $\ell = O\left(\log^2(u)\varepsilon^{-2}\delta^{-2}\right)$, and a distribution D_ε over noise vectors such that:*

1. *For $S \in H$ and $\eta \in D_\varepsilon$, given $Sx + \eta$ we can compute, in time $O(\ell)$, an estimate $\widehat{w}$ of $kx \circ wk_1$ that with probability $1 - 1/u$ satisfies $|\widehat{w} - kx \circ wk_1| < \delta kx \circ wk_1 + O\left(\log(u)\varepsilon^{-2}\delta^{-2}\right)$.*
2. *For every S in the support of H, $Sx + \eta$ is ε-differentially private over the choice of $\eta \in D_\varepsilon$, and can be computed in time $O(\ell kxk_0 \log(u) + \ell)$, including time for sampling η.*

The assumption that $\varepsilon = O(1)$ is not essential, and is only made to simplify our bounds (which do not improve for privacy parameter $\varepsilon = \omega(1)$). Without loss of generality, we can assume that parameter δ is

Reference	DP	Additive error	Rel. error	Initial. time	Space usage
Hardt & Talwar [82]	ε	$\tilde{\Omega}(n^{1/3}/\varepsilon)$	{	{	{
McGregor et al. [103]	ε	$\tilde{\Omega}(\sqrt{m}/e^{\varepsilon})$	{	{	{
Jayram & Woodru [89]	{	{	$1+\Omega(1)$	{	$\tilde{\Omega}(1/\delta^2)$
Kane et al. [94]	{	$O(1)$	$1+\Omega(1)$	$O(1)$	$O(1/\delta^2)$
Mir et al. [110]	ε	$O(n^{1/2-\Theta(1)}/\varepsilon^{\Theta(1)})$	$1+\Omega(1)$	$\exp((1/\varepsilon)^{\Theta(1)})$	$O((1/\varepsilon)^{\Theta(1)})$
Kenthapadi et al. [96]	(ε, δ)	$O(\sqrt{m}/\varepsilon)$	$1+\Omega(1)$	$\tilde{O}(u)$	$O(1/\delta^2)$
Stanojevic et al. [129]	ε	$O(\sqrt{\lvert A \cup B\rvert}/\varepsilon^2)$	{	$O(\lvert A\rvert + \lvert B\rvert)$	$O(\lvert A\rvert + \lvert B\rvert)$
This work	ε	$O(n^{2/3}/\varepsilon^{2/3})$	$1+\Omega(1)$	$O(1/\varepsilon^2\delta^2)$	$O\left(1/\varepsilon^2\delta^2\right)$

Figure 3.1: Selected lower bounds (top part) and upper bounds (bottom part) for estimating the (unweighted) size of the symmetric difference $m = |A \Delta B|$ from small sketches of sets $A, B \subseteq \{1, \ldots, u\}$. Bounds stated as $\tilde{\Omega}$ and $\tilde{O}$ are simplified by suppressing multiplicative factors polynomial in $\log(1/\varepsilon)$, $\log(1/\delta)$, $\log(1/\beta)$, and $\log u$. The non-private bounds in [89, 94] improve previous results by an $\tilde{O}(1)$ factor, we refer to their references for details. The space usage of [96] is measured in terms of real numbers; it is unclear how much space a private, discrete implementation would need.

such that the error is dominated by $\|x - w\|_1$ because reducing τ further cannot reduce error by more than a factor 2. In the unweighted case, setting $\tau = \sqrt[3]{\log(u)/(\varepsilon^2 m)}$ to balance relative and additive error we get error $O(m^{2/3}/\varepsilon^{2/3})$, where the O notations suppresses a polylogarithmic factor. This is polynomially related to known lower bounds described in section 3.2.3.

Applications Suppose that two players hold set A with corresponding characteristic vector $x_A \in \{0,1\}^u$ and B with characteristic vector $x_B \in \{0,1\}^u$. They jointly sample $S \in H$ and privately sample $\gamma_A, \gamma_B \in D_{\varepsilon}$ according to Theorem 3.1. Then $Sx_A + \gamma_A$ and $Sx_B + \gamma_B$ are ε-di erentially private. Furthermore, $(Sx_A + \gamma_B) + (Sx_B + \gamma_B) = (Sx_A + Sx_B) + (\gamma_A + \gamma_B)$, and we show in Section 3.4.3 that $\gamma_A + \gamma_B \in D_{\varepsilon'}$ with $\varepsilon' = \varepsilon^2/(2 + 2\varepsilon)$. In Section 3.5.2 we use this in conjunction with Theorem 3.1 to establish:

Corollary 3.1. *For accuracy parameter $\beta > 0$, consider an ε-di erentially private noisy KOR sketch for a set A and an ε-di erentially private noisy KOR sketch for a set B, based on the same linear sketch $S \in H$, sampled independently of A and B. We can compute an approximation $\hat{}$ of the weight of the symmetric di erence, such that with probability $1 - 1/u$:*

$$|\|x_{A \Delta B}\|_1 - \hat{}| < \|x_{A \Delta B}\|_1 + \text{poly}(1/\varepsilon, 1/\beta, \log u) .$$

In the special case where all weights w_j are 1, this reduces to estimating the *size* of the symmetric difference $A \Delta B$.

In Section 3.6 we describe how to modify our sketch to apply in a streaming setting. In this case, we estimate the size of the union of the input streams rather than the size of the symmetric di erence when merging two sketches.

3.2 Related Work

In the absence of privacy constraints, seminal estimators for (unweighted) set cardinality that support merging sketches (to produce a sketch of the union) are HyperLogLog [67], FM-sketches [68], and bottom-k (aka. k-minimum values) sketches [14]. Progress on making these estimators private for set operations include [135] (using FM-sketches) and [127], which builds a private cardinality estimator to estimate set intersection size using the bottom-k sketch. We note that these sketches do not achieve di erential privacy but are aimed at a weaker notion of privacy. Speci cally, they o er a one-sided guarantee that may reveal that

an individual element is *not* present in the dataset. To our best knowledge, a private version of HyperLogLog with provable bounds on accuracy has not been described in the literature.

The *weighted* version of cardinality estimation has been less studied. For (scaled) integer weights in $[W]$ there is a simple reduction that inserts element i with weight w_i by inserting the tuples $(i;1);\ldots;(i;w_i)$ into a standard cardinality estimator on the domain $U \quad [W]$, but this makes the obtained bounds depend on the number W of possible weights. Cohen et al. [39] showed that the class of cardinality estimators that rely on extreme order statistics (for example, HyperLogLog) can be e ciently extended to the weighted setting, even for real-numbered weights.

Note that the weighted F_0 estimation problem is di erent from F_1 and L_1 estimation in the context of set operations; for example, the union of two identical sets will have the same weighted F_0, whereas summing two identical vectors will produce a vector with twice the L_1 norm. In the rest of this section, we focus on the standard, unweighted setting.

3.2.1 Differentially private cardinality estimators

Already the seminal paper on pan-privacy [62] discusses di erentially private streaming algorithms for F_0 on insertion-only streams. Their sketch is not linear and does not allow deletions or subtraction of sketches. It is not clear if the sketch can be merged to produce a sketch for the union. Recent work by von Voigt et al. [138] has shown how to estimate the cardinality of a set using less space in a di erentially private manner using FM-sketches, using the Probabilistic Counting with Stochastic Averaging (PCSA) technique [68]. These sketches can be merged to obtain a sketch for the union of the input set with a slightly higher level of noise. Privacy is achieved by randomly adding ones to the sketch and only sketching a sample of the input dataset.

Bloom Filters have been studied extensively to obtain cardinality estimators under set operations (already implicit in [62]). Alaggan et al. [5] estimated set intersection size by combining a technique for computing similarity between sets, represented by Bloom lters in a di erentially private manner, named BLIP (BLoom-then- IP) lters [4] with a technique for approximating set intersection of two sets based on their Bloom Filter representation [25]. We note that [4] achieves privacy by ipping each bit of the Bloom lter with a certain probability, much like the technique we use to get privacy of our sketch. Stanojevic et al. [129] show how to estimate set intersection, union, and symmetric di erence for two sets by computing an estimate for the size of the union, and combined with the size of each set, they show how to compute an estimate for the size of the intersection and the symmetric di erence. They achieve privacy by ipping each bit with some probability, like in [4]. Also, RAPPOR [65] uses Bloom Filters with a Randomized Response technique to collect data from users in a di erentially private way but is mainly aimed at computing heavy hitters.

Though a bound on the expected worst-case error of privately estimating the size of a symmetric di erence $|A \triangle B|$ (as in Corollary 3.1) is not stated in any of these papers, an upper bound of $O(\sqrt{pm})$, where m is an upper bound on the size of the sets, follows from the discussion in [129] (for xed $"$). It seems that this magnitude of error is inherent to approaches using Bloom lters since it arises by balancing the error related to the noise and the error related to hash collisions in the Bloom lter. An advantage and special case of our noisy KOR sketch is that it can be used to estimate the size of the symmetric di erence directly, and therefore the error will depend only on the size of the symmetric di erence. It seems that with non-linear sketches, it would be necessary rst to estimate the size of the union and combine this with the size of each input set as exhibited in, for example, [129]. Hence, the error would depend on the size of the union of the input sets.

3.2.2 Differentially private sketches

Closely related to our work is the di erentially private Johnson-Lindenstrauss (JL) sketch by Kenthapadi et al. [96], in which the technique of adding noise to the sketch is also applied. Kenthapadi et al. add Gaussian noise, so to store and maintain a sketched vector, some kind of discretization would be needed (not discussed in their paper). Discretizing a real-valued private mechanism is non-trivial: Without su cient care, one might lose privacy due to rounding in an implementation, as argued by Mironov [111]. Even if a suitable

discretization of the mechanism in [96] would be possible (see [27] for a general discussion), it has several drawbacks compared to our method:

- It only achieves *approximate* di erential privacy as opposed to the pure di erential privacy of the noisy KOR sketch.
- The time needed to update the sketch when a set element is inserted or removed is not constant (in the main method described, it is linear in the sketch size).
- The time needed to initialize the sketch is linear in the size of the sketch matrix, which has u columns because the noise needs to be calibrated to the sensitivity of the JL sketch matrix, which requires linear time in the size of the sketch matrix. Alternatively, which is the suggestion in Kenthapadi et al., the sketch matrix is assumed to have low sensitivity, and noise is calibrated to this sensitivity. If a sketch matrix with a large entry is randomly chosen, the sensitivity of the sketch matrix is large, in which case the noise does not ensure privacy. So with a small probability, privacy is not preserved.

Another closely related work is the paper of Mir et al. [110], which also adds a noise vector after computing standard linear sketches for F_0 estimation to make the sketch di erentially private. They further initialize their sketches with random noise vectors to also get pan-privacy. The error bound obtained is similar to ours, and the sketch has a discrete representation, but their method is inferior in terms of time complexity. This is because they rely on the *exponential mechanism* [105], which is not computationally e cient. (Note that a preprint of the paper of Mir et al. [109] presented a computationally more e cient method. However, the sensitivity analysis in that paper has an error [117] that was corrected in the slower method published in [110].)

Our method is more computationally e cient and arguably simpler than the methods of [96, 110]. Our linear sketch is not a replacement for these sketches, though, since our sketch is over GF(2) rather than the reals (or integers).

3.2.3 Lower bounds.

Jayram and Woodru [89] show that, even with no privacy guarantee, to obtain error probability $1=u$ we need a sketch of $(\log(u)^2)$ bits to estimate F_0 with relative error 1 . It is easy to extend this lower bound to our setting, in which an additive error of c is allowed: Simply insert each item c times to increase the size of the set so that the additive error is negligible. Formally this requires us to extend the universe to $U \times \{1,\ldots,c\}$, such that the lower bound in terms of the original universe size becomes $(\log(u/c)^2)$. (We do not use this reduction to eliminate the additive error in our upper bound because the reduction increases the sensitivity of updates, destroying the di erential privacy properties.)

Hardt and Talwar [82] show that an ε-di erentially private sketch for F_0 must have additive error $(1/\varepsilon)$, which is comparable (up to polynomial and logarithmic factors) to the additive error we achieve.

Desfontaines et al. [45] show that it is not possible to preserve privacy in accurate cardinality estimators if we can merge several sketches without loss in accuracy. Our sketch will have an increase in noise when merging sketches and thus does not satisfy the requirement for cardinality estimators formulated in [45].

McGregor et al. [103] showed that in order to estimate the size of the intersection of two sets A and B, based on di erentially private sketches of A and B, an additive error of $(\sqrt{u}/e^{\varepsilon})$ is needed in the worst case when A and B are arbitrary subsets of $[u]$. The lower bound holds even in an interactive setting where the players (holding A and B, resp.) can communicate, and we require that the communication transcript is di erentially private. The hard input distribution uses sets with symmetric di erence of size (u) with high probability. Since $|A \cap B| = (|A| + |B| - |A \triangle B|)/2$, estimating the intersection size is no more di cult (up to constant factors in error) than estimating $|A|$, $|B|$, and $|A \triangle B|$. We can estimate $|A|$ and $|B|$ with error $O(1/\varepsilon)$ under di erential privacy, so it follows that estimating $|A \triangle B|$ under di erential privacy requires error $(\sqrt{u}/e^{\varepsilon})$. For a contrasting upper bound, [137, 113] suggest an algorithm estimating two-party set intersection size up to an additive error of $O(\sqrt{u}/\varepsilon)$ with high probability. A lower bound in terms of the size m of the symmetric di erence follows by setting $u = m$.

3.2.4 Noisy sketching.

In addition to the paper of Mir et al. [110], there is some previous work on sketching techniques in the presence of noise. Motivated by applications in learning theory, Awasthi et al. [9] considered recovery of a vector based on noisy 1-bit linear measurements. The resistance to noise demonstrated is analogous to what we show for the KOR sketch but technically quite di erent since the linear mapping is computed over the reals before a sign operation is applied.

In a very recent paper [34], Choi et al. propose a framework for releasing di erentially private estimates of various sketching problems in a distributed setting. This framework ensures that the estimates only have a multiplicative error factor. The technique relies on secure multiparty computation, and the sketch submitted by each participant is not private and so cannot be released. Further, the results of Choi et al. do not immediately allow for estimating size or weight of the symmetric di erence between two sets.

If the sketching matrix S itself is secret and randomly chosen from a distribution over matrices with entries in a nite eld, very strong privacy guarantees on the sketch Sx can be obtained while still allowing kxk_0 to be estimated from Sx with small error [18]. Blocki et al. [20] prove that the Johnson-Lindenstrauss transform is, in fact di erentially private, when keeping the sketch matrix secret. However, the condition that the sketch matrix is secret is a serious limitation for applications such as streaming and distributed cardinality estimation that require S to be stored or shared.

3.3 Preliminaries

For a set $A \subseteq U$, we let x_A denote the characteristic vector for A, de ned for $j \in [u]$ as

$$(x_A)_j = \begin{cases} 1; & j \in A \\ 0; & \text{otherwise} : \end{cases}$$

We write w_A (or w_{x_A}) for the weight vector for input set A such that

$$w_A = x_A \quad w$$

for xed, public weights $w_j \in (0; 1]$, and denotes the Hadamard product.

For vector $x = (x_1; :::; x_u)$ we de ne $kxk_p = \left(\sum_{j=1}^{u} jx_j j^p\right)^{1=p}$ as the p-norm of x. For $p = 0$, we de ne $kxk_0 = \sum_{j=1}^{u} \mathbf{1}[x_j \neq 0]$, often called the zero-"norm". F_0 denotes the 0^{th} frequency moment and represents the number of distinct elements in a stream (or a set). Frequency moments are well-known from the streaming literature; see for example [7].

Our sketch Sx_A is comprised of $\log(u)$ "levels", $S_i x_A$ for $i = 0; :::; \log(u) \quad 1$. We refer to Section 3.4.1 for a description of these levels. Let n denote the size of the binary vector representation of $S_i x_A$ for each i. Hence, the size of the noisy KOR sketch $Sx_A + '$ is $= n \log u$. Note that n is xed and depends on the privacy parameter $"$ and the accuracy parameter .

Finally, we assume that sets and vectors are stored in a sparse representation, such that we can list the non-zero entries in the input vector x in time $O(kxk_0)$.

3.3.1 Hashing-based subsampling

The sketch matrix S is de ned by several hash functions. For simplicity, we assume access to an oracle representing random hash functions, namely, that we can sample a fully random hash function, and it can be evaluated in constant time. We do not store the hash function as part of our sketch, so the space for our sketch does not include space required for storing the hash function. We believe it is possible to replace these hash functions with concrete, e cient hash functions that can be stored in small space while preserving the asymptotic bounds on accuracy, but to focus on privacy aspects, we have not pursued this direction. Importantly, the di erential privacy of our method holds for any choice of hash function and does not depend on the random oracle assumption.

To ensure that adding two sketches gives a sketch for the symmetric difference, both players must sample the same elements for each S_i. To ensure coordinated sampling, we use a hash function, so the same elements from U are sampled by both players. We use the following (standard) subsampling technique: let S be the family of all fully random hash functions from U into $[0;1]$. Let $s \in S$ uniformly at random. We sample an element j from the input set at level $i = 0;:::;\log(u) - 1$ if and only if $s(j) \in \left(w_j/2^{i+1}; w_j/2^i\right]$. We refer the reader to the survey of Woodruff [144] for more details on subsampling.

3.3.2 The Differential Privacy Model

In section 3.5.1 we prove that our protocol obtains ε-differential privacy. As the inputs are binary vectors and the sketch is over GF(2), neighboring inputs are input vectors that differ in exactly one entry, i.e., vectors with Hamming distance 1. The sketch is over GF(2) and has sensitivity 1, which motivates the use of Randomized Response as privatizing technique. We refer to Section 2.1.2 for the details about differential privacy.

We may think of the protocol for estimating the weight of symmetric difference as working in the *local* model of differential privacy (Section 2.1.5), as each player adds noise to their own sketch. We note that our sketch would *also* work in a model where vectors supplied by the users are combined using a black-box *secure multiparty aggregation* [77, 108]. In this setting, only the sketch for the symmetric difference would be released, and so only *this* sketch would need to be differentially private.

In Section 3.4.3 we discuss how to compute estimates for the union and intersection of two input sets by combining estimates for the weights of each set together with an estimate for the weight of the symmetric difference. By each party releasing the weight of their own set via the Laplace mechanism (Lemma 2.3), we may compute such differentially private estimates of the union and intersection with error of the same magnitude as for the symmetric difference.

3.4 Techniques

3.4.1 Sketch Description

In this section, we describe the noisy KOR sketch in detail. The description is self-contained, but we refer the interested reader to [40] for more background on (linear) sketches. As mentioned, our sketch combines the techniques from [99] with hashing-based subsampling to achieve a sketch that is robust against adding noise, as long as we know how much noise was added.

Recall that for input vector $x \in \{0;1\}$ and public weight vector $w \in (0;1]^u$ we simplify notation by defining $w_x := x \cdot w$. The goal is to estimate *the weight of* x, $\|w_x\|_1$.

We first give the intuition behind the $n \times u$-matrices S_i, that our sketch S is comprised of: Suppose that we have a rough estimate $\hat{E}$ of $\|w_x\|_1$, accurate within a constant factor. Then we can obtain a more precise estimate by sampling (using a hash function) a fraction $n/\hat{E}$ of the elements, for some parameter n, and computing the sketch from [99] of size n for the sampled elements. This gives an approximation of the number of sampled elements, which in turn gives an approximation of $\|w_x\|_1$ with small relative error. Since we do not know $\|w_x\|_1$ within a constant factor { especially in the setting where we are interested in the size of the symmetric difference { we use hashing-based subsampling to sample each element j from the input set with probability $w_j/2^{i+1}$ for $i = 0;:::;\log(u) - 1$. Thus for each i, we sample elements corresponding to approximately a $1/2^{i+1}$ fraction of the weight and compute the sketch from [99] of size n for the sampled elements. For one of these i we are guaranteed to sample approximately a fraction $n/\|w_x\|_1$ of the input weight assuming that $\|w_x\|_1 > n$. For this i, we can obtain a precise estimate of $\|w_x\|_1$ from the sketch.

We now define S_i formally. We first describe the sketch from [99] as a linear sketch over GF(2). Let F be the family of all hash functions from universe U into $[n]$, and pick $h \in F$ uniformly at random. The hash function h uniquely defines an $n \times u$-matrix K, where

$$K_{k;j} = \begin{cases} 1; & \text{if } h(j) = k \\ 0; & \text{otherwise :} \end{cases}$$

We combine this with the following sampling technique:

Let S be the family of all hash functions from U to $[0;1]$. Sample $s \in S$ uniformly at random. The hash function s defines a $u \times u$-diagonal matrix M_i for each $i = 0;:::;\log(u) - 1$, defined by

$$(M_i)_{j;j} = \begin{cases} 1; & \text{if } s(j) \in (w_j=2^{i+1}; w_j=2^i] \\ 0; & \text{otherwise.} \end{cases}$$

The matrix-vector product $M_i x$ represents subsample of input vector x, where we sample each element with probability $w_j=2^{i+1}$.

We are finally ready to define S_i as $S_i = K M_i$, which is an $n \times u$-matrix over GF(2). By definition:

$$(S_i)_{k;j} = \begin{cases} 1; & (h(j) = k) \wedge (s(j) \in (w_j=2^{i+1}; w_j=2^i]) \\ 0; & \text{otherwise.} \end{cases}$$

The KOR sketch can be represented as an $n \log(u) \times u$-matrix S, formed by stacking $S_1;:::;S_{\log(u)}$.

Let D_η be a distribution over vectors from $\{0;1\}^{n \log(u)}$, where each entry is 1 independently with probability p. We show in Section 3.5.2 that it suffices to set $p = 1=(2 + \varepsilon)$. Sample the noise (or *pertubation*) vector $\tau \in D_\eta$ independently and uniformly at random. The *noisy* KOR sketch of x is then computed (over GF(2)) as:

$$Sx + \tau:$$

3.4.2 Estimation

Next, we describe how to compute a weight estimate from a sketch $Sx + \tau$. Let $w_x := x \cdot w$ and let τ_i be the restriction of τ to the entries that are added to $S_i x$ when adding τ to Sx. To compute an estimate for $\|w_x\|_1$, for each $i = 0;:::;\log(u) - 1$ count the number of 1s in $S_i x + \tau_i$, $Z_i = \|S_i x + \tau_i\|_0$ and compute the interval:

$$I_i = \begin{cases} [0; u] & \text{if } Z_i \geq (1 - \delta)n=2 \\ \left[2^i n \ln\left(\frac{\frac{1}{2}-\eta+1}{1-\frac{2Z_i}{(1+\delta)n}}\right); 2^i n \ln\left(\frac{\frac{1}{2}-\eta+1}{1-\frac{2Z_i}{(1-\delta)n}}\right)\right] & \text{otherwise.} \end{cases} \tag{3.1}$$

where $\delta < \frac{1=n}{7e^3(2=\varepsilon+1)}$: Compute the intersection $I = \bigcap_{i=0}^{\log(u)-1} I_i$ and check if the maximum value in I is within a factor $(1 + \gamma)$ of the minimum value in I for

$$\gamma = \frac{6\delta\left(e^3\left(\frac{2}{\varepsilon}+1\right) - 1\right)}{1 + \delta - 2\delta\left(e^3\left(\frac{2}{\varepsilon}+1\right)\right)}:$$

If that is the case, every element in I is a good estimate for $\|w_x\|_1$ (having relative error at most $(1 + \gamma)$) with high probability. Otherwise, $\|w_x\|_1$ is small with high probability, and we let the estimate for $\|w_x\|_1$ be 0. We analyze the accuracy of this estimator in Section 3.5.

3.4.3 Application to symmetric difference

In this section, we describe a differentially private protocol to compute an estimate for the weight of the symmetric difference between sets held by two parties. First, we show that the sum of two noisy KOR sketches, $Sx_A + \tau$ and $Sx_B + \tau'$, is a noisy KOR sketch for the symmetric difference, $S(x_{A \Delta B}) + (\tau + \tau')$, which has the same properties as $Sx_A + \tau$ and $Sx_B + \tau'$, but for $\varepsilon' < \varepsilon$ as more noise is added.

Lemma 3.1. *The addition (over GF(2)) of two noisy KOR sketches with perturbation vectors $\tau \in D_\eta$ and $\tau' \in D_\eta$, respectively, is a noisy KOR sketch for the symmetric difference of the input sets with noise $\tau + \tau' \in D_{\eta'}$ for $\eta' = \eta^2=(2 + 2\eta)$.*

Proof. Let x_A and x_B be the input vectors from each of the two players. Let S be as defined in Section 3.4.1, and define τ'; as the noise vectors for the noisy KOR sketches for x_A and x_B, respectively. We have (over GF(2)) that

$$(Sx_A + \tau') + (Sx_B +) = (Sx_A + Sx_B) + (\tau' +) = S(x_A + x_B) + (\tau' +):$$

This is exactly the noisy KOR sketch for the symmetric difference with perturbation $\tau' +$. Note that we observe a 1 in an entry of $\tau' +$ with probability $p' = p(1 - p) + (1 - p)p = 2p(1 - p)$. We show in Section 3.5.2 that we can let $p = \frac{1}{2+\varepsilon}$. Observe that

$$p' = \frac{1}{2+\varepsilon'} = \frac{2}{2+\varepsilon}\left(1 - \frac{1}{2+\varepsilon}\right)$$

which implies that $\varepsilon' = \varepsilon^2/(2 + 2\varepsilon)$. □

By Lemma 3.1 we can treat a sketch for the symmetric difference exactly like a sketch for input vector x although with a different privacy parameter ε'. Hence, Theorem 3.1 gives us Corollary 3.1, restated here for convenience:

Corollary 3.1. *For accuracy parameter > 0, consider an ε-differentially private noisy KOR sketch for a set A and an ε-differentially private noisy KOR sketch for a set B, based on the same linear sketch $S \in H$, sampled independently of A and B. We can compute an approximation $\hat{}$ of the weight of the symmetric difference, such that with probability $1 - 1/u$:*

$$|\|x_A - x_B\|_1 - \hat{}| < \|x_A - x_B\|_1 + \mathrm{poly}(1/\varepsilon, 1/ , \log u) :$$

Note that the additive error in Corollary 3.1 still depends polynomially on ε even for privacy parameter ε', which is explained by the fact that $\varepsilon' = \varepsilon^2/(2 + 2\varepsilon)$.

Finally, we assumed that $\|wx_A\|_1$ and $\|wx_B\|_1$ were released with Laplace noise, which gives an expected additive error of $O(1/\varepsilon)$ for each of $\|wx_A\|_1$ and $\|wx_B\|_1$ [60]. We can use the following equations to get estimates for the union, intersection and difference:

$$\|wx_{A \cup B}\|_1 = \frac{\|wx_A\|_1 + \|wx_B\|_1 + \|wx_{A \triangle B}\|_1}{2},$$
$$\|wx_{A \cap B}\|_1 = \frac{\|wx_A\|_1 + \|wx_B\|_1 - \|wx_{A \triangle B}\|_1}{2}$$
$$\|wx_{A \setminus B}\|_1 = \frac{\|wx_A\|_1 + \|wx_{A \triangle B}\|_1 - \|wx_B\|_1}{2} :$$

That is, the error is bounded by half the error of the estimate of the symmetric difference size plus $O(1/\varepsilon)$.

3.5 Proof of Theorem 3.1

In this section, we give a proof of Theorem 3.1, restated here for convenience:

Theorem 3.1 (Noisy KOR sketch)**.** *Let $w \in (0,1]^u$ be given. For every choice of $0 < < 1$ and $\varepsilon = O(1)$ there exists a distribution H over GF(2)-linear sketches mapping a vector $x \in \{0,1\}^u$ to $\{0,1\}^{\ }$, where $= O\left(\log^2(u)/\varepsilon^{\ 2}{}^{\ 2}\right)$, and a distribution D_ε over noise vectors such that:*

1. *For $S \in H$ and $\tau' \in D_\varepsilon$, given $Sx + \tau'$ we can compute, in time $O(\ )$, an estimate $\hat{w}$ of $\|x - w\|_1$ that with probability $1 - 1/u$ satisfies $|\hat{w} - \|x - w\|_1| < \|x - w\|_1 + O\left(\log(u)/\varepsilon^{\ 2}{}^{\ 2}\right)$.*

2. *For every S in the support of H, $Sx + \tau'$ is ε-differentially private over the choice of $\tau' \in D_\varepsilon$, and can be computed in time $O(\|x\|_0 \log(u) +)$, including time for sampling τ'.*

3.5.1 Noise level and Differential Privacy Guarantees

We first show that the noisy KOR sketch $Sx + \eta$ satisfies ε-differential privacy, which proves part 2 of Theorem 3.1. Intuitively, removal/insertion of a single element can change only a single entry in the sketch, as the element is inserted into only a single level.

Lemma 3.2. *If $p \geq \left(\frac{1}{e^{\varepsilon}+1}, \frac{1}{2}\right)$ then $Sx + \eta$ is ε-differentially private.*

Proof. Let A and B be neighboring input sets with corresponding characteristic vectors, x_A and x_B. *Neighboring* here means that one set is a subset of the other and the sizes differ by 1. By symmetry of differential privacy, we can without loss of generality assume that A is the smaller set. Suppose that $B \setminus A = \{z\}$. The element z can only affect $S_i x$ for i where z is sampled. Recall that there is at most one such i. If z is never sampled, then $Sx_A = Sx_B$ and privacy is trivial. So assume $i \in \{0, \ldots, \log(u) - 1\}$ such that $s(z) \in \left(w_z/2^{i+1}, w_z/2^i\right]$. We limit our attention to $S_i x_A + \eta_i$, where we think of η_i as the restriction of the $n\log(u)$-dimensional random vector $\eta \sim D_\varepsilon$ to the entries that would be added to $S_i x_A$ when adding η to Sx_A. We show that $S_i x_A + \eta_i$ is ε-differentially private. Then the entire sketch, $Sx_A + \eta$, is ε-differentially private.

Inserting z into the sketch implies that $S_i x_A$ and $S_i x_B$ will differ in exactly one entry. That is, $\|S_i x_A - S_i x_B\|_0 = 1$. Fix a noisy sketch, H_i. There exist unique vectors η'_i and η_i, such that $H_i = S_i x_A + \eta'_i = S_i x_B + \eta_i$. Note that $\|\eta'_i - \eta_i\|_0 = 1$. Let $\|\eta'_i\|_0 = r$. Then $\|\eta_i\|_0 = r'$ for $r' \in \{r+1, r-1\}$. Conditioned on $\|\eta'_i\|_0 = r$ and $\|\eta_i\|_0 = r'$, the probabilities of randomly drawing exactly these randomness vectors are, respectively:

$$(1-p)^{n-r} p^r \qquad \text{and} \qquad (1-p)^{n-r'} p^{r'}.$$

Let $\varepsilon = O(1)$ be given. By Section 2.1.2 we know that it suffices to show that for any fixed output $H_i = S_i x_A + \eta'_i = S_i x_B + \eta_i$, we have

$$e^{-\varepsilon} \leq \frac{\Pr\left[\text{observe } H_i \text{ from } A\right]}{\Pr\left[\text{observe } H_i \text{ from } B\right]} = \frac{\Pr\left[\text{observe } S_i x_A + \eta'_i \text{ from } A\right]}{\Pr\left[\text{observe } S_i x_B + \eta_i \text{ from } B\right]} \leq e^{\varepsilon},$$

where the probability is over the randomness in η'_i and η_i.

Hence, to obtain differential privacy it suffices that for every possible value of r and $r' \in \{r+1, r-1\}$

$$e^{-\varepsilon} \leq \frac{(1-p)^{n-r} p^r}{(1-p)^{n-r'} p^{r'}} = \frac{1}{(1-p)^{r-r'} p^{r'-r}} \leq e^{\varepsilon},$$

which is satisfied for $1/2 > p \geq 1/(e^{\varepsilon} + 1)$, since $p < 1/2$ by assumption. □

3.5.2 Bounding accuracy

In this section, along with Section 3.5.3, we prove the first part of Theorem 3.1. Let input vector x be given and let $w_x = \|x\|_1 - w$. We will mainly consider each $S_i x$ isolated, so let the (binary) randomness vector η'_i be the n-dimensional restriction of η as described in the proof of Lemma 3.2. First, we state two useful lemmas.

Lemma 3.3. *For each $i = 0, \ldots, \log(u) - 1$ let $L_i = \|S_i x\|_0$ and $Z_i = \|S_i x + \eta'_i\|_0$. Then:*

$$\mathop{\mathbb{E}}_{h,S}[L_i] = \frac{n}{2}\left(1 - \prod_{j:x_j=1}\left(1 - \frac{w_j}{2^i n}\right)\right) \tag{3.2}$$

$$\mathop{\mathbb{E}}_{h,S,\eta'_i \sim D_p}[Z_i] = \frac{n}{2}\left(1 - (1-2p)\prod_{j:x_j=1}\left(1 - \frac{w_j}{2^i n}\right)\right) \tag{3.3}$$

Proof. We refer to Section 3.8 for the proof. □

Lemma 3.4. *For $i = 0, \ldots, \log(u) - 1$ let $Z_i = kS_i x + \ '_i k_0$. For any $0 < \ < 1$, we have with probability at least $1 - 6\log(u)e^{-\frac{2\ ^2 p^3 n}{6^2\ ^3}}$ that for all $i = 0, \ldots, \log(u) - 1$ simultaneously:*

$$(1 - \ ) \mathop{\mathsf{E}}_{\substack{h\ F;\\ s\ S;\\ '_i\ D_p}}[Z_i] < Z_i < (1 + \ ) \mathop{\mathsf{E}}_{\substack{h\ F;\\ s\ S;\\ '_i\ D_p}}[Z_i].$$

Proof. We refer to Section 3.8 for the proof. □

First, we consider the case when $1 < n < kw_x k_1$. In Lemma 3.5 we state that in this case, with high probability, we get an error of at most a factor $(1 + \ )$ for a well-chosen $\ $, where $\ $ is a function of the privacy parameter $"$, the accuracy parameter $\ $ and the size of the universe, u. For convenience, de ne

$$I_i(p) = \begin{cases} [0; u] & \text{if } Z_i \quad (1 - \ )n/2 \\ \left[2^i n \ln\left(\frac{1 - \frac{2p}{2Z_i}}{1 - \frac{2Z_i}{(1+\ )n}}\right); 2^i n \ln\left(\frac{1 - \frac{2p}{2Z_i}}{1 - \frac{2Z_i}{(1-\ )n}}\right)\right] & \text{otherwise} \end{cases} \tag{3.4}$$

and $\hat{w}_x := 2^i n \ln\left(1/\prod_{j:x_j=1}\left(1 - \frac{w_j}{2^i n}\right)\right)$. We prove our result in two steps:

1. If $\hat{w}_x \ 2\ I_i(p)$ for all $i = 0, \ldots, \log(u) - 1$, then there is some i such that any value from (3.4) estimates $\hat{w}_x$ up to a factor $(1 + \ )$, where $\ $ is a function of $\ $ and $"$.
2. $kw_x k_1 - \hat{w}_x \ \left(1 + \frac{1}{2^i n}\right) kw_x k_1$ for each i. Speci cally, $kw_x k_1 - \hat{w}_x \ \left(1 + \frac{1}{n}\right) kw_x k_1$ for all i.

Hence, we choose $\ $ independent of i such that $(1 + \ )\left(1 + \frac{1}{n}\right) \quad (1 + \ )$ for at least one of the intervals $I_i(p)$. We pick $\ $ to work for the i where $kw_x k_1/(2^i n) \ 2\ [1; 2)$ as this corresponds to having an input of size between n and $2n$ (we obtain this input size by the sampling from x in S_i). If $kw_x k_1 \quad n$, there is such an i, and we can identify it by checking that the endpoints of the interval are su ciently close together, as described in Section 3.4.2. We consider the case when $kw_x k_1 < n$ in Section 3.5.3 where we show that in this case, the error is bounded by an additive factor of $O(n)$.

Lemma 3.5. *Assume $kw_x k_1 > n > 1$, and $\ > \frac{1}{n}$. With probability at least $1 - 6\log(u)e^{-\frac{2\ ^2 p^3 n}{108}}$ there exists an $i \ 2\ f0, \ldots, \log(u) - 1g$ such that any element from $I_i(p)$ is a $(1 + \ )$-approximation to $kw_x k_1$ for*

$$< \frac{\left(\ - \frac{1}{n}\right)(1 - \ - 2p)}{7e^3}:$$

Speci cally, i where $\frac{kw_x k_1}{2^i n} \ 2\ [1; 2)$, gives these guarantees.

Proof Sketch. We give an informal sketch of the proof and refer to Section 3.8 for the formal proof. We rst remark that for $\ $ as described, Lemma 3.4 implies that if $kw_x k_1/(2^i n) \quad 2$, then $Z_i < (1 - \ )n/2$ with high probability. Hence, it su ces to consider the intervals from (3.4) of the form

$$I_i(p) = \left[2^i n \ln\left(\frac{1 - \frac{2p}{2Z_i}}{1 - \frac{2Z_i}{(1+\ )n}}\right); 2^i n \ln\left(\frac{1 - \frac{2p}{2Z_i}}{1 - \frac{2Z_i}{(1-\ )n}}\right)\right]:$$

De ne

$$\hat{w}_x := 2^i n \ln\left(\frac{1}{\prod_{j:x_j=1}\left(1 - \frac{w_j}{2^i n}\right)}\right):$$

From Lemma 3.3, we have

$$\prod_{j:x_j=1}\left(1 - \frac{w_j}{2^i n}\right) = \frac{1 - \frac{2\,\mathsf{E}[Z_i]}{n}}{1 - 2p}:$$

Assume that the bounds in Lemma 3.4 are satis ed. We remove this assumption shortly. By the bounds in Lemma 3.4, $\hat{w}_x \in I_i(p)$ for all i. We show that $\hat{w}_x$ is contained in an interval, which is slightly bigger than $I_i(p)$ whenever $kw_xk_1/(2^i n) \in [1;2)$ and show that the endpoints of this interval are within a factor $(1+\ )$ of each other, where is a function of . Clearly, then $I_i(p)$ is also su ciently small for this i. Denote this interval $I_i(p)$. Any element from $I_i(p)$ is a $(1+\ )$-approximation to $\hat{w}_x$. Removing the assumption that the bounds in Lemma 3.4 hold, we simply get a small error probability and conclude that with probability at least $1 - 6\log(u)e^{-2p^3 n/108}$ we have $\hat{w}_x \in I_i(p)$ for all i, and thus any value from $I_i(p)$ is a $(1+\ )$ estimation to $\hat{w}_x$ with high probability. Observing that $kw_xk_1 < \hat{w}_x < \left(1+\frac{1}{n}\right) kw_xk_1$ for any i, we choose in terms of such that $(1+\ )\left(1+\frac{1}{n}\right) < (1+\ )$. Then any value from $I_i(p)$ is a $(1+\ )$-approximation for kw_xk_1. We formally choose when giving the technical details in Section 3.8. We remark that the assumption $kw_xk_1/(2^i n) \in [1;2)$ allows us to choose independent of i, such that we can compute $I_i(p)$ for all i with a single value of . □

Observing that $\frac{1}{2+\varepsilon} > \frac{1}{e^{\varepsilon}+1}$ for $\varepsilon > 0$, we let $p = 1/(2+\varepsilon)$ and observe that for $I_i := I_i(1/(2+\varepsilon))$ with the choice of described in Lemma 3.5, we get the interval I_i in (3.1).

3.5.3 Putting things together

In this section, we consider the accuracy in the remaining case where $kw_xk_1 \quad n$. We also analyze the running time. Combining with Section 3.5.1, this completes the proof of Theorem 3.1.

Note that if $\varepsilon > 1$, we can start our protocol by dividing ε by a suitable constant, c such that $\varepsilon' = \varepsilon/c < 1$. Changing ε by a constant will change our bounds by a constant factor as well. Hence, we can, without loss of generality, assume $\varepsilon < 1$. We can also, without loss of generality, assume $u > 10$, as this will at most increase the failure probability and space by a constant factor.

We rst show a su cient upper bound on the sketch size $= n\log u$. Observe that $p > 1/4$ and let $c = 7e^3$ be a constant. Then we want $e^{-2p^3 n/108} < 1/u^2$ as this ensures a failure probability of at most $6\log(u)/u^2 < 1/u$: Noting that

$$(1 \quad 2p)^2 = \left(1 \quad \frac{2}{2+\varepsilon}\right)^2 = \left(\frac{1}{2/\varepsilon+1}\right)^2 = \frac{1}{4/\varepsilon^2 + 4/\varepsilon + 1} > \frac{\varepsilon^2}{20};$$

we have

$$e^{-\frac{2p^3 n}{108}} < e^{-\frac{\left(\frac{1}{n}\right)(1 \quad 2p)^2 n/4^3}{7e^3}\cdot\frac{1}{108}} = e^{-\frac{\left(\frac{1}{n}\right)^2 \frac{1}{(2/\varepsilon+1)^2} n}{4^3 c^2 \ 108}} < e^{-\frac{\left(\frac{1}{n}\right)^2 \varepsilon^2 n}{20\ 4^3 c^2\ 108}} < 1/u^2$$

when letting $n = O\left(\log(u)^{\ 2}\varepsilon^{\ 2}\right)$. Hence, the size of the sketch is

$$= \log(u) \quad n = O\left(\frac{\log^2(u)}{\varepsilon^2\ 2}\right):$$

Note that this n satis es the requirement $> 1/n$ from Lemma 3.5.

We argue about the error: Note that if $kw_xk_1 \quad n$, then if one of the intervals I_i is su ciently small and $\hat{w}_x \in I_i$ for all $i = 0;:::;\log(u) \quad 1$, then $\hat{w}_x \in I = \bigcap_{i=0}^{\log(u)\ 1} I_i$ and I is also su ciently small to give the wanted estimate. So by Lemma 3.5, we can check if the endpoints of I are within a factor at most $(1+\ )$ of each other, and if so, with probability $1 \quad 1/u$ any value from I is within a factor $(1+\ )$ of kw_xk_1. If I is too big, then none of the intervals I_i was su ciently small implying that our assumption that $kw_xk_1/(2^i n) \in [1;2)$ does not hold for any i. And so, with probability $1 \quad 1/u$ we have $kw_xk_1 < n$. We refer to Section 3.8 for the formal proof of Lemma 3.5. Our protocol sets the estimate of kw_xk_1 to 0 leading to an additive error of $O(n)$ when I was too big. This means that we get an additive error of at most $n = O\left(\log(u)^{\ 2}\varepsilon^{\ 2}\right)$, as required.

Finally, we comment on the running times: For the first part of Theorem 3.1, we note that in order to compute the estimate, we need to count the number of ones in $S_i x + \tau'_i$ for each $i = 0, \ldots, \log(u) - 1$, compute the intervals I_i and their intersection and check if it is sufficiently small. Counting the number of ones in all $S_i x + \tau'_i$ is the bottleneck and requires time $O(\quad)$. For the second part of Theorem 3.1, note that we can initialize the randomness vector τ' in time $O(\quad)$ and we can hash vector x in time $O(\|x\|_0 \log(u))$ assuming that we can iterate over x in time $O(\|x\|_0)$. Combining with Lemma 3.5 and Lemma 3.2, we have completed the proof of Theorem 3.1.

3.6 Distributed Streaming Implementation

In a streaming setting, we want a sketch that can be updated, and two sketches can be merged to give a sketch for the union of the input streams, while we cannot guarantee that there are no duplicates in the input stream. Our sketch does not immediately apply in this case, as items with an even number of occurrences would "cancel out". Therefore, such items would never be represented in the sketch, as the sketch is over GF(2). This issue can easily be fixed: the idea is to add another layer of sampling, such that we sample each *occurrence* of a data item with probability 1/2. Hence, we treat identical items independently on each occurrence and so ensure that an entry in the sketch is 1 with probability 1/2, regardless of the number of copies of identical items and collisions with other items. We refer to this as the *pre-sampled* sketch. The intuition is that the number of copies of an item inserted in the pre-sampled sketch is even or odd with probability 1/2. By Chernoff bounds, the fraction of elements that are sampled an odd number of times is very close to 1/2 with high probability. Thus it is natural to consider the estimator that is two times the estimator described in Section 3.4.2.

To understand this in more detail, we argue that merging two (non-private) pre-sampled sketches over GF(2) gives a sketch for the *union* of the two input sets. Suppose $z \in A \cup B$, $h(z) = k$ and that z is sampled at level i. We argue that $\Pr[(S_i x_{A \cup B})_k = 1] = 1/2$. Note that

$$(S_i x_{A \cup B})_k = 1 \quad , \quad (S_i x_A)_k \neq (S_i x_B)_k.$$

Further, we have that if $z \in A$, then $\Pr[(S_i x_A)_k = 1] = 1/2$ regardless of the number of other elements hashing to k at level i. If no elements from A hash to entry k at level i, then $\Pr[(S_i x_A)_k = 1] = 0$. We have

$$\Pr[(S_i x_{A \cup B})_k = 1] = \Pr[(S_i x_A)_k = 1]\Pr[(S_i x_B)_k = 0] + \Pr[(S_i x_A)_k = 0]\Pr[(S_i x_B)_k = 1],$$

which is 1/2 whenever $z \in A \cup B$.

3.7 Open Problems

An immediate question is, how to get a better additive error than described in Figure 3.1 in terms of the size of the symmetric difference m and the accuracy and privacy parameters and ε. Whereas the noisy KOR sketch has additive error $O(m^{2/3}/\varepsilon^{2/3})$, one may ask whether we can improve on the exponent of 2/3 to get closer to the well-known lower bound [103] of $\tilde{\Omega}(\sqrt{m}/\varepsilon)$.

In Section 3.4.3 we saw that combining sketches $(Sx + \tau') - (Sy + \quad) = S(x - y) + (\tau' - \quad)$ lead to a sketch with noise $\tau' - \quad$, which need not have the same distribution as τ' and . Hence, the combined sketch $S(x - y) + (\tau' - \quad)$ may have a higher noise level (and so better privacy guarantees and worse accuracy guarantees) than the original sketches $Sx + \tau'$ and $Sy +$. Therefore it is interesting to understand how noise behaves under such addition/subtraction, and in particular, how fast the amount of noise grows when combining multiple sketches. Suppose that we combine n linear sketches: $\sum_{j=1}^{n} Sx_j + \tau'_j = S\left(\sum_{j=1}^{n} x_j\right) + \sum_{j=1}^{n} \tau'_j$. By accepting lower privacy guarantees for the individual sketches, or by combining differential privacy with secure aggregation as discussed in Section 2.3.3, we may choose τ'_j appropriately to ensure that $\sum_j \tau'_j$ follows a noise distribution which ensures that the combined sketch achieves a specific privacy guarantee.

As the noise distribution for $'_j$ will usually depend on n, an especially interesting question concerns when the number of sketches merged is unknown, in which case it is not clear how to avoid a signi cant increase in the noise level (and so decreasing accuracy).

As mentioned, our noisy KOR sketch can be used to estimate the size of the union of streaming data and multisets. Without the privacy constraint, it is known that we can use (non-private) linear sketches to estimate the size of the symmetric di erence, but it would be interesting to understand how to estimate the size of the symmetric di erence for such data *with* privacy.

3.8 Technical Details

In this section, we give the technical details and proofs omitted in the previous sections.

3.8.1 Omitted Proofs for Expected Number of Ones in Sketch

Lemma 3.3. *For each $i = 0; :::; \log(u)$ 1 let $L_i = kS_i x k_0$ and $Z_i = kS_i x + '_i k_0$. Then:*

$$\underset{s\ S}{\underset{h\ F;}{\mathsf{E}}}[L_i] = \frac{n}{2}\left(1 \quad \prod_{j:x_j=1}\left(1 \quad \frac{w_j}{2^i n}\right)\right) \tag{3.2}$$

$$\underset{'_i\ D_p}{\underset{s\ S;}{\underset{h\ F;}{\mathsf{E}}}}[Z_i] = \frac{n}{2}\left(1 \quad (1 \quad 2p)\prod_{j:x_j=1}\left(1 \quad \frac{w_j}{2^i n}\right)\right) \tag{3.3}$$

Proof. Let A be the input set with corresponding weight vector w. Let $v_i \ 2 \ Z^n{}_0$ be a vector such that for each $k \ 2 \ [n]$

$$(v_i)_k = \sum_{j \ 2 A} \mathbf{1}\left[\frac{s(j)}{w_j} \ 2 \ \left(1=2^{i+1}; 3=2^{i+1}\right]\right] \quad \mathbf{1}[h(j) = k]:$$

That is, each entry $(v_i)_k$ is the number of candidates for entry k in the sketch at level i, i.e., the number of items j that hash to k and satisfy $\frac{s(j)}{w_j} \ 2 \ \left(1=2^{i+1}; 3=2^{i+1}\right]$. Since $s(j)$ is uniform, we have for such a candidate

$$\Pr_{s\ S}\left[s(j) \ 2 \ \left(w_j=2^{i+1}; 2w_j=2^{i+1}\right] \ \middle| \ s(j) \ 2 \ \left(w_j=2^{i+1}; 3w_j=2^{i+1}\right]\right] = \frac{1}{2}:$$

If there is at least one candidate for entry k then, by the Principle of Deferred Decisions, the probability that we sample an odd number of these is 1=2 and so for $i = 0; :::; \log(u)$ 1

$$\underset{s\ S}{\underset{h\ F}{\Pr}}[(S_i x_A)_k = 1 \ j \ (v_i)_k \ 6= 0] = \frac{1}{2} \qquad \text{and} \qquad \underset{s\ S}{\underset{h\ F}{\Pr}}[(S_i x_A)_k = 1 \ j \ (v_i)_k = 0] = 0:$$

As

$$\Pr_{s\ S}\left[\frac{s(j)}{w_j} \ 2 \ \left(1=2^{i+1}; 3=2^{i+1}\right]\right] = \Pr_{s\ S}\left[s(j) \ 2 \ \left(w_j=2^{i+1}; 3w_j=2^{i+1}\right]\right] = \frac{w_j}{2^i};$$

we have

$$\underset{s\ S}{\underset{h\ F}{\Pr}}[(v_i)_k \ 6= 0] = 1 \quad \prod_{j \ 2 A}\left(1 \quad \frac{w_j}{2^i n}\right):$$

We conclude that

$$\underset{s\ S}{\underset{h\ F;}{\Pr}}[(S_i x_A)_k = 1] = \frac{1 \quad \prod_{j \ 2 A}\left(1 \quad \frac{w_j}{2^i n}\right)}{2}:$$

and letting $L_i = \sum_{k=1}^{n} (S_i x_A)_k$, we get

$$\mathop{\mathbb{E}}_{\substack{h \\ s}}^{F;} [L_i] = \frac{n}{2}\left(1 \quad \prod_{j \in A}\left(1 \quad \frac{w_j}{2^i n}\right)\right)$$

We similarly compute an expression for $\mathbb{E}_{h \quad F;s \quad S;' _i \quad D_p}[Z_i]$. Let $' _i$ be the restriction of a randomness vector $' \quad D''$ to the entries that are added to $S_i x_A$ when adding $'$ to $S x_A$. We see that

$$\begin{aligned}
\Pr_{\substack{h \\ s \\ ' _i}}^{F; \\ S; \\ D_p}\left[(S_i x_A + ' _i)_k = 1\right] &= \Pr_{\substack{h \\ s \\ ' _i}}^{F; \\ S; \\ D_p}\left[(S_i x_A + ' _i)_k = 1 \mid (S_i x_A)_k = 1\right] \Pr_{\substack{h \\ s}}^{F; \\ S}\left[(S_i x_A)_k = 1\right] \\
&\quad + \Pr_{\substack{h \\ s \\ ' _i}}^{F; \\ S; \\ D_p}\left[(S_i x_A + ' _i)_k = 1 \mid (S_i x_A)_k = 0\right] \Pr_{\substack{h \\ s}}^{F; \\ S}\left[(S_i x_A)_k = 0\right] \\
&= (1 \quad p) \Pr_{\substack{h \\ s}}^{F; \\ S}\left[(S_i x_A)_k = 1\right] + p \Pr_{\substack{h \\ s}}^{F; \\ S}\left[(S_i x_A)_k = 0\right] \\
&= (1 \quad p) \frac{1}{2}\left(1 \quad \prod_{j \in A}\left(1 \quad \frac{w_j}{2^i n}\right)\right) + p \left(1 \quad \frac{1 \quad \prod_{j \in A}\left(1 \quad \frac{w_j}{2^i n}\right)}{2}\right) \\
&= \frac{1}{2} \quad \left(\frac{1}{2} \quad p\right) \prod_{j \in A}\left(1 \quad \frac{w_j}{2^i n}\right)
\end{aligned}$$

showing that

$$\mathop{\mathbb{E}}_{\substack{h \\ s \\ ' _i}}^{F; \\ S; \\ D_p}[Z_i] = \frac{n}{2}\left(1 \quad (1 \quad 2p) \prod_{j \in A}\left(1 \quad \frac{w_j}{2^i n}\right)\right):$$

□

3.8.2 Omitted Proofs for Concentration Bounds for Number of Ones in Sketch

Lemma 3.4. *For $i = 0; \ldots; \log(u) \quad 1$ let $Z_i = \|S_i x + ' _i\|_0$. For any $0 < \quad < 1$, we have with probability at least $1 \quad 6\log(u) e^{\frac{2p^3 n}{6^2 3}}$ that for all $i = 0; \ldots; \log(u) \quad 1$ simultaneously:*

$$(1 \quad) \mathop{\mathbb{E}}_{\substack{h \\ s \\ ' _i}}^{F; \\ S; \\ D_p}[Z_i] < Z_i < (1 + \ ) \mathop{\mathbb{E}}_{\substack{h \\ s \\ ' _i}}^{F; \\ S; \\ D_p}[Z_i]:$$

Before proving Lemma 3.4, we mention the following lemma:

Lemma 3.6. *Let $L_i = \|S_i x_A\|_0$. For any $0 < \ ' < 1$, we have with probability at least $1 \quad 4\log(u) e^{2 \ ' ^2 n}$*

$$\mathop{\mathbb{E}}_{\substack{h \\ s}}^{F; \\ S}[L_i] \quad 2 \ ' n \quad L_i \quad \mathop{\mathbb{E}}_{\substack{h \\ s}}^{F; \\ S}[L_i] + 2 \ ' n$$

for all $i = 0; \ldots; \log(u) \quad 1$ simultaneously.

Proof. Let A be the input set and w the corresponding weight vector. Let $v_i \in \mathbb{Z}^n{}_0$ be a vector such that for each $k \in [n]$

$$(v_i)_k = \sum_{j \in A} \mathbf{1}\left[\frac{s(j)}{w_j} \in \left(1=2^{i+1}; 3=2^{i+1}\right]\right] \quad \mathbf{1}[h(j) = k]$$

so $(v_i)_k$ is the number of candidates for entry k in the sketch at level i. Let $V_i = kv_ik_0 = \sum_{k=1}^{n} \mathbf{1}[(v_i)_k \neq 0]$. V_i is a sum of negatively associated random variables (for definition and argument see Section 4.1 in [53]), so by Theorem 4.3 in [53], we can use the Hoeffding bound to see that with probability at least $1 - 2e^{-2n'^2}$ we have for any $i = 0, \ldots, \log(u) - 1$

$$\mathrm{E}[V_i] - {}'n \leq V_i \leq \mathrm{E}[V_i] + {}'n. \tag{3.5}$$

Let $L_i = kS_ix_Ak_0 = \sum_{k=1}^{n}(S_ix_A)_k$ denote the number of ones in the linear sketch. For fixed V_i, L_i is a sum of independent random variables with (by the principle of deferred decisions)

$$\Pr\left[(S_ix_A)_k = 1 \mid (v_i)_k \neq 0\right] = \frac{1}{2}, \qquad \Pr\left[(S_ix_A)_k = 1 \mid (v_i)_k = 0\right] = 0.$$

So for any fixed $V_i = t$

$$\mathrm{E}\left[L_i \mid V_i = t\right] = \frac{t}{2}. \tag{3.6}$$

Furthermore, as L_i is a sum of independent random variables for a fixed choice of V_i, we can use the Hoeffding bound: with probability at least $1 - 2e^{-2n'^2}$

$$\mathrm{E}\left[L_i \mid V_i = t\right] - {}'n \leq L_i|_{V_i = t} \leq \mathrm{E}\left[L_i \mid V_i = t\right] + {}'n,$$

where $L_i|_{V_i = t}$ means the value of L_i when we assume that $V_i = t$. Combining this with (3.5) and (3.6) a union bound gives with probability at least $1 - 4e^{-2n'^2}$

$$\frac{\mathrm{E}[V_i] - {}'n}{2} - {}'n \leq L_i \leq \frac{\mathrm{E}[V_i] + {}'n}{2} + {}'n. \tag{3.7}$$

Simultaneously, (3.5) and (3.6) gives

$$\frac{\mathrm{E}[V_i] - {}'n}{2} \leq \mathrm{E}[L_i] \leq \frac{\mathrm{E}[V_i] + {}'n}{2}, \tag{3.8}$$

which implies

$$2\,\mathrm{E}[L_i] - {}'n \leq \mathrm{E}[V_i] \leq 2\,\mathrm{E}[L_i] + {}'n. \tag{3.9}$$

Note that in the union bound from (3.7), we already assumed that (3.5) was satisfied, so (3.9) is trivially satisfied under the union bound without changing the probability guarantees. Hence, inserting (3.9) into (3.7), we have

$$\frac{2\,\mathrm{E}[L_i] - 2\,{}'n}{2} - {}'n \leq L_i \leq \frac{2\,\mathrm{E}[L_i] + 2\,{}'n}{2} + {}'n. \tag{3.10}$$

which finally shows that with probability at least $1 - 4e^{-2n'^2}$ we have

$$\mathrm{E}[L_i] - 2\,{}'n \leq L_i \leq \mathrm{E}[L_i] + 2\,{}'n.$$

A union bound over the $\log(u)$ values of i concludes the proof. □

We are now ready to prove Lemma 3.4.

Proof of Lemma 3.4. Fix i. Let $L_i = kS_ix_Ak_0$ and $Z_i = kS_ix_A + {}'_ik_0$. We let $Z_i|_{L_i = t}$ be the number of ones in $S_ix_A + {}'_i$, assuming that $L_i = t$. For any fixed value $t \in \{0, \ldots, n\}$ of L_i, we have

$${}'_i\,\mathrm{E}_{D_p}\left[Z_i|_{L_i = t}\right] = (1 - p)\,t + p(n - t) = np + (1 - 2p)t. \tag{3.11}$$

By Lemma 3.6, with probability at least $1 - 4\log(u)e^{-2\varepsilon'^2 n}$ we have for all $i = 0, \ldots, \log(u) - 1$

$$\mathop{\mathrm{E}}_{\substack{h \in F;\\ s \in S;\\ \varepsilon' \in D_p}}[Z_i] \geq np + (1 - 2p)\left(\mathop{\mathrm{E}}_{\substack{h \in F;\\ s \in S}}[L_i] - 2\varepsilon' n\right) \tag{3.12}$$

$$\mathop{\mathrm{E}}_{\substack{h \in F;\\ s \in S;\\ \varepsilon' \in D_p}}[Z_i] \leq np + (1 - 2p)\left(\mathop{\mathrm{E}}_{\substack{h \in F;\\ s \in S}}[L_i] + 2\varepsilon' n\right) \tag{3.13}$$

Furthermore, for any fixed S_i, let $Z_i|_{S_i}$ denote the number of ones in $S_i x_A + \varepsilon'_i$, conditioned on this choice of S_i. We note that fixing S_i is equivalent to fixing L_i as L_i is uniquely determined by S_i and the input. $Z_i|_{S_i}$ is a sum of independent random variables, where the randomness comes from the perturbation. So for any $0 < \delta < 1$, a Chernoff bound gives

$$\Pr_{\varepsilon'_i \in D_p}\left[Z_i|_{S_i} > (1+\delta)\,\mathrm{E}\left[Z_i|_{S_i}\right] \vee Z_i|_{S_i} < (1-\delta)\,\mathrm{E}\left[Z_i|_{S_i}\right]\right] \leq 2e^{-2\delta^2 \mathrm{E}[Z_i|_{S_i}]/3} \tag{3.14}$$

where $\mathrm{E}\left[Z_i|_{S_i}\right]$ is over $\varepsilon'_i \in D_p$. By (3.11), $\mathrm{E}_{\varepsilon'_i \in D_p}\left[Z_i|_{S_i}\right] \geq np$ for any choice of S_i, so $2e^{-2\delta^2 pn/3}$ is an upper bound on (3.14). Moreover, (3.14) holds for all $i = 0, \ldots, \log(u) - 1$ simultaneously with probability at most $2\log(u)e^{-2\delta^2 pn/3}$. We conclude that

$$\Pr_{\varepsilon'_i \in D_p}\left[\forall i : (1-\delta)\,\mathrm{E}\left[Z_i|_{S_i}\right] < Z_i|_{S_i} < (1+\delta)\,\mathrm{E}\left[Z_i|_{S_i}\right]\right] \geq 1 - 2\log(u)e^{-2\delta^2 pn/3} \tag{3.15}$$

Combining (3.12), (3.13) and (3.15) and letting $\delta' = \delta$, we have by a union bound that for all levels i simultaneously, where the expectation is over $h \in F$ and $s \in S$

$$Z_i \geq (1-\delta')(np + (1-2p)(\mathrm{E}[L_i] - 2\varepsilon' n))$$
$$Z_i \leq (1+\delta')(np + (1-2p)(\mathrm{E}[L_i] + 2\varepsilon' n));$$

with probability at least

$$1 - \left(4\log(u)e^{-2n\varepsilon'^2} + 2\log(u)e^{-\varepsilon'^2 pn/3}\right) \geq 1 - 6\log(u)e^{-\varepsilon'^2 pn/3}.$$

By Lemma 3.3, this is equivalent to

$$Z_i \geq (1-\delta')(\mathrm{E}[Z_i] - 2(1-2p)\varepsilon' n) \tag{3.16}$$
$$Z_i \leq (1+\delta')(\mathrm{E}[Z_i] + 2(1-2p)\varepsilon' n). \tag{3.17}$$

where the expectation is over $h \in F; s \in S$ and $\varepsilon'_i \in D_p$. We pick a suitable ε':

$$\varepsilon' = \frac{p}{6} \quad \Rightarrow \quad 2(1-2p)\varepsilon' n = (1-2p)\frac{p}{3}n \quad \Rightarrow \quad 2(1-2p)\varepsilon' n \leq \frac{(1-2p)}{3}\mathrm{E}[Z_i].$$

Hence, let $\varepsilon' = \frac{p}{6}$. Inserting into (3.16) and (3.17) we have

$$Z_i \geq \left(1 - \frac{p}{6}\right)\left(\mathrm{E}[Z_i] - \frac{(1-2p)}{3}\mathrm{E}[Z_i]\right)$$
$$Z_i \leq \left(1 + \frac{p}{6}\right)\left(\mathrm{E}[Z_i] + \frac{(1-2p)}{3}\mathrm{E}[Z_i]\right)$$

where $\mathrm{E}[Z_i]$ is over $h \quad F; s \quad S$ and $' _i \quad D_p$.

We conclude that with this choice of , with probability at least $1 \quad 6\log(u)e^{\frac{{}^2p^3n}{6^2\ 3}}$

$$(1 \quad) \underset{\substack{h\ F;\\ s\ S;\\ ' _i\ D_p}}{\mathrm{E}}[Z_i] \quad Z_i \quad (1+ \ ) \underset{\substack{h\ F;\\ s\ S;\\ ' _i\ D_p}}{\mathrm{E}}[Z_i]:$$

□

3.8.3 Omitted Proofs for Size of Interval for Input Size

Before proving Lemma 3.5, we give a technical lemma:

Lemma 3.7. *For any* $0 < \ < \frac{1}{\frac{2e^3}{1\ 2p} \quad 1}$ *any value*

$$\hat{m} \ 2 \left[2^i n \ln\left(\frac{1 \quad 2p}{1 \quad \frac{2Z_i}{(1+ \ )n}} \right); 2^i n \ln\left(\frac{1 \quad 2p}{1 \quad \frac{2Z_i}{(1 \quad)n}} \right) \right] \tag{3.18}$$

satis es

$$\hat{m} \quad (1 \quad) 2^i n \ln\left(\frac{1}{\prod_{j\ 2A}\left(1 \quad \frac{w_j}{2^i n}\right)} \right)$$

$$\hat{m} \quad (1+ \ ) 2^i n \ln\left(\frac{1}{\prod_{j\ 2A}\left(1 \quad \frac{w_j}{2^i n}\right)} \right)$$

for

$$= \frac{6 \ \left(\frac{e^3}{1\ 2p} \quad 1\right)}{(1 \quad) \quad 2 \ \left(\frac{e^3}{1\ 2p} \quad 1\right)}$$

with probability at least $1 \quad 6\log(u)e^{\ {}^2p^3n=108}$ *for the i where* $\frac{kwk_1}{2^i n} \ 2 \ [1; 2]$.

Proof. By Lemma 3.3

$$\prod_{j\ 2A}\left(1 \quad \frac{w_j}{2^i n}\right) = \frac{1 \quad \frac{2\,\mathrm{E}[Z_i]}{n}}{1 \quad 2p}$$

and so by Lemma 3.4, with probability at least $1 \quad 6\log(u)e^{\ {}^2p^3n=108}$ we have for any $0 < \ < 1$ that for all $i = 0; :::; \log(u) \quad 1$ simultaneously.

$$\frac{1 \quad \frac{2Z_i}{(1 \quad)n}}{1 \quad 2p} < \prod_{j\ 2A}\left(1 \quad \frac{w_j}{2^i n}\right) < \frac{1 \quad \frac{2Z_i}{(1+ \ )n}}{1 \quad 2p}: \tag{3.19}$$

For convenience, we consider the slightly bigger interval { note that if (3.19) is satis ed, then so is this interval:

$$\frac{1 \quad \frac{2(1+ \ )\,\mathrm{E}[Z_i]}{(1 \quad)n}}{1 \quad 2p} < \prod_{j\ 2A}\left(1 \quad \frac{w_j}{2^i n}\right) < \frac{1 \quad \frac{2(1 \quad)\,\mathrm{E}[Z_i]}{(1+ \ )n}}{1 \quad 2p};$$

where the left-hand side can be reordered as

$$\left(1 \quad \frac{2}{1} \left(\frac{1}{(1 \quad 2p)\prod_{j\ 2A}\left(1 \quad \frac{w_j}{2^i n}\right)} \quad 1 \right)\right) \prod_{j\ 2A}\left(1 \quad \frac{w_j}{2^i n}\right) \tag{3.20}$$

and the right-hand side as

$$\left(1 + \frac{2}{1+} \left(\frac{1}{(1 \quad 2p)\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)} \quad 1\right)\right)\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right): \tag{3.21}$$

We will bound this interval further using the following claim:

Claim 3.1. *De ne*

$$:= \frac{2}{1} \left(\frac{e^{2+\frac{1}{2^i}\frac{1}{1}n}}{1 \quad 2p} \quad 1\right):$$

Whenever $\frac{kwk_1}{2^i n} < 2$*, the interval de ned by (3.20) and (3.21) is contained in*

$$\left[(1 \qquad)\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right); (1+ \quad)\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)\right]$$

Proof of Claim. As $\frac{2}{1+} < \frac{2}{1}$, we increase (3.21) to

$$\left(1 + \frac{2}{1} \left(\frac{1}{(1 \quad 2p)\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)} \quad 1\right)\right)\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right):$$

Observing that when $\frac{kwk_1}{2^i n} \quad 2$

$$\frac{2}{1}\left(\frac{1}{(1 \quad 2p)\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)} \quad 1\right) \quad \frac{2}{1}\left(\frac{e^{\frac{kwk_1}{2^i n}+\frac{kwk_1}{(2^i n)^2}}}{1 \quad 2p} \quad 1\right) \quad \frac{2}{1}\left(\frac{e^{2+\frac{1}{2^i}\frac{1}{1}n}}{1 \quad 2p} \quad 1\right) =:$$

we have the result. □

Applying the claim, we consider the interval:

$$2^i n \ln\left(\frac{1}{\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)}\right) \quad 2^i n \ln\left(\frac{1}{(1+ \quad)\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)}\right) \tag{3.22}$$

$$2^i n \ln\left(\frac{1}{\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)}\right) \quad 2^i n \ln\left(\frac{1}{(1 \qquad)\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)}\right): \tag{3.23}$$

We remind the reader that by construction, this interval contains the target interval (3.18).

We consider the ratio between the end-points of the interval de ned by (3.22) and (3.23). Observe that

$$\frac{2^i n \ln\left(\frac{1}{(1 \qquad)\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)}\right)}{2^i n \ln\left(\frac{1}{(1+ \quad)\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)}\right)} = \frac{\ln\left(\frac{1}{\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)}\right) \quad \ln(1 \qquad)}{\ln\left(\frac{1}{\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)}\right) \quad \ln(1+ \quad)} \quad \frac{\ln\left(\frac{1}{\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)}\right) + \frac{1}{1}}{\ln\left(\frac{1}{\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)}\right)}$$

$$= 1 + \frac{\left(1 + \frac{1}{1}\right)}{\ln\left(\frac{1}{\prod_{j\,2\,A}\left(1 \quad \frac{w_j}{2^i n}\right)}\right)}$$

where the inequality follows from

$$\frac{x}{1+x} \quad \ln(1+x) \quad x; \qquad x > \quad 1:$$

For $< 1=2$, we have

$$\frac{\left(1 + \frac{1}{1}\right)}{\ln\left(\frac{1}{\prod_{j2A}\left(1 - \frac{w_j}{2^i n}\right)}\right)} < \frac{3}{\ln\left(\frac{1}{\prod_{j2A}\left(1 - \frac{w_j}{2^i n}\right)}\right)} < \frac{3}{\frac{kwk_1}{2^i n}}$$

Observe that as $\frac{kwk_1}{2^i n}$ increases, it gets easier to satisfy this inequality. But we remind ourselves of the Claim, where we required $\frac{kwk_1}{2^i n} < 2$. So the interval in (3.22) and (3.23) does not necessarily contain the target interval (3.18) for larger values of $\frac{kwk_1}{2^i n}$. Assume further that $\frac{kwk_1}{2^i n}$ 1. Then

$$\frac{3}{\frac{kwk_1}{2^i n}} < \frac{3}{1}:$$

So, we conclude that with probability at least $1 - 6\log(u)e^{-2p^3 n=108}$, any value in the target interval (3.18) is within a factor $1 + \frac{3}{1}$ of $2^i n \ln\left(\left(\prod_{j2A}\left(1 - \frac{w_j}{2^i n}\right)\right)^{1}\right)$.

Inserting the value of , we obtain an estimate within a factor of

$$1 + \frac{6\left(\frac{e^{2+\frac{1}{2^i 1 n}}}{1 - 2p} - 1\right)}{(1 -)2\left(\frac{e^{2+\frac{1}{2^i 1 n}}}{1 - 2p} - 1\right)} < 1 + \frac{6\left(\frac{e^3}{1 - 2p} - 1\right)}{(1 -)2\left(\frac{e^3}{1 - 2p} - 1\right)}:$$

Thus it su ces that

$$< \frac{1}{\frac{2e^3}{1 - 2p} - 1}:$$

□

We are now ready to prove Lemma 3.5:

Lemma 3.5. *Assume $kw_x k_1 > n > 1$, and $> \frac{1}{n}$. With probability at least $1 - 6\log(u)e^{-\frac{2p^3 n}{108}}$ there exists an $i \in \{0; :::; \log(u) - 1\}$ such that any element from $I_i(p)$ is a $(1 +)$-approximation to $kw_x k_1$ for*

$$< \frac{\left(\frac{1}{n}\right)(1 - 2p)}{7e^3}:$$

Speci cally, i where $\frac{kw_x k_1}{2^i n} \in [1; 2)$, gives these guarantees.

Proof. We will choose in terms of the accuracy parameter , such that with high probability any estimate from the interval

$$\left[2^i n \ln\left(\frac{1 - 2p}{1 - \frac{2Z_i}{(1+)n}}\right); 2^i n \ln\left(\frac{1 - 2p}{1 - \frac{2Z_i}{(1-)n}}\right)\right] \tag{3.24}$$

is within a factor $(1 +)$ of kwk_1. We do this in a few steps: First, we show that any value from (3.24) is a good estimate of

$$2^i n \ln\left(\frac{1}{\prod_{j2A}\left(1 - \frac{w_j}{2^i n}\right)}\right): \tag{3.25}$$

As $2^i n \ln\left(\frac{1}{e^{\frac{kwk_1}{2^i n}}}\right) = kwk_1$ and

$$\frac{2^i n \ln\left(\frac{1}{\prod_{j \in A}(1 - \frac{w_j}{2^i n})}\right)}{2^i n \ln\left(\frac{1}{e^{\frac{kwk_1}{2^i n}}}\right)} \quad \frac{\ln\left(e^{\frac{kwk_1}{2^i n} + \frac{kwk_1}{(2^i n)^2}}\right)}{\ln\left(e^{\frac{kwk_1}{2^i n}}\right)} = 1 + \frac{1}{2^i n}$$

where we used the Taylor expansion of the exponential function, we have

$$kwk_1 \quad 2^i n \ln\left(\frac{1}{\prod_{j \in A}(1 - \frac{w_j}{2^i n})}\right) \quad \left(1 + \frac{1}{2^i n}\right) kwk_1.$$

So a good estimate for (3.25) will allow for a good estimate of kwk_1. The technical lemma, Lemma 3.7, shows that as long as kwk_1 is sufficiently large, that is, there is an i such that $\frac{kwk_1}{2^i n} \in [1, 2)$, we get a suitable estimate for (3.25) with the interval (3.24) with high probability.

Hence, any value from (3.24) is within a factor $(1 + \ )$ of kwk_1 for

$$< \frac{(\quad 1=n)(1 \quad 2p)}{7e^3} < \frac{1=n}{7\left(\frac{e^3}{1 \quad 2p} \quad 1\right)} < \frac{\frac{1}{2^i n}}{7\left(\frac{e^3}{1 \quad 2p} \quad 1\right)}$$

for $> \frac{1}{n}$. We will choose n in terms of such that this is always satisfied. Clearly, this value of is significantly smaller than the requirement from Lemma 3.7, which concludes the proof. □

Acknowledgements

We thank Shuang Song and Abhradeep Guha Thakurta for feedback on a previous version of this manuscript.

Chapter 4

Differentially Private Euclidean Distance Approximation

This chapter is based on the paper *Improved Di erentially Private Euclidean Distance Approximation* by Nina Mesing Stausholm [130].

4.1 Introduction

The Euclidean distance between real-valued vectors is an important measure with applications in various elds such as nearest-neighbor search, clustering, numerical linear algebra, just to mention a few. Matrices satisfying the Johnson-Lindenstrauss Lemma (Lemma 2.4), the so-called Johnson-Lindenstrauss matrices, can be used to compute linear sketches of input vectors that allow for estimating the Euclidean distance between two real-valued input vectors from their sketches: For k u-sketch matrix S and inputs $x; y$ 2 R^u held by di erent parties, one can estimate the Euclidean distance between x and y as $k\,Sx$ $Sy\,k_2^2$ = $k\,S(x$ $y)\,k_2^2$. By the Johnson-Lindenstrauss Lemma, this estimate is within a factor (1) of kx yk_2^2 with high probability. We henceforth use *transform* and *projection* interchangeably and refer to random projections satisfying the Johnson-Lindenstrauss Lemma as *Johnson-Lindenstrauss projections*, or simply *JL projections*. We will even misuse this convention slightly, as we also use this name for projections that preserve Euclidean norm in *expectation*, as de ned in De nition 4.2.

As the input x may contain sensitive information, we wish to ensure that the released sketch is di erentially private and therefore add noise to the sketch Sx. For noise vector ' D^k, we denote by Sx + ' the noisy counter-part to Sx. The main goal of this chapter is to analyze the privacy and utility guarantees of the noisy sketch Sx + '.

That is, for noise distribution D and noise vectors '; 2 D^k, we analyze whether we can privately and accurately estimate kx yk_2^2 from Sx + ' and Sy + . The main questions of interest are: *How much noise do we need to add?* and *What utility guarantees can we achieve?* We de ned di erential privacy in Section 2.1.2 and mention common choices for noise distribution D in Section 4.3.2.

4.1.1 Differentially Private Random Projections

The results presented in this chapter improve on the work of Kenthapadi et al. [96], in which it was shown how to construct an (";)-di erentially private version of a JL transform allowing for high accuracy estimators for squared Euclidean distance. The idea applied by Kenthapadi et al. is simple: Let S_{iid} be the i.i.d. normally distributed JL transform where each entry is drawn from the standard Normal distribution. For input vector x 2 $[0; 1]^u$, add Gaussian noise to each entry of $S_{iid}x$.

Kenthapadi et al. prove Theorems 4.1 and 4.2. We remark that these results extend naturally to $x; y$ 2 R^u.

Theorem 4.1 ([96]). *Let S_{iid} be a $k \times u$-projection matrix with i.i.d. entries from the standard Normal distribution and let $x, y \in [0,1]^u$ be input vectors. Let $\eta, \eta' \sim N(0, \sigma^2)^k$ be noise vectors. If $\varepsilon = \varepsilon'\sqrt{\log(1/\delta)}$, $\varepsilon' < \ln(1/\delta)$ and $k > 2(\ln(u) + \ln(2/\delta))$, then $S_{iid}x + \eta$ is (ε', δ)-differentially private.*

Note 4.1. *Kenthapadi et al. show that for $k > 2\ln(u) + 2\ln(1/\delta')$, the ℓ_2-sensitivity of S_{iid} is greater than 2 with probability at most δ'. We will assume that the ℓ_2-sensitivity of S_{iid} is computed exactly in an initializing step, as discussed in Section 4.2.1, and hence avoid this assumption on k. From [89, 92] we know that for any $\varepsilon, p \in (0,1/2)$ $k = \Theta(\varepsilon^{-2}\log(1/p))$ is optimal in the non-private case. We use this value of k and discuss the optimal k for the noisy construction in Section 4.6.2. We also remark that the ε in Theorem 4.1 can be exchanged with $\varepsilon_2 \cdot \sqrt{2\log(1.25/\delta)}$ by a later result from [63] (See Lemma 4.2), where Δ_2 is the (exact) ℓ_2-sensitivity of S_{iid}.*

Theorem 4.2 ([96]). *Let S_{iid} be a $k \times u$-projection matrix with i.i.d. entries from the standard Normal distribution and $x, y \in [0,1]^u$ be input vectors. Let $\eta, \eta' \sim N(0, \sigma^2)^k$ be noise vectors, where σ is independent of the realization of S_{iid}. Define*

$$\hat{E}_{iid} := \|(S_{iid}x + \eta) - (S_{iid}y + \eta')\|_2^2 - 2k\sigma^2.$$

Then

1. *$\hat{E}_{iid}$ is an unbiased estimator for $\|x - y\|_2^2$.*
2. *$\mathrm{Var}\left[\hat{E}_{iid}\right] = \frac{2}{k}\|x - y\|_4^4 + 8\sigma^2\|x - y\|_2^2 + 8\sigma^4 k$.*

Note 4.2. *Letting σ be independent of the realization of S_{iid} might lead to complete loss of privacy if the ℓ_2-sensitivity of S_{iid} is much higher than 1, as argued in Section 4.2.1. Hence, we let σ be a function of the exact ℓ_2-sensitivity of S_{iid}, Δ_2, as discussed in Note 4.1.*

4.1.2 Contributions

An immediate idea to achieve a speed-up is to apply the techniques of Kenthapadi et al. to a JL transform, which is faster than the i.i.d. normally distributed JL transform. We show such a result for a private Fast Johnson-Lindenstrauss Transform (FJLT) [3] in Section 4.5.2, but remark that the privacy issue of Kenthapadi et al. mentioned in Note 4.2 carries over if we simply exchange the i.i.d. normally distributed JL transform for the FJLT. We discuss how to address this issue in Section 4.5 to obtain a differentially private version of FJLT, where σ does not depend on the ℓ_2-sensitivity of the transform (which could be very large). Kenthapadi et al. leave open the question of whether we can obtain better results with Laplace noise. We answer this question by proving that we can indeed obtain an ε-differentially private estimator for squared Euclidean distances, which has better variance for certain parameters. Specifically, we show the following main theorem:

Theorem 4.3. *For any $0 < \varepsilon, p < 1/2$ and any integer $u > 0$ there exists a random $k \times u$-projection S for $k = \Theta(\varepsilon^{-2}\log(1/p))$ with sparsity $s = O(\varepsilon^{-1}\log(1/p))$ and a distribution D over $\mathbb{R}$ such that for any $x, y \in \mathbb{R}^u$ and $\eta, \eta' \sim D^k$ we define:*

$$\hat{E}_{SJLT} := \|(Sx + \eta) - (Sy + \eta')\|_2^2 - \frac{2ks}{\varepsilon'^2}\sigma^2.$$

Then

1. *$\hat{E}_{SJLT}$ is an unbiased estimator for $\|x - y\|_2^2$.*
2. *$\mathrm{Var}\left[\hat{E}_{SJLT}\right] \leq \frac{2}{k}\|x - y\|_4^4 + O\left(\frac{s}{\varepsilon'^2}\|x - y\|_2^2 + \frac{s^2}{\varepsilon'^4}k\right)$.*
3. *The sketch $(S, Sx + \eta)$ is ε-differentially private.*

4. *For a data stream, we can update the sketch* $(S; Sx + \ ')$ *in time* $O(s)$.

5. $Sx + \ '$ *can be computed in time* $O(skxk_0 + k)$. *Given* $Sx + \ '$ *and* $Sy +$, $\hat{E}_{SJLT}$ *can be computed in time* $O(k)$.

The noise distribution D will depend on the sparsity of S, but it is crucial that D is otherwise independent of S. We state our improvements over the work of Kenthapadi et al. [96]:

- Recall that the projection S_{iid} of Kenthapadi et al. has constant $`_2$-sensitivity with high probability. Under this assumption, we combine Theorems 4.1, 4.2 and Note 4.1 to see that

$$\text{Var}[\hat{E}_{iid}] = \frac{2}{k}kx \quad yk_2^4 + O\left(\frac{\log(1=\ )}{"^2}kx \quad yk_2^2 + \frac{\log^2(1=\ )}{"^4}k\right);$$

  and so $\hat{E}_{SJLT}$ improves over $\hat{E}_{iid}$ in terms of variance whenever $< e^{\ s} = p^{O(1=\ )}$ (see Section 4.7). In the case where S_{iid} has higher sensitivity, our results give an even better improvement.

- Kenthapadi et al. have an additional initialization cost of $O(uk)$ to compute the sensitivity of the projection matrix. We refer the reader to Section 4.2.1 for a detailed discussion.

- Our estimator $\hat{E}_{SJLT}$ is more e cient as the update time, i.e., time to compute $Sx + \ '$, is $O(skxk_0 + k)$ rather than $O(kkxk_0 + k)$ for $s = o(k)$.

- Rather than *approximate* di erential privacy, we achieve *pure* di erential privacy (See Section 2.1.2 for comments on the di erence).

Our improved e ciency in Theorem 4.3 relies on the sparsity of the Sparser JL transforms by Kane & Nelson [93], henceforth referred to as *the SJLT*. We remark that the results of Kenthapadi et al. extend naturally to these JL transforms, and thus they would obtain the same e ciency for $> p^{O(1=\ )}$. We do, although, give the analysis proving that these transforms can indeed be used. Using a SJLT instead of the i.i.d. normally distributed transform, the work of Kenthapadi et al. would also avoid the initialization cost.

Related to our analysis for the SJLT, we remark that our main result is, in fact, a special case of an even more general result: we give a class of length preserving linear transformations that allow for e cient, private estimators for Euclidean distance with a high level of utility. The FJLT and SJLT are merely examples of such linear transformations. We de ne what is meant by *length preserving* in section 4.3.3 and prove our general, technical results in Section 4.4. In Section 4.5 we give two di erentially private versions of FJLT, and in Section 4.6, we prove Theorem 4.3 by applying the technical results to the SJLT with noise from the Laplace distribution. Finally, we compare the work of Kenthapadi et al. with our private FJLT and SJLT in Section 4.7.

4.2 Related Work

Di erential privacy is usually achieved by adding random noise to the output of a query to obfuscate the exact result before publishing the result. This idea is easily extended to vector outputs by simply adding noise to each entry of the output vector. This technique has been studied extensively in previous work; see for example [84, 104, 110, 119].

We consider a distributed setting, where party i adds noise $' _i \quad D^k$ to the projection Sx_i of input vector x_i and releases the noisy projection $Sx_i + \ ' _i$ for future distance estimation. All parties must use the same randomized matrix S and noise drawn from the same distribution D. It is crucial that the projection matrix is public and only the noise be kept secret.

4.2.1 Versions of Johnson-Lindenstrauss Transformations

We refer to the classical JL transform by Indyk & Motwani [87] as the *i.i.d. normally distributed JL transform*. As the name suggests, the random projection matrix consists of i.i.d. entries from the standard Normal distribution.

The *sparsity* of the random projection, i.e., the number of non-zero entries per column, is an important tool in speeding up dimensionality reduction. Ailon & Chazelle [3] presented a JL transform with a sparser projection matrix with a mixture of normally distributed entries and 0s. This transform is commonly known as *The Fast Johnson-Lindenstrauss Transform* or, in short, *FJLT*. We describe the transform in detail in Section 4.5.1. The sparsity not only a ects the sensitivity of the transformation (see De nition 4.1) but also the time required to compute the projection of an input vector x. For a random projection S with sparsity s, we can compute Sx in time $O(skxk_0)$. Kane & Nelson [93] show that the JL transform of Dasgupta et al. [43] requires sparsity $s = \tilde{}(\ ^1 \log^2(1=p))$, and Nelson & Nguyen showed that this sparsity is optimal up to a factor $O(\log(1=))$ [116]. Kane & Nelson [93] also give two sparser constructions with $s = (\ ^1 \log(1=p))$ for embedding into $k = (\ ^2 \log(1=p))$ dimensions. These transformations are commonly known as *The Sparser JL Transforms* and we will henceforth refer to them as *SJLT*. We describe SJLT in Section 4.6.1.

Differentially Private JL Construction

Kenthapadi et al. [96], which was also discussed in Section 4.1.1, give a private estimator for Euclidean distance relying on the i.i.d. normally distributed JL transform. A drawback of their construction is that the $`_2$-sensitivity is only 1 in *expectation*, so the sensitivity *might* not be small. This is the case if the random projection has even a single very large entry. The authors suggest drawing noise calibrated to a low sensitivity projection matrix independently of the actual projection matrix S_{iid}. However, with a small probability, S_{iid} *does not* have low sensitivity, in which case the noise is not ensured to provide di erential privacy. Kenthapadi et al. "hide" the probability of drawing a high-sensitivity projection under , but for a xed S_{iid}, either the noise provides privacy, or certain inputs would always be distinguishable, even in the presence of noise calibrated to low sensitivity. An alternative solution is to compute the sensitivity of the xed S_{iid} and calibrate the noise to the actual sensitivity. Hence, initialization requires time $O(uk)$. Kenthapadi et al. state without proof that their results extend to the JL transformations from [1, 43]. Xu et al. [147] extend the work of [96] with experimental comparisons with JTree [32], PrivBayes [150], PriView [122] and PrivateSVM [124].

4.2.2 Differentially Private Linear Transformations

Mir et al. (PODS11) [110] suggest a general framework for generating pan-private linear transformations by initializing with noise from the exponential mechanism. The work argues how to create a $"$-pan private estimator for (squared) Euclidean distance with multiplicative error $(1 +)$ and additive error $\text{poly}(\log u; " \ ^1; \ ^1; \log(q \ ^1)) + O(Z)$, with probability at least $1 \ q$, where Z is an upper bound on the entries of the input vector. The technique used by Mir et al. can be used for private dimensionality reduction but is computationally ine cient as the sketch relies on the exponential mechanism for noise addition.

In an earlier (unpublished) version of the same work, [109], Mir et al. analyze the *cropped* second moment for a parameter , de ned for input vector $x \ 2 \ Z^u$ as $\sum_{i=1}^{u} \min\{x_i^2; \ g\}$. In this work, Mir et al. show a $2"$-di erentially private estimator with additive error $O_"(\ ^1 \overline{u})$ with high probability. Di erential privacy is achieved by an application of Randomized Response [141]. As our error depends on $kx \ yk_2$ and $\overline{k} < \overline{u}$, we see an improvement when x and y are sparse. The problems are not directly comparable as the cropped second moment of Mir et al. applies to integer inputs, whereas we consider inputs over the reals.

4.2.3 When Data is Known in Advance

If input data is known in advance, there are other techniques to achieve di erential privacy. A curator with access to all data can compute the *exact* distances (up to the error incurred by the JL embedding) and add

noise precisely calibrated to this distance. This technique often incurs less noise but is not applicable in our setting, as data is split among several parties and may not all be available at once.

Blocki et al. [20] show that, as long as the projection matrix is kept secret, the i.i.d. normally distributed JL transform allows for differentially private estimates of distances with the accuracy guarantees from the Johnson-Lindenstrauss Lemma. Upadhyay [136] proves that this technique does not generally work to preserve privacy for sparser JL projections. As we consider a distributed setting, keeping the projection matrix secret is unattainable. Bhaskar et al. [17] introduce *noiseless privacy* where the output is always exact, rather than a noisy approximation. The privacy guarantees are similar to differential privacy but rely on assumptions about the distribution of the data and auxiliary information, whereas differential privacy aims for a higher level of generality.

Representing Noise from Continuous Noise Distributions

We will assume that noise is drawn from either the continuous Laplace or Gaussian distribution, which, however, may introduce practical issues. Mironov [111] described how privacy might be lost due to floating-point error when sampling noise from a continuous distribution. As an alternative to the continuous Laplace distribution, Mironov suggests the *Snapping mechanism* which incurs an additional error of approximately $_1="$ compared to noise from Lap($_1="$), where $_1$ is the $`_1$-sensitivity of the query.

[47] improve over the Snapping Mechanism, by drawing noise from a discrete distribution, differing from the Laplace distribution by at most a factor $(1 + \frac{1+2="}{2^k})$ for a fixed integer k, which controls the accuracy of the discretization. It suffices to use $k \geq [10; 45]_Z$.

A discrete, "hole-free" alternative to the Gaussian distribution, requiring only expected constant time is suggested in [47]. The distribution builds on the Binomial distribution with parameters n and $p = 1=2$ and the work of [24] to give a distribution which for large n differs from the Gaussian distribution by at most $O(\log^{1:5}(n)=\sqrt{n})$.

In a very recent paper, Canonne et al. [27] describe a discretization of the Gaussian distribution supported on Z whose variance is at most that of the corresponding continuous Gaussian distribution, and hence allows for identical or slightly better utility. Simultaneously, the discretization has sub-Gaussian tails compared to the corresponding continuous Gaussian distribution and essentially the same privacy guarantees. We refer to the discussion in [27] for further reading on discretizations of the Laplace and Gaussian distributions.

4.2.4 Lower bounds

McGregor et al. [103] show that any protocol for estimating Hamming distance (and so for inner product, which again leads to a protocol for estimating squared Euclidean distance) of two binary k-dimensional vectors in a differentially private manner incurs an additive error of $\tilde{}(\sqrt{k})$, which is contrasted by the observation that simple Randomized Response [141] allows for error $O(\sqrt{k})$. The error lower bound implies a $\tilde{}(k)$ lower bound for the variance of the noisy estimator. In contrast, our variance of the noise added (we may disregard the variance introduced by the JL projection, as this error occurs even in the non-private version) depends on $kx \quad yk_2^2 \quad u$ and k (for binary input vectors).

Independently from the work of McGregor et al., Mir et al. [110] show a similar lower bound of additive error $(\sqrt{k})$ for estimating inner product for binary vectors in a pan-private setting. The lower bound by McGregor et al. implies a lower bound for pan-private algorithms, which is weaker than the lower bound of Mir et al. in the case of single-pass algorithms and dynamic data. Hardt & Talwar [82] show that an $"$-differentially private algorithm for the second frequency moment F_2 requires an additive error factor of $(1="^2)$, which is comparable to our result (up to polynomial and logarithmic factors).

4.3 Preliminaries

4.3.1 Sensitivity

We consider a setting where the inputs are vectors $x; y \in \mathrm{R}^u$. Hence, two vectors x and y are neighbors (see De nition 2.2) if $kx \quad yk_1 \quad 1$.

Recall De nition 2.3, where we de ned the sensitivity of a general query. We now restrict this de nition to queries that are linear transformation:

Definition 4.1 ($\grave{}_p$-sensitivity of transformation [96])**.** *For $p \quad 1$, the $\grave{}_p$-sensitivity of a linear transformation $S : \mathrm{R}^u \to \mathrm{R}^k$ is*

$$_p(S) = \max_{x;y2\mathrm{R}^u:\ kx \quad yk_1 \quad 1} kSx \quad Syk_p = \max_{1 \ j \ u} \left(\sum_{i=1}^{k} jS_{i;j}j^p \right)^{1=p} = \max_{1 \ j \ u} kS_{;j}k_p$$

where $S_{;j}$ is the j^{th} column of S.

Note 4.3. *The de nition follows from the observation that any vector of $\grave{}_1$-norm 1 (which is the case for neighboring vectors) can be represented as a convex linear combination of basis vectors.*

4.3.2 Mechanisms In Differential Privacy

Recall the two fundamental techniques in di erential privacy, the Laplace and Gaussian mechanisms as de ned in Lemmas 2.3 and 2.5. We will use these mechanisms extensively in our analysis and repeat them here in terms of linear transforms for convenience:

Lemma 4.1 (Laplace Mechanism [60])**.** *For linear transformation $S \in \mathrm{R}^{k \ u}$ and input $x \in \mathrm{R}^u$, the Laplace Mechanism with parameter outputs $Sx + '$ for $' \quad \mathrm{Lap}(0;)^k$. Let $_1$ be the $\grave{}_1$-sensitivity of S. The Laplace Mechanism with parameter $_1" \ ^1$ preserves "-di erential privacy.*

Lemma 4.2 (Gaussian Mechanism [57, 63])**.** *For linear transform $S \in \mathrm{R}^{k \ u}$ and input $x \in \mathrm{R}^u$, the Gaussian Mechanism with parameter outputs $Sx + '$ for $' \quad N(0; \ ^2)^k$. Let $_2$ be the $\grave{}_2$-sensitivity of S. The Gaussian Mechanism with parameter $_2" \ ^1\sqrt{2\log(1:25=)}$ preserves $(";\ )$-di erential privacy.*

4.3.3 Length Preserving Property

Our technical results in Section 4.4 rely on linear transforms with the *Length Preserving Property (LPP)*:

Definition 4.2 (Length Preserving Property (LPP))**.** *A random $k \quad u$-projection S satis es the* Length Preserving Property *if for any $x \in \mathrm{R}^u$ we have*

$$\mathrm{E}_S \left[kSxk_2^2 \right] = kxk_2^2:$$

Note that if S satis es LPP, then S also preserves Euclidean distances and inner products, as

$$hx; yi = \frac{kxk_2^2 + kyk_2^2 \quad kx \quad yk_2^2}{2}:$$

4.4 Supporting Lemmas

We now show our general, technical lemmas which will be useful for proving Theorem 4.3. Let S be a random $k \quad u$-matrix with LPP as de ned in De nition 4.2 and let $x; y \in \mathrm{R}^u$. Let D be a zero-mean distribution and $'; \quad D^k$ noise vectors. Let $' \quad D$. We de ne

$$\hat{E}_{gen} := k(Sx + ') \quad (Sy + \ )k_2^2 \quad 2k\,\mathrm{E}_D[\ '^2]:$$

Our technical results are as follows:

Lemma 4.3. *We have*

1. $\hat{E}_{gen}$ *is an unbiased estimator for* $kx - yk_2^2$.
2. *The variance of* $\hat{E}_{gen}$ *is*

$$\mathrm{Var}\left[\hat{E}_{gen}\right] = \mathrm{Var}\left[kSx - Syk_2^2\right] + 8\,\mathrm{E}_D[\,'^2]kx - yk_2^2 + 2k\,\mathrm{E}_D[\,'^4] + 2k\,\mathrm{E}_D[\,'^2]^2$$

Proof. See Section 4.9.1. □

Hence, the variance of $\hat{E}_{gen}$ is close to the variance of the non-private estimator but has an additional noise term depending on the output dimension k and the Euclidean distance of the input vectors. The following result describes the privacy guarantees of $\hat{E}_{gen}$:

Lemma 4.4. *Let* $_1$ *and* $_2$ *be the* $`_1$*- and* $`_2$*-sensitivities of* S*, respectively. Let* > 0 *be given and de ne*

$$m := \min\left\{ \ _1; \ \ _2\sqrt{\ln(1=\ )}\right\}:$$

There is a distribution D *such that*

1. *The sketch* $(S; Sx + \ ')$ *is di erentially private.*
2. $\mathrm{Var}\left[\hat{E}_{gen}\right] = \mathrm{Var}\left[kSx - Syk_2^2\right] + O\left(\frac{m^2}{"^2}kx - yk_2^2 + \frac{m^4}{"^4}k\right):$
3. *Given* Sx *and* Sy*, the estimate* $\hat{E}_{gen}$ *can be computed in time* $O(k)$.

Proof. We show that it su ces to let D be either the Normal or Laplace distribution for well-chosen parameters. We start with the following useful note:

Note 4.4. *Let* $n!!$ *be the product of the numbers* $1; :::; n$ *that have the same parity as* n*. For* L Lap( ) *and* G $N(0; \ ^2)$*, we have*

$$8n\ 2\ \mathbb{N}: \ \mathrm{E}_D[L^n] = \frac{n!}{(\ _1)^n} \qquad \text{and} \qquad \text{for even } n: \ \mathrm{E}_D[G^n] = (n - 1)!! \ ^n:$$

By Lemma 4.2, the noisy projection $Sx + \ '$ is $("; \ )$-di erentially private for $D = N(0; \ ^2)$ with $\frac{}{"^2}\sqrt{2\ln(1{:}25=\ )}$. By the post-processing property of di erential privacy, $\hat{E}_{gen}$ is also $("; \ )$-di erentially private. From Lemma 4.3 and Note 4.4

$$\mathrm{Var}\left[\hat{E}_{gen}\right] = \mathrm{Var}\left[kSx - Syk_2^2\right] + O\left(\frac{\frac{2}{2}\ln\left(\frac{1}{\ }\right)}{"2}kx - yk_2^2 + \frac{\frac{4}{2}\ln^2\left(\frac{1}{\ }\right)}{"4}k\right): \tag{4.1}$$

Similarly, by Lemma 4.1 $\hat{E}_{gen}$ is $"$-di erentially private for $D = \mathrm{Lap}(\ _1=")$, and from Lemma 4.3 and Note 4.4 we get

$$\mathrm{Var}\left[\hat{E}_{gen}\right] = \mathrm{Var}\left[kSx - Syk_2^2\right] + O\left(\frac{\frac{2}{1}}{"2}kx - yk_2^2 + \frac{\frac{4}{1}}{"4}k\right): \tag{4.2}$$

Finally, we can draw noise from the Laplace or the Normal distribution in constant time. □

Note 4.5. *As seen in the proof of Lemma 4.4, letting* $D = Lap(\ _1=")$ *gives* $m = \ _1$ *and letting* $D = N(0; \ ^2)$ *for* $\ _2 " \ ^1\sqrt{2\ln(1{:}25=\ )}$ *gives* $m = \ _2\sqrt{\ln(1=\ )}$*. We wish to choose the* D *which minimizes* $\mathrm{Var}[\hat{E}_{gen}]$*. disregarding constants, (4.2) is upper bounded by (4.1) when*

$$_1 < \ _2\sqrt{\ln(1=\ )} \quad , \qquad < e^{\ \frac{2}{1}=\ \frac{2}{2}}: \tag{4.3}$$

Hence, when (4.3) is satis ed, we let D *be the Laplace distribution { that is,* $D = \mathrm{Lap}(\ _1=")$ *and otherwise let* $D = N(0; \ ^2)$ *with* $\ _2 " \ ^1\sqrt{2\log(1{:}25=\ )}$.

4.5 Private Fast Johnson-Lindenstrauss Transform

We now discuss a private version of the Fast Johnson-Lindenstrauss transform (FJLT) by Ailon & Chazelle [3]. We first remind the reader of the non-private transform in Section 4.5.1 and then give two private versions in Section 4.5.2.

4.5.1 Description of (non-private) Fast Johnson-Lindenstrauss Transform (FJLT)

FJLT [3] is a random distribution of linear mappings $: \mathrm{R}^u \to \mathrm{R}^k$ with $k = O(\log(1/p)/ ^2)$, such that for $; p \in (0; 1/2)$, with probability at least $1 \quad p$

$$(1 \quad)kkxk_2^2 \quad k \ xk_2^2 \quad (1 + \ )kkxk_2^2:$$

For given values of $u; \ ; p$, we describe how to obtain the random mapping as the product of three real-valued matrices, $P; H$ and D:

- D is a random $u \quad u$-diagonal matrix with D_{jj} drawn independently from $f \ 1; +1g$ with probability 1/2.
- H is a $u \quad u$-normalized Hadamard matrix such that for $f; j \in [u]$

$$H_{fj} = \frac{1}{\sqrt{u}}(\ 1)^{hf \ 1;j \ 1i}$$

  where $hf; j i$ is the dot-product between vectors expressing f and j in binary representation.

- P is a random $k \quad u$-matrix whose entries are independently either normally distributed or 0. Specifically, for

$$q = \min\left\{ \quad \left(\frac{\log^2(1/p)}{u}\right); 1\right\}$$

  we let P_{if} be drawn (independently) from $N(0; q \ ^1)$ with probability q and $P_{if} = 0$ with probability $1 \quad q$ for $i \in [k]$ and $f \in [u]$.

The transform is defined as

$$:= PHD:$$

To formalize, we get the following lemma:

Lemma 4.5 (Lemma 2.1 from [3])**.** *Let $; p \in (0; 1/2)$ and let be a random $k \quad u$-projection matrix as described above. Let $x \in \mathrm{R}^u$. With probability at least $1 \quad p$, the following two events occur:*

- $(1 \quad)kkxk_2^2 \quad k \ xk_2^2 \quad (1 + \ )kkxk_2^2$.
- *The mapping $: \mathrm{R}^u \to \mathrm{R}^k$ requires time*

$$O\left(u \log u + uq\frac{\log(1/p)}{ ^2}\right); \quad \text{for} \quad q = \min\left\{ \quad \left(\frac{\log^2(1/p)}{u}\right); 1\right\}:$$

Proof. See [3]. □

We will henceforth concern ourselves with the *normalized* FJLT, $1/\sqrt{k}$, such that

$$(1 \quad)kxk_2^2 \quad 1/kk \ xk_2^2 \quad (1 + \ )kxk_2^2:$$

Lemma 4.6. *The normalized FJLT satisfies LPP (see Definition 4.2).*

Proof. See Section 4.9.2. □

Lemma 4.7. *Let $x; y \in \mathrm{R}^u$ and let be the FJLT as described above. Then*

$$\mathrm{Var}[1/kk \ x \quad yk_2^2] \quad \frac{3}{k}kx \quad yk_2^4:$$

Proof. The proof follows directly from Lemma 4.11 in Section 4.9.2. □

4.5.2 Private FJLT

We now argue how to construct a differentially private version of FJLT by adding Gaussian noise to the input.

If we simply exchange the i.i.d. normally distributed JL transform for FJLT in the work of Kenthapadi et al. [96], we get the following result.

Corollary 4.1. *Let be a random k u-FJLT and let x; y 2* $\mathbb{R}^u$ *be input vectors. Let* $_2$ *be the* ℓ_2*-sensitivity of and let ' ;* $N(0; \ ^2)^k$ *for* $_2$" $^1\sqrt{2\log(1.25/)}$ *be noise vectors. Define*

$$\hat{E}_{FJLT_o} := \frac{1}{k}k(\ x + ') \quad (\ y + \ )k_2^2 \quad 2k \ ^2$$

- $\hat{E}_{FJTL_o}$ *is an unbiased estimator for* $kx \quad yk_2^2$.
- *The estimator has variance*

$$\mathrm{Var}\left[\hat{E}_{FJLT_o}\right] \quad \frac{3}{k}kx \quad yk_2^4 + O\left(k \ ^4 + \ ^2kx \quad yk_2^2\right):$$

- *The sketch* $(\ ; \ x + ')$ *is* $(" ; \ )$*-differentially private.*
- *The sketch* $x + '$ *can be computed in time*

$$O\left(\max\left\{u\log u; \frac{uq\log(1/p)}{^2}\right\}\right); \qquad \text{for} \quad q = \min f \ (\log^2(1/p)/u); 1g:$$

Proof. That the estimator is unbiased and the variance follow from Lemmas 4.3, 4.6 and 4.7. Privacy follows from Lemmas 4.2 and 4.4. Running time follows from Lemmas 4.5 and 4.4. □

Note 4.6. *The* ℓ_2*-sensitivity of the (normalized) projection is concentrated around 1, which justifies the choice of Gaussian noise. Although the* ℓ_2*-sensitivity of the normalized FJLT is concentrated around 1, the sensitivity of* could *(with a small probability) be very large, so the sketch* $x + '$ *suffers from the same initialization cost as the work of Kenthapadi et al. (see Section 4.2.1).*

We now introduce a private version of FJLT, where we perturb the *input*. This version avoids the issue described in Note 4.6, but will inevitably introduce error depending on the input size.

Lemma 4.8. *Let be a random k u-FJLT and let x; y 2* $\mathbb{R}^u$ *be input vectors. Let ' ;* $N(0; \ ^2)^u$ *for* $" \ ^1\sqrt{2\log(1.25/)}$ *be noise vectors. Define*

$$\hat{E}_{FJLT_i} := \frac{1}{k}k \ (x + ') \qquad (y + \ )k_2^2 \quad 2u \ ^2$$

- $\hat{E}_{FJTL_i}$ *is an unbiased estimator for* $kx \quad yk_2^2$.
- *The estimator has variance*

$$\mathrm{Var}\left[\hat{E}_{FJLT_i}\right] \quad \frac{3}{k}kx \quad yk_2^4 + O\left(\frac{u^2 \ ^4}{k} + u \ ^2kx \quad yk_2^2\right):$$

- *The sketch* $(\ ; \ (x + '))$ *is* $(" ; \ )$*-differentially private.*
- *The sketch* $(x + ')$ *can be computed in time*

$$O\left(\max\left\{u\log u; \frac{uq\log(1/p)}{^2}\right\}\right); \qquad \text{for} \quad q = \min f \ (\log^2(1/p)/u); 1g:$$

Proof. For proofs that the estimator is unbiased and for the variance, see Section 4.9.3. We remark that the factor u on the last term in the variance is a by-product of applying to the noise. Privacy follows directly from the Gaussian mechanism (see Lemma 4.2), as the ℓ_2-sensitivity is at most 1 (as we perturb the input vectors). As noise can be added in time $O(u)$, the time required to compute the sketch follows from Lemma 4.5. □

Note 4.7. *By spherical symmetry of the Normal distribution,* $(x + \eta)$ *and* $x + P\eta$*, where* P *is de ned in Section 4.5.1, are identically distributed. Hence, one could add the same amount of noise after the Hadamard transform to get a di erentially private sketch { that is, compute* $P(HDx + \eta)$*. Thus, for a given projection* P*, suppose column* j *is all zeros, then we can immediately set* $\eta_j = 0$*. This way, we may save a bit of randomness in an application.*

4.6 Private Sparser Johnson-Lindenstrauss Transform

We now turn to the question of perturbation using Laplace noise rather than Gaussian noise. We present and analyze a private sketch based on the SJLT and conclude Theorem 4.3 in Section 4.6.2. The main observation about this sketch is that we perturb the *output* vectors rather than the input vectors, so the amount of noise depends on k rather than u while avoiding the initialization cost that was inherent to the work of Kenthapadi et al. as well as Corollary 4.1. We compare the work of Kenthapadi et al., our private FJLT from Lemma 4.8 and our private SJLT from Theorem 4.3 in Section 4.7.

Theorem 4.3 is proven by combining the technical Lemmas 4.3 and 4.4 with the SJLT. These transforms are more e cient than the suggestions from [96] due to their sparsity. We remark that this is just one example of linear transformations where our results can be applied. It should also be noted that the results of Kenthapadi et al. are transferable to the SJLT, although the results were only proven for the i.i.d. normally distributed JL transform, whereas we give the analysis here.

4.6.1 Description of (non-private) Sparser Johnson-Lindenstrauss Transforms (SJLT)

We rst describe the SJLT from [93]. We focus on the c)-construction and remark that similar arguments applies for the b)-construction. Let $k =$ ($^2 \log(1=p)$) and let $x \in \mathrm{R}^u$ be an input vector. Let $h_1, \ldots, h_s : [u] \to [k/s]$ and $\eta_1, \ldots, \eta_s : [u] \to \{-1, +1\}$ be independent, random hash functions from $O(\log(1/p))$-wise independent families. De ne $\delta_{ri}(j) = 1[h_r(j) = i]$. Then $\mathrm{E}[\delta_{ri}(j)^2] = \mathrm{E}[\delta_{ri}(j)] = \frac{s}{k}$. The projection matrix S is de ned by

$$S_{(i,r),j} = \frac{1}{\sqrt{s}} \eta_r(j)\, \delta_{ri}(j)$$

for $i = 1, \ldots, k/s$ and $r = 1, \ldots, s$. Hence, entry $i' = i + r \cdot [k]$ in the resulting embedding Sx can be described as

$$(Sx)_{i'} = (Sx)_{(i,r)} = \frac{1}{\sqrt{s}} \sum_{j=1}^{u} \eta_r(j)\, \delta_{ri}(j) x_j .$$

We can think of Sx as a vector consisting of s blocks, each of length k/s. The ith block describes the projection of x under h_i and η_i.

Lemma 4.9. *The SJLT as described above satisfy LPP from De nition 4.2.*

Proof. The proof is a simple calculation and can be found in Section 4.9.4. □

Lemma 4.10. *Let* $x, y \in \mathrm{R}^u$ *and let* S *be the SJLT as described above. Then*

$$\mathrm{Var}\left[\|Sx - Sy\|_2^2\right] \quad \frac{2}{k}\|x - y\|_2^4 .$$

Proof. The proof can be found in Section 4.9.4. □

4.6.2 Private SJLT

We are now ready to prove our main theorem, Theorem 4.3. Combining Lemmas 4.3, 4.9 and 4.10, we obtain the following corollary.

Corollary 4.2. *Let S be a random k u-SJLT and let $x; y \in \mathbb{R}^u$ be input vectors. Let $'$; D^k be noise vectors where each entry is drawn from a zero-mean distribution D. Then*

$$\hat{E}_{SJLT_D} := k(Sx + ') \quad (Sy +)k_2^2 \quad 2k\,\mathrm{E}_D[{'}^2]$$

is an unbiased estimator for $kx \quad yk_2^2$ with variance

$$\mathrm{Var}\left[\hat{E}_{SJLT_D}\right] \quad \frac{2}{k}kx \quad yk_2^4 + 8\,\mathrm{E}_D\left[{'}^2\right]kx \quad yk_2^2 + 2k\left(\mathrm{E}_D[{'}^4] + \mathrm{E}_D\left[{'}^2\right]^2\right): \tag{4.4}$$

We have yet to choose D to ensure di erential privacy of this estimator as well as argue about the e ciency. We start by discussing the optimal value of the output dimension, k.

Optimal Projection Dimension

For the non-private SJLT, the optimal projection dimension is $k = \left(\ ^2 \log(1=p)\right)$. One may ask what k is optimal in the private case. The analysis and our optimal k are very similar to the ndings in [96]: we see that the variance in (4.4) is minimized for $k = \left(p\frac{kx \quad yk_2^2}{\mathrm{E}[{'}^4]+\mathrm{E}[{'}^2]^2}\right)$. By the same argument as in [96], generally, no xed value of k will be optimal for the entire input domain, although there might be exceptions, when certain properties of the data are known. As in the work of Kenthapadi et al., if we have input domain X, then we may let $= \max_{x2X} fkxk_2^2g$ to obtain $k = (\ ''^2 = \ _1^2)$ for $D = \mathrm{Lap}(\ _1='')$. Note that k might not be optimal for all input vectors. We assume that is unknown and may be very large { in particular, we consider vectors over the reals { and thus proceed with $k = (\ ^2 \log(1=p))$.

Efficiency

Let S be a SJLT with sparsity $s = O\left(\ ^1 \log(1=p)\right)$ and let input $x \in \mathbb{R}^u$ be given. The embedding Sx can be computed in time $O(skxk_0)$. Assuming that we sample from $N(0; \ ^2)$ and $\mathrm{Lap}(\ )$ in constant time, random noise vector $' \quad D$ for $D = \mathrm{Lap}(\ _1='')$ or $D = N(0; \ ^2)$ for $\ _2'' \ ^1\sqrt{2\log(1:25= \ )}$ can be added in time $O(k)$ to give $Sx + '$. From [47] we know that we can at least sample from discretizations in expected constant time, so this assumption seems reasonable. For given $Sx + '$ and $Sy + \ $, the estimator $\hat{E}_{SJLT}$ can be computed in time $O(k)$.

Putting Everything Together

The SJLT as described in Section 4.6.1, where $k = (\ ^2 \log(1=p))$ and $s = O(\ ^1 \log(1=p))$, has $`_1$-sensitivity $\ _1 = \sqrt{s}$ and $`_2$-sensitivity $\ _2 = 1$. Hence, consider Corollary 4.2 with $D = \mathrm{Lap}(\sqrt{s}='')$. Lemma 4.1 ensures that $\hat{E}_{SJLT}$ is $''$-di erentially private. Combining with Section 4.6.2 nishes the proof of Theorem 4.3. If instead we let $D = N(0; \ ^2)$ for $\ '' \ ^1\sqrt{2\ln(1:25= \ )}$ in Corollary 4.2, $\hat{E}_{SJLT}$ is $(''; \ )$-di erentially private and achieves the same variance as the work of Kenthapadi et al., while we gain a speed-up as well as avoid the initialization cost. Finally, we remark that by Note 4.5, we minimize the variance of $\hat{E}_{SJLT}$ by letting $D = \mathrm{Lap}(\sqrt{s}='')$ whenever $< e \ ^s$.

4.7 Comparison

We now compare Lemma 4.8 and Theorem 4.3 with the work of Kenthapadi et al. We rst compare the running times to see for what parameters the private FJLT is faster than the private SJLT and then compare the variances for the two methods to get the speed-variance trade-o . Finally, we compare with the results of Kenthapadi et al.

Recall that our private FJLT can be computed in time

$$O\left(\max\left\{u\log u;\frac{\log^3(1=p)}{2}\right\}\right);$$

and the private SJLT can be computed in time bounded by $O(su)$ (for dense vectors) where $s = O\left(\log(1=p)\quad^{1}\right)$. Observing that

$$O(su) > O(u\log u)\qquad,\qquad u < e^{O(s)} = \frac{1}{p^{O(1=\ )}}$$

and

$$O(su) > O\left(\frac{\log^3(1=p)}{2}\right)\qquad,\qquad u > O\left(\frac{\log^2(1=p)}{}\right);$$

we conclude that our private FJLT is indeed faster than the private SJLT whenever

$$O\left(\frac{\log^2(1=p)}{}\right) < u < \frac{1}{p^{O(1=\ )}}: \tag{4.5}$$

We now turn to comparing the variances of the private versions of FJLT and SJLT: Recall from Lemma 4.8 that the private FJLT has variance

$$\mathrm{Var}\left[\hat{E}_{FJLT_i}\right]\quad\frac{3}{k}kx\quad yk_2^4 + O\left(u\ ^2kx\quad yk_2^2 + \frac{u^2\ \ ^4}{k}\right):$$

while, as seen in Theorem 4.3, the private SJLT has variance

$$\mathrm{Var}\left[\hat{E}_{SJLT}\right]\quad\frac{2}{k}kx\quad yk_2^4 + O\left(\frac{s}{\prime\prime_2}kx\quad yk_2^2 + \frac{s^2}{\prime\prime^4}k\right):$$

For simplicity, we will disregard the variance incurred by the transforms and limit ourselves to considering the terms incurred by the noise addition. The private SJLT, in particular, achieves a better variance than the private FJLT whenever each of the error terms incurred by the noise is bounded by the corresponding error term from the FJLT. That is:

$$O\left(\frac{u^2\ \ ^4}{k}\right) = O\left(\frac{u^2\log^2(1=\ )}{\prime\prime^4 k}\right) > O\left(\frac{s^2k}{\prime\prime^4}\right)\qquad\text{and}$$
$$O\left(u\ ^2kx\quad yk_2^2\right) = O\left(\frac{u\log(1=\ )}{\prime\prime_2}kx\quad yk_2^2\right) > O\left(\frac{s}{\prime\prime_2}kx\quad yk_2^2\right):$$

Treating each of the inequalities separately, we analyze for what values of this is the case:

$$O\left(\frac{u^2\log^2(1=\ )}{\prime\prime^4 k}\right) > O\left(\frac{s^2k}{\prime\prime^4}\right)\qquad,\qquad\log(1=\ ) > O\left(\frac{sk}{u}\right)\qquad,\qquad\frac{1}{e^{O(sk=u)}} >$$

and

$$O\left(\frac{u\log(1=\ )}{\prime\prime_2}kx\quad yk_2^2\right) > O\left(\frac{s}{\prime\prime_2}kx\quad yk_2^2\right)\qquad,\qquad\log(1=\ ) > O\left(\frac{s}{u}\right)\qquad,\qquad\frac{1}{e^{O(s=u)}} > :$$

Hence, in particular, the private SJLT has smaller variance than the private FJLT when

$$< \min\left\{1=e^{O(s=u)};1=e^{O(sk=u)}\right\} = 1=e^{O(sk=u)} = 1=e^{O\ \ \frac{\log^2(1=p)}{3u}} = p^{O\ \ \frac{\log(1=p)}{3u}}:$$

The variance of the estimator from Theorem 4.2 by Kenthapadi et al. was

$$\mathrm{Var}[\hat{E}_{iid}] = \frac{2}{k}\|x - y\|_2^4 + O\left(\frac{1}{}\|x - y\|_2^2 + \frac{1}{^4 k}\right):$$

An argument similar to the one above proves that the variance of our private SJLT improves over the variance of Kenthapadi et al. when $< e^{-s} = p^{O(1=)}$. Clearly, Kenthapadi et al. always achieve better variance than our private FJLT due to the dependence on u, which was inherent from perturbing the input rather than the output, and we may assume $k < u$.

Hence, we see a trade-o in running time versus variance for certain values of input dimension u.

To sum up the above discussion, suppose that $< p^{O(1=)}$. Then the private SJLT obtains the best variance out of all the methods. If u satis es (4.5), then the private FJLT achieves the best running time, and otherwise, the private SJLT improves over the private FJLT in terms of both variance and running time.

4.8 Open Problems

One interesting question is why we get the separation between Laplace and Gaussian noise depending on the size of : if is su ciently small, then Laplace noise is preferable over Gaussian noise. This observation suggests that there exists a di erentially private mechanism, which adds optimal noise for the entire parameter space, i.e., a mechanism which adds noise from a distribution achieving error at most that incurred by Laplace and Gaussian noise (simultaneously) for all and all " < 1.

4.9 Technical Details

In this section, we give the technical details and proofs omitted in the previous sections.

4.9.1 Omitted Proofs for Supporting Lemmas

Lemma 4.3. *We have*

1. $\hat{E}_{gen}$ *is an unbiased estimator for* $\|x - y\|_2^2$.
2. *The variance of* $\hat{E}_{gen}$ *is*

$$\mathrm{Var}\left[\hat{E}_{gen}\right] = \mathrm{Var}\left[\|Sx - Sy\|_2^2\right] + 8\,\mathrm{E}_D[\,'^2]\|x - y\|_2^2 + 2k\,\mathrm{E}_D[\,'^4] + 2k\,\mathrm{E}_D[\,'^2]^2$$

Proof. We start by showing 1). For simpler notation, we de ne $z := x - y$. By independence and since $\mathrm{E}_D[\,'_i] = 0$ for all i,

$$\begin{aligned}
\mathrm{E}_{S;D}\left[\left\|(Sx + \,') - (Sy + \,)\right\|_2^2\right] &= \mathrm{E}_{S;D}\left[\sum_{i=1}^{k}((Sx + \,')_i - (Sy + \,)_i)^2\right] \\
&= \mathrm{E}_{S;D}\left[\sum_{i=1}^{k}\left((Sz)_i^2 + (\,'_i - \,_i)^2 + 2(\,'_i - \,_i)(Sz)_i\right)\right] \\
&= \mathrm{E}_S\left[\sum_{i=1}^{k}(Sz)_i^2\right] + 2\sum_{i=1}^{k}\mathrm{E}_D\left[\,'^2\right] = k\|z\|_2^2 + 2k\,\mathrm{E}_D\left[\,'^2\right]
\end{aligned}$$

where we in the last step used that S has the LPP. So clearly, re-inserting $z = x - y$

$$\mathrm{E}_{S;D}\left[\left\|(Sx + \,') - (Sy + \,)\right\|_2^2 - 2k\,\mathrm{E}_D\left[\,'^2\right]\right] = \|x - y\|_2^2:$$

We turn to proving 2):

$$\mathrm{Var}\left[\hat{E}_{gen}\right] = \mathrm{E}_{S;D}\left[\hat{E}^2_{gen}\right] \quad \mathrm{E}_{S;D}\left[\hat{E}_{gen}\right]^2 = \mathrm{E}_{S;D}\left[\hat{E}^2_{gen}\right] \quad kx \quad yk_2^4; \tag{4.6}$$

so we analyze the rst term:

$$\begin{aligned}\mathrm{E}_{S;D}\left[\hat{E}^2_{gen}\right] &= \mathrm{E}_{S;D}\left[\left(\|(Sx + \ ') \quad (Sy + \ )\|_2^2 \quad 2k\,\mathrm{E}_D\left[\ '^2\right]\right)^2\right] \\ &= \mathrm{E}_{S;D}\left[\|(Sx + \ ') \quad (Sy + \ )\|_2^4\right] + 4k^2\,\mathrm{E}_D\left[\ '^2\right]^2 \\ &\quad 4k\,\mathrm{E}_D\left[\ '^2\right]\mathrm{E}_{S;D}\left[\|(Sx + \ ') \quad (Sy + \ )\|_2^2\right]\end{aligned} \tag{4.7}$$

The last term in (4.7) equals

$$4k\,\mathrm{E}_D\left[\ '^2\right]\left(kx \quad yk_2^2 + 2k\,\mathrm{E}_D\left[\ '^2\right]\right) = 4k\,\mathrm{E}_D\left[\ '^2\right]kx \quad yk_2^2 + 8k^2\,\mathrm{E}_D\left[\ '^2\right]^2 \tag{4.8}$$

Claim 4.1. *The rst term in (4.7) equals*

$$\begin{aligned}\mathrm{E}_{S;D}\left[\|(Sx + \ ') \quad (Sy + \ )\|_2^4\right] = \mathrm{E}_S\left[kS(x \quad y)k_2^4\right] + 4(k+2)\,\mathrm{E}_D\left[\ '^2\right]kx \quad yk_2^2 + 2k\,\mathrm{E}_D[\ '^4] \\ + 2k(1+2k)\,\mathrm{E}_D\left[\ '^2\right]^2\end{aligned}$$

The proof of the claim is straightforward but tedious and thus left out here. It is proven formally at the end of this section.

Inserting (4.8) and Claim 4.1 into (4.7), we get

$$\begin{aligned}\mathrm{E}_{S;D}\left[\hat{E}^2_{gen}\right] &= \mathrm{E}_{S;D}\left[\left(\|(Sx + \ ') \quad (Sy + \ )\|_2^2 \quad 2k\,\mathrm{E}_D\left[\ '^2\right]\right)^2\right] \\ &= \mathrm{E}_S\left[kS(x \quad y)k_2^4\right] + 4(k+2)\,\mathrm{E}_D\left[\ '^2\right]kx \quad yk_2^2 + 2k\,\mathrm{E}_D[\ '^4] + 2k(1+2k)\,\mathrm{E}_D\left[\ '^2\right]^2 \\ &\quad + 4k^2\,\mathrm{E}_D\left[\ '^2\right]^2 \quad 4k\,\mathrm{E}_D\left[\ '^2\right]kx \quad yk_2^2 \quad 8k^2\,\mathrm{E}_D\left[\ '^2\right]^2 \\ &= \mathrm{E}_S\left[kS(x \quad y)k_2^4\right] + 8\,\mathrm{E}_D\left[\ '^2\right]kx \quad yk_2^2 + 2k\,\mathrm{E}_D[\ '^4] + 2k\,\mathrm{E}_D\left[\ '^2\right]^2:\end{aligned}$$

Inserting this expression into (4.6) proves that the variance is

$$\begin{aligned}\mathrm{Var}\left[\hat{E}_{gen}\right] &= \mathrm{E}_S\left[kS(x \quad y)k_2^4\right] + 8\,\mathrm{E}_D\left[\ '^2\right]kx \quad yk_2^2 + 2k\,\mathrm{E}_D[\ '^4] + 2k\,\mathrm{E}_D\left[\ '^2\right]^2 \quad kx \quad yk_2^4 \\ &= \mathrm{Var}\left[kS(x \quad y)k_2^2\right] + 8\,\mathrm{E}_D\left[\ '^2\right]kx \quad yk_2^2 + 2k\,\mathrm{E}_D[\ '^4] + 2k\,\mathrm{E}_D\left[\ '^2\right]^2;\end{aligned}$$

again using that S satis es LPP. □

Proof of Claim 4.1

Proof. For a simpler notation, we de ne $z := x \quad y$. By simply unfolding the expression, we see that

$$\begin{aligned}
\mathrm{E}_{S;D}\left[\left\|(Sx + ') \quad (Sy + \quad)\right\|_2^4\right] &= \sum_{i;`=1}^{k} \mathrm{E}_S\left[(Sz)_i^2(Sz)^2\right] + \sum_{i;`=1}^{k}\left(\mathrm{E}_D\left[{}'^2_i\right] + \mathrm{E}_D\left[\ {}^2_i\right]\right)\mathrm{E}_S\left[(Sz)^2\right] + 0 \\
&\quad + \sum_{i;`=1}^{k} \mathrm{E}_S\left[(Sz)_i^2\right]\left(\mathrm{E}_D\left[{}'^2\right] + \mathrm{E}_D\left[\ {}^2\right]\right) + \sum_{i=1}^{k}\mathrm{E}_D\left[('_i \quad _i)^4\right] \\
&\quad + \sum_{i\neq `}\mathrm{E}_D\left[('_i \quad _i)^2\right]\mathrm{E}_D\left[('_` \quad _`)^2\right] \\
&\quad + 2\sum_{i=1}^{k}\mathrm{E}_S\left[(Sz)_i\right]\mathrm{E}_D\left[('_i \quad _i)^3\right] \\
&\quad + 0 + 2\sum_{`=1}^{k}\mathrm{E}_D\left[('_` \quad _`)^3\right]\mathrm{E}_S\left[(Sz)_`\right] \\
&\quad + 4\sum_{i=1}^{k}\mathrm{E}_D\left[('_i \quad _i)^2\right]\mathrm{E}_S\left[(Sz)_i^2\right] \\
&\quad + 4\sum_{i\neq `}\underbrace{\mathrm{E}_D\left[('_i \quad _i)('_` \quad _`)\right]}_{=0}\mathrm{E}_S\left[(Sz)_i(Sz)_`\right]
\end{aligned}$$

where we used that $\mathrm{E}['_i] = \mathrm{E}[\ _i] = 0$ for all $i = 1;:::;k$ and that the noise is drawn independently of S.

Recalling that $\mathrm{E}_D['^2_i] = \mathrm{E}_D[\ ^2_i] = \mathrm{E}_D['^2]$ for all i, we obtain

$$\begin{aligned}
\mathrm{E}_{S;D}\left[\left\|(Sx + ') \quad (Sy + \quad)\right\|_2^4\right] &= \sum_{i;`=1}^{k}\mathrm{E}_S\left[(Sz)_i^2(Sz)^2\right] + 4k\,\mathrm{E}_D\left['^2\right]kzk_2^2 + k\left(2\,\mathrm{E}_D['^4] + 6\,\mathrm{E}_D['^2]^2\right) \\
&\quad + \sum_{i\neq `}4\,\mathrm{E}_D\left['^2\right]^2 + 2\sum_{i=1}^{k}\mathrm{E}_S\left[(Sz)_i\right]\left(\mathrm{E}_D\left['^3\right] \quad \mathrm{E}_D\left['^3\right]\right) \\
&\quad + 2\sum_{`=1}^{k}\mathrm{E}_S\left[(Sz)_`\right]\left(\mathrm{E}_D\left['^3\right] \quad \mathrm{E}_D\left['^3\right]\right) + 4\sum_{i=1}^{k}2\,\mathrm{E}_D\left['^2\right]\mathrm{E}_S\left[(Sz)_i^2\right]
\end{aligned}$$

which simpli es to

$$\begin{aligned}
&= \mathrm{E}_S\left[kSzk_2^4\right] + 4k\,\mathrm{E}_D\left['^2\right]kzk_2^2 + 2k\,\mathrm{E}_D['^4] + 6k\,\mathrm{E}_D['^2]^2 + 4(k^2 \quad k)\,\mathrm{E}_D\left['^2\right]^2 + 8\,\mathrm{E}_D\left['^2\right]kzk_2^2 \\
&= \mathrm{E}_S\left[kSzk_2^4\right] + 4(k+2)\,\mathrm{E}_D\left['^2\right]kzk_2^2 + 2k\,\mathrm{E}_D['^4] + 2k(1+2k)\,\mathrm{E}_D\left['^2\right]^2
\end{aligned}$$

Re-inserting $z = x \quad y$, we conclude that

$$\begin{aligned}
\mathrm{E}_{S;D}\left[\left\|(Sx + ') \quad (Sy + \quad)\right\|_2^4\right] &= \mathrm{E}_S\left[kS(x \quad y)k_2^4\right] + 4(k+2)\,\mathrm{E}_D\left['^2\right]kx \quad yk_2^2 + 2k\,\mathrm{E}_D['^4] \\
&\quad + 2k(1+2k)\,\mathrm{E}_D\left['^2\right]^2 :
\end{aligned}$$

□

4.9.2 Omitted Proofs for FJLT

Primitives

This section will give some primitives that will be useful in the next section. Non-trivial arguments can be found in Section 4.9.2. We let $= PHD$ be the FJLT transform as described in Section 4.5.1 and $x; y \ 2 \ \mathrm{R}^u$

be any real vectors. Let $X \quad N(0; q^{-1})$. Then for any $i; n \in [k]$ and $j; \ell \in [u]$

$$\begin{aligned}
&\mathrm{E}_P[P_{ij}] = 0 \\
&\mathrm{E}_P\left[P_{ij}^2\right] = q \quad \mathrm{E}_X\left[X^2\right] = 1 \\
&\mathrm{E}_P\left[P_{ij}^4\right] = q \quad \mathrm{E}_X\left[X^4\right] = q \quad 3q^{\ 2} = \frac{3}{q} \\
&\mathrm{E}_D[D_{jj}] = 0 \\
&\mathrm{E}_D[D_{jj}^2] = D_{jj}^2 = 1 \\
&\mathrm{E}[\ _{ij}] = \sum_{f=1}^{u} \mathrm{E}_P[P_{if}]\, H_{fj}\, \mathrm{E}_D[D_{jj}] = 0
\end{aligned}$$

$$\mathrm{E}\left[\ _{ij}\ _{n\ell}\right] = \begin{cases} \mathrm{E}\left[\ _{ij}\right]\mathrm{E}\left[\ _{n\ell}\right] = 0; & i \neq n;\ j \neq \ell \\ \mathrm{E}_P[(PH)_{ij}]\,\mathrm{E}_P[(PH)_{nj}] = 0; & i \neq n;\ j = \ell \\ \mathrm{E}\left[\ _{ij}\right]\mathrm{E}\left[\ _{i\ell}\right] = 0; & i = n;\ j \neq \ell \\ \mathrm{E}\left[\ _{ij}^{\ 2}\right] = 1; & i = n;\ j = \ell \end{cases}$$

$$\mathrm{E}\left[\ _{ij}^{\ 2}\ _{n\ell}^{\ 2}\right] = \begin{cases} 1; & i \neq n \\ \frac{3}{qu} + 1 \quad \frac{3}{u}; & i = n;\ j \neq \ell \\ \frac{3}{qu} + 3 \quad \frac{3}{u}; & i = n;\ j = \ell \end{cases} \tag{4.9}$$

$$\mathrm{E}\left[\ _{ij}\ _{i\ell}\ _{nv}\ _{nw}\right] = \begin{cases} \mathrm{E}\left[\ _{ij}^{\ 2}\ _{nv}^{\ 2}\right]; & j = \ell;\ v = w \\ \mathrm{E}\left[\ _{ij}^{\ 2}\ _{i\ell}^{\ 2}\right]; & i = n;\ (j = v;\ \ell = w) \_ (j = w;\ \ell = v) \\ 0; & \text{otherwise} \end{cases}$$

$$\mathrm{E}\left[(\ x)_i(\ y)_n\right] = \sum_{j,\ell=1}^{u} x_j y_\ell\, \mathrm{E}\left[\ _{ij}\ _{n\ell}\right] = \begin{cases} 0; & i \neq n \\ \sum_{j=1}^{u} x_j y_j; & i = n \end{cases}$$

$$\mathrm{E}\left[(\ x)_i^2(\ y)_n^2\right] = kxk_2^2 kyk_2^2; \qquad i \neq n \tag{4.10}$$

$$\mathrm{E}\left[(\ x)_i^2(\ y)_i^2\right] = \frac{3}{u}\left(\frac{u}{3} + \left(\frac{1}{q} \quad 1\right)\right)\left(kxk_2^2 kyk_2^2 + 2hx; yi^2\right) \quad \frac{6}{u}\left(\frac{1}{q} \quad 1\right)\sum_{j=1}^{u} x_j^2 y_j^2 \tag{4.11}$$

Arguments for the Primitives Stated Above:

Argument for Equation (4.9)

We use that

$$\mathrm{E}\left[\ _{ij}^{\ 2}\ _{n\ell}^{\ 2}\right] = \sum_{f;g;h;s=1}^{u} \mathrm{E}_P[P_{if}P_{ig}P_{nh}P_{ns}]H_{fj}H_{gj}H_{h\ell}H_{s\ell}\,\mathrm{E}_D[D_{jj}^2 D_{\ell\ell}^2]$$

and so for $i \neq n$

$$\sum_{f;h=1}^{u} \mathrm{E}_P\left[P_{if}^2\right]\mathrm{E}_P\left[P_{nh}^2\right] H_{fj}^2 H_{h\ell}^2 = 1$$

and for $i = n$

$$\begin{aligned}\sum_{f,g,h,s=1}^{u} \mathrm{E}_P[P_{if}P_{ig}P_{nh}P_{ns}]H_{fj}H_{gj}H_{h\ell}H_{s\ell} &= \sum_{f=1}^{u}\mathrm{E}_P[P_{if}^4]H_{fj}^2H_{f\ell}^2 + \sum_{f\neq h=1}^{u}\mathrm{E}_P[P_{if}^2]\mathrm{E}_P[P_{ih}^2]H_{fj}^2H_{h\ell}^2 \\ &\quad + 2\sum_{f\neq g,\ell=1}^{u}\mathrm{E}_P[P_{if}^2]\mathrm{E}_P[P_{ig}^2]H_{fj}H_{gj}H_{f\ell}H_{g\ell} \\ &= \mathrm{E}_P[P_{if}^4]/u + (u^2 - u)/u^2 + 2(\langle H_j, H_\ell\rangle^2 - 1/u) \\ &= \mathrm{E}_P[P_{if}^4]/u + 1 - 3/u + 2\langle H_j, H_\ell\rangle^2 \\ &= \begin{cases}\mathrm{E}_P[P_{if}^4]/u + 1 - 3/u, & j \neq \ell \\ \mathrm{E}_P[P_{if}^4]/u + 3 - 3/u, & j = \ell\end{cases}\end{aligned}$$

because $\langle H_j, H_\ell\rangle = \begin{cases}0, & j \neq \ell \\ 1, & j = \ell.\end{cases}$

Argument for Equations (4.10) and (4.11)

We used that

$$\mathrm{E}\left[(Dx)_i^2(Dy)_n^2\right] = \sum_{j,v=1}^{u} x_j^2y_v^2\,\mathrm{E}\left[\sigma_{ij}^2\sigma_{nv}^2\right] + 2\sum_{j\neq\ell=1}^{u} x_j^2y_\ell^2\,\mathrm{E}\left[\sigma_{ij}\sigma_{i\ell}\sigma_{nj}\sigma_{n\ell}\right]$$

where we for $i \neq n$ get $\|x\|_2^2\|y\|_2^2$ and for $i = n$ get

$$\begin{aligned}&\sum_{j=1}^{u} x_j^2y_j^2\,\mathrm{E}\left[\sigma_{ij}^4\right] + \sum_{j\neq v=1}^{u} x_j^2y_v^2\,\mathrm{E}\left[\sigma_{ij}^2\sigma_{iv}^2\right] + 2\sum_{j\neq\ell=1}^{u} x_jx_\ell y_jy_\ell\,\mathrm{E}\left[\sigma_{ij}^2\sigma_{i\ell}^2\right] \\ &= \sum_{j=1}^{u} x_j^2y_j^2\left(\frac{3}{qu} + 3 - \frac{3}{u}\right) + \sum_{j\neq v=1}^{u} x_j^2y_v^2\left(\frac{3}{qu} + 1 - \frac{3}{u}\right) + 2\sum_{j\neq\ell=1}^{u} x_jx_\ell y_jy_\ell\left(\frac{3}{qu} + 1 - \frac{3}{u}\right) \\ &= \sum_{j=1}^{u} x_j^2y_j^2\left(\frac{3}{qu} + 3 - \frac{3}{u}\right) + \left(\|x\|_2^2\|y\|_2^2 - \sum_{j=1}^{u} x_j^2y_j^2\right)\left(\frac{3}{qu} + 1 - \frac{3}{u}\right) \\ &\qquad + 2\left(\langle x, y\rangle^2 - \sum_{j=1}^{u} x_j^2y_j^2\right)\left(\frac{3}{qu} + 1 - \frac{3}{u}\right) \\ &= \frac{3}{u}\left(\frac{u}{3} + \left(\frac{1}{q} - 1\right)\right)\left(\|x\|_2^2\|y\|_2^2 + 2\langle x, y\rangle^2\right) - \frac{6}{u}\left(\frac{1}{q} - 1\right)\sum_{j=1}^{u} x_j^2y_j^2\end{aligned}$$

Proof of FJLT Satisfying LPP

Lemma 4.6. *The normalized FJLT satisfies LPP (see Definition 4.2).*

Proof. Applying the primitives from above, we get

$$\mathrm{E}\left[\frac{1}{k}\|Dx\|_2^2\right] = \frac{1}{k}\mathrm{E}\left[\sum_{i=1}^{k}(Dx)_i^2\right] = \frac{1}{k}\sum_{i=1}^{k}\sum_{j,\ell=1}^{u}\mathrm{E}\left[\sigma_{ij}\sigma_{i\ell}\right]x_jx_\ell = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{u}\mathrm{E}\left[\sigma_{ij}^2\right]x_j^2 = \|x\|_2^2.$$

□

Variance under FJLT

For convenience, we prove the following result, as it will be useful in this form for several other proofs. Note that Lemma 4.7 follows directly from Lemma 4.11.

Lemma 4.11. *Let k u-matrix $= PHD$, where P_{ij} is $N(0; q^{-1})$ with probability q and and 0 otherwise. For input vector $'$ D^u for a real-valued distribution D:*

$$\mathrm{Var}[k \quad 'k_2^2] \quad \frac{3}{k}\mathrm{E}_{,'}\left[k\,'k_2^2\right]:$$

For x 2 R^u, we get

$$\mathrm{Var}[k \quad xk_2^2] \quad \frac{3}{k}kxk_2^2:$$

Proof.

$$\begin{aligned}
\mathrm{Var}[k \quad 'k_2^2] &= \mathrm{E}_{;'}[k \quad 'k_2^4] \quad \mathrm{E}_{;'}[k \quad 'k_2^2]^2 = \sum_{i;n=1}^{k} \mathrm{E}_{;'}\left[(\quad ')_i^2(\quad ')_n^2\right] \quad k^2\,\mathrm{E}_{,'}\left[k\,'k_2^4\right] \\
&= \sum_{i=1}^{k}\mathrm{E}_{;'}\left[(\quad ')_i^4\right] + \sum_{i\in n=1}^{k}\mathrm{E}\left[(\quad ')_i^2(\quad ')_n^2\right] \quad k^2\,\mathrm{E}_{,'}\left[k\,'k_2^4\right] \\
&= \frac{9k}{u}\left(\frac{u}{3}+\left(\frac{1}{q} \quad 1\right)\right)\mathrm{E}_{,'}\left[k\,'k_2^4\right] \quad \frac{6k}{u}\left(\frac{1}{q} \quad 1\right)\mathrm{E}_{,'}\left[k\,'k_4^4\right] \quad k^2\,\mathrm{E}_{,'}\left[k\,'k_2^4\right] \\
&= 3k\left(\frac{2}{3}+\frac{3}{u}\left(\frac{1}{q} \quad 1\right)\right)\mathrm{E}_{,'}\left[k\,'k_2^4\right] \quad \frac{6k}{u}\left(\frac{1}{q} \quad 1\right)\mathrm{E}_{,'}\left[k\,'k_4^4\right]:
\end{aligned}$$

which again implies

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{k}k \quad 'k_2^2\right] = \frac{3}{k}\left(\frac{2}{3}+\frac{3}{u}\left(\frac{1}{q} \quad 1\right)\right)\mathrm{E}_{,'}\left[k\,'k_2^4\right] \quad \frac{6}{uk}\left(\frac{1}{q} \quad 1\right)\mathrm{E}_{,'}\left[k\,'k_4^4\right] \\
\frac{3\,\mathrm{E}_{,'}\left[k\,'k_2^4\right]}{k}\left(\frac{2}{3}+\frac{1}{u}\left(\frac{1}{q} \quad 1\right)\right) \quad \frac{3}{k}\mathrm{E}_{,'}\left[k\,'k_2^4\right]
\end{aligned}$$

when q $\frac{1}{u=3+1}$. □

4.9.3 Omitted Proofs for Private FJLT

Estimator and Variance for Private FJLT

Lemma 4.12. *We have*

1. $\hat{E}_{FJLT_i}$ *is an unbiased estimator for* $kx \quad yk_2^2$.
2. $\mathrm{Var}[\hat{E}_{FJLT_i}] \quad \frac{3}{k}kx \quad yk_2^4 + O\left(\frac{u^2 \ ^4}{k} + u \ ^2kx \quad yk_2^2\right)$.

Proof. We repeatedly apply the primitives of Section 4.9.2 and Section 4.9.2.

We start by proving 1). Observe that

$$\begin{aligned}
\mathrm{E}\left[k \ (x+') \quad (y+ \ )k_2^2\right] &= \mathrm{E}\left[k \ (x \quad y)+ \ ('\quad \ )k_2^2\right] = \mathrm{E}\left[k \ (x \quad y)k_2^2\right] + \mathrm{E}\left[k \ ('\quad \ )k_2^2\right] \\
&= kkx \quad yk_2^2 + k\,\mathrm{E}_{';\ }\left[k'\quad k_2^2\right]
\end{aligned}$$

Since $';\quad N(0;\ ^2)^u$, we have $'\quad N(0;2\ ^2)^u$ and so

$$\mathrm{E}_{,'}\left[k'\quad k_2^2\right] = \sum_{j=1}^{u}\mathrm{E}_{';\ }[('_j\quad _j)^2] = 2u\ ^2:$$

We conclude that

$$\hat{E}_{FJLT_i} = 1/kk\ (x + ') \quad (y +)k_2^2 \quad 2u\ ^2$$

is an unbiased estimator for $kx \quad yk_2^2$.

We turn to proving 2). Note that

$$\mathrm{Var}\left[1/kk\ (x \quad y) + \ (' \quad)k_2^2 \quad 2u\ ^2\right] = \frac{1}{k^2}\mathrm{Var}\left[k\ (x \quad y) + \ (' \quad)k_2^2\right];$$

so it su ces to consider the RHS. For readability, we will do the analysis for x and $'$, and eventually substitute x for $x \quad y$ and $'$ for $' \quad$, recalling that if $' \quad N(0;\ ^2)$, then $' \quad N(0; 2\ ^2)$.

For any $x; ' \ 2\ \mathrm{R}^u$

$$\begin{aligned}\mathrm{E}_{;'}\left[k\ x + \quad 'k_2^2\right]^2 &= \left(\mathrm{E}\ \left[k\ xk_2^2\right] + \mathrm{E}\ \left[k\ 'k_2^2\right]\right)^2 \\ &= \mathrm{E}\ \left[k\ xk_2^2\right]^2 + \mathrm{E}\ \left[k\ 'k_2^2\right]^2 + 2\,\mathrm{E}\ \left[k\ xk_2^2\right]\mathrm{E}\ \left[k\ 'k_2^2\right]\end{aligned}$$

By the triangle inequality, we see that

$$\begin{aligned}\mathrm{E}_{;'}\left[k\ x + \quad 'k_2^4\right] &= \mathrm{E}_{;'}\left[\left(k\ x + \quad 'k_2^2\right)^2\right] \quad \mathrm{E}_{;'}\left[\left(k\ xk_2^2 + k\ 'k_2^2 + 2k\ xk_2k\ 'k_2\right)^2\right] \\ &= \mathrm{E}\ \left[k\ xk_2^4\right] + \mathrm{E}_{;'}\left[k\ 'k_2^4\right] + 6\,\mathrm{E}_{;'}\left[k\ xk_2^2k\ 'k_2^2\right]\end{aligned}$$

Where the last equality follows from the zero-meaned $'$ leading to several terms canceling out.

Hence, the variance is bounded by

$$\begin{aligned}\mathrm{Var}\left[k\ x + \quad 'k_2^2\right] \quad \mathrm{E}\ \left[k\ xk_2^4\right] + \mathrm{E}_{;'}\left[k\ 'k_2^4\right] + 6\,\mathrm{E}_{;'}\left[k\ xk_2^2k\ 'k_2^2\right] \quad \mathrm{E}\ \left[k\ xk_2^2\right]^2 \quad \mathrm{E}_{;'}\left[k\ 'k_2^2\right]^2 \\ 2\,\mathrm{E}\ \left[k\ xk_2^2\right]\mathrm{E}_{;'}\left[k\ 'k_2^2\right]\end{aligned}$$

which again implies

$$\begin{aligned}\mathrm{Var}\left[1/kk\ x + \quad 'k_2^2\right] \quad \mathrm{Var}\ \left[1/kk\ xk_2^2\right] + \mathrm{Var}_{;'}\left[1/kk\ 'k_2^2\right] + \frac{6}{k^2}\mathrm{E}_{;'}\left[k\ xk_2^2k\ 'k_2^2\right] \\ \frac{2}{k^2}\mathrm{E}\ \left[k\ xk_2^2\right]\mathrm{E}_{;'}\left[k\ 'k_2^2\right]\end{aligned} \tag{4.12}$$

For the last term we have

$$\mathrm{E}\ \left[k\ xk_2^2\right]\mathrm{E}_{;'}\left[k\ 'k_2^2\right] = 2k^2kxk_2^2\,\mathrm{E}_{'}[k\,'k_2^2] = 2k^2u\ ^2kxk_2^2$$

and for the second to last term, we get:

$$\begin{aligned}\mathrm{E}_{;'}\left[k\ xk_2^2k\ 'k_2^2\right] &= \sum_{i;n=1}^{k}\mathrm{E}_{;'}[(\ x)_i^2(\ ')_n^2] = \sum_{i=1}^{k}\mathrm{E}_{;'}[(\ x)_i^2(\ ')_i^2] + \sum_{i\in n=1}^{k}\mathrm{E}_{;'}[(\ x)_i^2(\ ')_n^2] \\ &= \frac{3k}{u}\left(\frac{u}{3} \quad \left(1 \quad \frac{1}{q}\right)\right)\left(kxk_2^2\,\mathrm{E}\left[k\,'k_2^2\right] + 2\,\mathrm{E}\left[hx;\ 'i^2\right]\right) + \frac{6k}{u}\left(1 \quad \frac{1}{q}\right)\sum_{j=1}^{u}x_j^2\,\mathrm{E}\left[\ 'j^2\right] \\ &\quad + (k^2 \quad k)kxk_2^2\,\mathrm{E}_{'}[k\,'k_2^2] \\ &= \frac{3k}{u}\left(\frac{u}{3} \quad \left(1 \quad \frac{1}{q}\right)\right)\left(kxk_2^2\,\mathrm{E}\left[k\,'k_2^2\right] + 2kxk_2^2\,\mathrm{E}_{'}[\ '^2]\right) + \frac{6k}{u}\left(1 \quad \frac{1}{q}\right)kxk_2^2\,\mathrm{E}\left[\ '^2\right] \\ &\quad + (k^2 \quad k)kxk_2^2\,\mathrm{E}_{'}[k\,'k_2^2] \\ &= k\left(k \quad \frac{3}{u}\left(1 \quad \frac{1}{q}\right)\right)kxk_2^2u\ ^2 + 2kkxk_2^2\ ^2 \\ &= kkxk_2^2\ ^2\left(ku + 2 \quad 3\left(1 \quad \frac{1}{q}\right)\right)\end{aligned}$$

Inserting into (4.12) and applying Section 4.9.2, we see that

$$\begin{aligned}\mathrm{Var}[1=kk\ \ x+\ \ 'k_2^2] &\quad \frac{3}{k}kxk_2^4+O\left(\frac{u^2\ \ ^4}{k}\right)+\frac{6}{k}kxk_2^2\ \ ^2\left(ku+2\ \ \ 3\left(1\ \ \ \frac{1}{q}\right)\right)\ \ \ 4u\ \ ^2kxk_2^2\\ &=\frac{3}{k}kxk_2^4+O\left(\frac{u^2\ \ ^4}{k}\right)+\frac{2}{k}kxk_2^2\ \ ^2\left(ku+6+9\left(\frac{1}{q}\ \ \ 1\right)\right)\end{aligned}$$

Substituting x for $x\ \ \ y$ and $'$ for $'\ \ \ \ $ proves that

$$\mathrm{Var}[1=kk\ \ (x+\ ')\ \ \ \ (y+\ \ )k_2^2\ \ \ 2u\ \ ^2]\ \ \ \frac{3}{k}kx\ \ \ yk_2^4+O\left(\frac{u^2\ \ ^4}{k}+u\ \ ^2kxk_2^2+\frac{^2}{qk}kxk_2^2\right)$$

Recalling that $q=\min\left\{\ \ \left(\frac{\log k}{u}\right);1\right\}$ we get

$$\frac{3}{k}kx\ \ \ yk_2^4+O\left(\frac{u^2\ \ ^4}{k}+u\ \ ^2kxk_2^2\right)$$

concluding the proof. □

4.9.4 Omitted Proofs for SJLT

Proof of SJLT Satisfying LPP

Lemma 4.9. *The SJLT as described above satisfy LPP from De nition 4.2.*

Proof. We show the result here for the c)-construction. A similar proof shows the result for the b)-construction.

$$\begin{aligned}\mathrm{E}_S\left[kSxk_2^2\right]=\mathrm{E}_S\left[\sum_{i=1}^{k=s}\sum_{r=1}^{s}(Sx)_{(i;r)}^2\right]&=\frac{1}{s}\mathrm{E}_{h;'}\left[\sum_{i=1}^{k=s}\sum_{r=1}^{s}\left(\sum_{j=1}^{u}\ '_r(j)\ _{ri}(j)x_j\right)^2\right]\\ =\frac{1}{s}\mathrm{E}_S\left[\sum_{i=1}^{k=s}\sum_{r=1}^{s}\sum_{j;`=1}^{u}\ '_r(j)\,'_r(`)\ _{ri}(j)\ _{ri}(`)x_jx_`\right]&=\frac{1}{s}\sum_{j=1}^{u}x_j^2\sum_{i=1}^{k=s}\sum_{r=1}^{s}\mathrm{E}_h[\ _{ri}(j)]=kxk_2^2\end{aligned}$$

because $'_r(j)$ and $'_r(`)$ are independent for $j\ \ \ `$ and $\mathrm{E}_{'}['_r(j)]=0$. □

Proof of Variance of (non-private) SJLT

The following lemma will be useful throughout this section. The proof is immediate from the de nition of .

Lemma 4.13.

$$\mathrm{E}[\ _{ri}(j)\ _{tn}(`)]=\begin{cases}\mathrm{E}[\ _{ri}(j)]\mathrm{E}[\ _{tn}(`)]; & j\ \ \ `\\ \mathrm{E}[\ _{ri}(j)^2]; & r=t;\ i=n;\ j=`\\ \mathrm{E}[\ _{ri}(j)]\mathrm{E}[\ _{ti}(j)]; & r\ \ \ t;\ i=n;\ j=`\\ 0; & r=t;\ i\ \ \ n;\ j=`\\ \mathrm{E}[\ _{ri}(j)]\mathrm{E}[\ _{tn}(j)]; & r\ \ \ t;\ i\ \ \ n;\ j=`\end{cases}=\begin{cases}s^2=k^2; & j\ \ \ `\\ s=k; & r=t;\ i=n;\ j=`\\ s^2=k^2; & r\ \ \ t;\ i=n;\ j=`\\ 0; & r=t;\ i\ \ \ n;\ j=`\\ s^2=k^2; & r\ \ \ t;\ i\ \ \ n;\ j=`\end{cases}$$

where we recalled that

$$h_r(j)=i\ \wedge\ h_r(j)=n;\qquad ,\qquad i=n:$$

We now prove Lemma 4.10. We state it here for convenience.

Lemma 4.10. *Let $x; y \in \mathbb{R}^u$ and let S be the SJLT as described above. Then*

$$\mathrm{Var}\left[\|Sx\|_2^2 - \|Sy\|_2^2\right] \leq \frac{2}{k}\|x\|_2^4 \|y\|_2^4.$$

Proof. Throughout the proof, we will apply Lemma 4.13 without further comment. By linearity of S and since S satisfies LPP, it is sufficient to show that for $x \in \mathbb{R}^u$

$$\mathrm{Var}\left[\|Sx\|_2^2\right] = \mathbb{E}_S\left[\|Sx\|_2^4\right] - \left(\mathbb{E}_S\left[\|Sx\|_2^2\right]\right)^2 = \mathbb{E}_S\left[\|Sx\|_2^4\right] - \|x\|_2^4 \leq \frac{2}{k}\|x\|_2^4.$$

We will consider the first term:

$$\begin{aligned}\mathbb{E}_S\left[\|Sx\|_2^4\right] &= \mathbb{E}_S\left[\left(\sum_{i=1}^{k/s}\sum_{r=1}^{s}(Sx)_{(i,r)}^2\right)^2\right] = \mathbb{E}_S\left[\left(\sum_{i=1}^{k/s}\sum_{r=1}^{s}\left(\sum_{j=1}^{u}\frac{1}{s}x_j\,\sigma_r(j)\,\delta_{ri}(j)\right)^2\right)^2\right] \\ &= \frac{1}{s^2}\mathbb{E}_S\left[\left(\sum_{i=1}^{k/s}\sum_{r=1}^{s}\sum_{j,\ell=1}^{u}x_j x_\ell\,\sigma_r(j)\sigma_r(\ell)\,\delta_{ri}(j)\,\delta_{ri}(\ell)\right)^2\right]\end{aligned} \tag{4.13}$$

Letting

$$a = \sum_{i=1}^{k/s}\sum_{r=1}^{s}\sum_{j=1}^{u}x_j^2\,\delta_{ri}(j); \qquad \text{and} \qquad b = \sum_{i=1}^{k/s}\sum_{r=1}^{s}\sum_{j\neq\ell}^{u}x_j x_\ell\,\sigma_r(j)\sigma_r(\ell)\,\delta_{ri}(j)\,\delta_{ri}(\ell);$$

we can express (4.13) as

$$\frac{1}{s^2}\mathbb{E}_S\left[(a+b)^2\right] = \frac{1}{s^2}\mathbb{E}_S\left[a^2 + b^2 + 2ab\right] \tag{4.14}$$

The proofs of the following claims are straightforward but tedious, and thus we leave them out here. They can be found later in this section.

Claim 4.2.

$$\mathbb{E}_S\left[a^2\right] \leq s^2\|x\|_2^4.$$

Claim 4.3.

$$\mathbb{E}_S\left[b^2\right] = \frac{2s^2}{k}\left(\|x\|_2^4 - \|x\|_4^4\right).$$

Claim 4.4.

$$2\,\mathbb{E}_S[ab] \geq 0$$

Inserting Claims 4.2-4.4 into (4.14), we conclude that

$$\mathbb{E}_S\left[\|Sx\|_2^4\right] \leq \|x\|_2^4 + \frac{2}{k}\left(\|x\|_2^4 - \|x\|_4^4\right)$$

finally proving that

$$\mathrm{Var}\left[\|Sx\|_2^2\right] \leq \frac{2}{k}\left(\|x\|_2^4 - \|x\|_4^4\right) \leq \frac{2}{k}\|x\|_2^4.$$

□

Proof of Claims

Throughout this section, we apply Lemma 4.13 repeatedly without further comment.

Claim 4.2.

$$\mathsf{E}_S\left[a^2\right] = s^2 k x k_2^4.$$

Proof.

$$\begin{aligned}
\mathsf{E}_h\left[\left(\sum_{i=1}^{k=s}\sum_{r=1}^{s}\sum_{j=1}^{u} x_j^2\ {}_{ri}(j)\right)^2\right] &= \sum_{i,n=1}^{k=s}\sum_{r,t=1}^{s}\sum_{j,\grave{}=1}^{u} x_j^2 x_{\grave{}}^2\,\mathsf{E}_h[\ {}_{ri}(j)\ {}_{tn}(\grave{})] \\
&= \sum_{i,n=1}^{k=s}\sum_{r,t=1}^{s}\sum_{j=1}^{u} x_j^4\,\mathsf{E}_h[\ {}_{ri}(j)\ {}_{tn}(j)] + \sum_{i,n=1}^{k=s}\sum_{r,t=1}^{s}\sum_{j\neq\grave{}=1}^{u} x_j^2 x_{\grave{}}^2\,\mathsf{E}_h[\ {}_{ri}(j)\ {}_{tn}(\grave{})] \\
&= \sum_{i=1}^{k=s}\sum_{r=1}^{s}\sum_{j=1}^{u} x_j^4\frac{s}{k} + \sum_{i=1}^{k=s}\sum_{r\neq t=1}^{s}\sum_{j=1}^{u} x_j^4\frac{s^2}{k^2} + \sum_{i\neq n=1}^{k=s}\sum_{r=1}^{s}\sum_{j=1}^{u} x_j^4 \quad 0 \\
&\qquad + \sum_{i\neq n=1}^{k=s}\sum_{r\neq t=1}^{s}\sum_{j=1}^{u} x_j^4\frac{s^2}{k^2} + \sum_{i,n=1}^{k=s}\sum_{r,t=1}^{s}\sum_{j\neq\grave{}=1}^{u} x_j^2 x_{\grave{}}^2\frac{s^2}{k^2} \\
&= skxk_4^4 + \frac{s(s^2 \quad s)}{k}kxk_4^4 + \left(1 \quad \frac{s}{k}\right)(s^2 \quad s)kxk_4^4 + s^2\left(kxk_2^4 \quad kxk_4^4\right) \\
&= s^2 k x k_2^4
\end{aligned}$$

□

Claim 4.3.

$$\mathsf{E}_S\left[b^2\right] = \frac{2s^2}{k}\left(kxk_2^4 \quad kxk_4^4\right).$$

Proof.

$$\begin{aligned}
&\mathsf{E}_{h,'}\left[\left(\sum_{i=1}^{k=s}\sum_{r=1}^{s}\sum_{j\neq\grave{}} x_j x_{\grave{}}\ '_r(j)\ '_r(\grave{})\ {}_{ri}(j)\ {}_{ri}(\grave{})\right)^2\right] \\
&= \sum_{i,n=1}^{k=s}\sum_{r,t=1}^{s}\sum_{\substack{j\neq\grave{}\\ v\neq w}} x_j x_{\grave{}} x_v x_w\,\mathsf{E}_{'}[\ '_r(j)\ '_r(\grave{})\ '_t(v)\ '_t(w)]\,\mathsf{E}_h[\ {}_{ri}(j)\ {}_{ri}(\grave{})\ {}_{tn}(v)\ {}_{tn}(w)]
\end{aligned}$$

We remark that, as $j \neq \grave{}$, we have

$$\mathsf{E}_{'}[\ '_r(j)\ '_r(\grave{})\ '_t(v)\ '_t(w)] = \begin{cases} 1; & r = t \wedge ((j = v \wedge \grave{} = w) \_ (j = w \wedge \grave{} = v)) \\ 0; & \text{otherwise.} \end{cases}$$

This leaves us with

$$\begin{aligned}
2\sum_{i,n=1}^{k=s}\sum_{r=1}^{s}\sum_{j\neq\grave{}} x_j^2 x_{\grave{}}^2\,\mathsf{E}_h[\ {}_{ri}(j)\ {}_{ri}(\grave{})\ {}_{rn}(j)\ {}_{rn}(\grave{})] &= 2\sum_{i=1}^{k=s}\sum_{r=1}^{s}\sum_{j\neq\grave{}} x_j^2 x_{\grave{}}^2\,\mathsf{E}_h\left[\ {}_{ri}(j)^2\right]\mathsf{E}_h\left[\ {}_{ri}(\grave{})^2\right] \\
&= 2\sum_{i=1}^{k=s}\sum_{r=1}^{s}\sum_{j\neq\grave{}} x_j^2 x_{\grave{}}^2\frac{s^2}{k^2} = \frac{2s^2}{k}\left(kxk_2^4 \quad kxk_4^4\right)
\end{aligned}$$

where we used that $\mathsf{E}_h[\ {}_{ri}(j)\ {}_{rn}(j)] = 0$ if $i \neq n$. □

Claim 4.4.

$$2\,\mathrm{E}_S[ab] = 0$$

Proof. Observe that

$$\mathrm{E}_S[ab] = \sum_{i,v=1}^{k=s} \sum_{r,t=1}^{s} \sum_{\substack{j=1 \\ v\neq w}}^{u} x_j^2 x_v x_w \,\mathrm{E}_{\cdot}[\,'_t(v)\,'_t(w)]\,\mathrm{E}_h[\,_{ri}(j)\,\,_{tn}(v)\,\,_{tn}(w)] = 0.$$

because the signs are independent. □

Chapter 5

Noise Distributions for Differential Privacy

This chapter is based on the paper *The Arete Distribution for Di erentially Private Noise Addition* by Rasmus Pagh & Nina Mesing Stausholm, which was in submission at the time of writing.

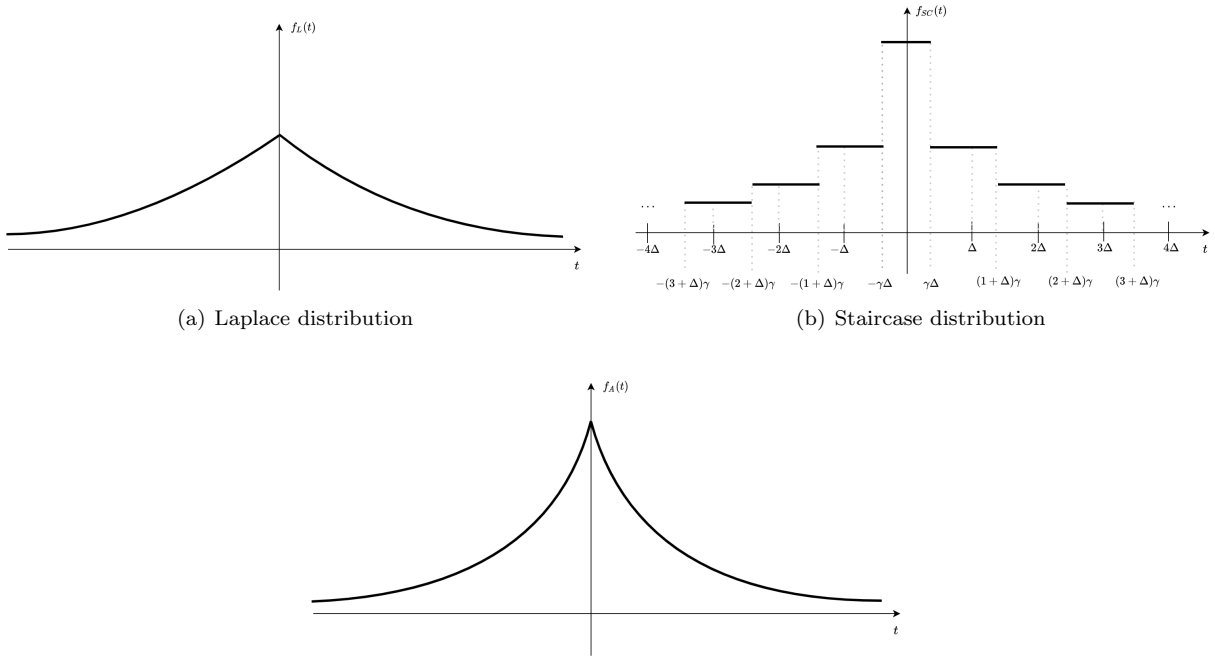
5.1 Introduction

Di erential privacy is generally achieved by adding random noise to a query output to perturb the true value, where the amount of noise is scaled according to the privacy parameter ". For real-numbered queries with sensitivity , Laplace distributed noise has become something of a standard for ensuring "-di erential privacy (see Lemma 2.3), leading to expected error (=") and variance (^{2}="^{2}). Geng, Kairouz, Oh, and Viswanath suggested an alternative to the Laplace mechanism named the *Staircase* mechanism [69, 70] (see Lemma 5.4). This mechanism adds noise from the so-called Staircase distribution, which can be parameterized to obtain error ($e^{-"=2}$) or variance ($^2e^{-2"=3}$), thus outperforming the Laplace mechanism for " larger than some constant.

In this chapter, we propose a new noise distribution, the *Arete*[1] distribution, with expected absolute value and variance, exponentially decreasing in ", and thus comparable to that of the Staircase distribution up to constant factors in ". The Arete distribution is, like the Laplace and Staircase distributions, oblivious of the data and the query output except for the sensitivity of the query, and we show that a mechanism adding random noise from the Arete distribution with suitable parameters is "-di erentially private. Figure 5.1 illustrate the shape of each of the three distributions (see Figure 5.2 for a plot of the Arete and Laplace density functions for " = 6 and " = 8). The Arete distribution has a continuous density function implying that the privacy level decreases more smoothly with sensitivity, in contrast to the Staircase distribution (see discussion in Section 5.2). Moreover, the Arete distribution is in nitely divisible (see De nition 5.3), and so we can divide the noise required to achieve di erential privacy between multiple players by letting each player draw independent *noise shares*, whose sum has the Arete distribution. Speci cally, we consider the shu e model of di erential privacy and secure multiparty aggregation. We argue that the Arete distribution allows for an improvement to the works of [73, 76] in the sense that we achieve error exponentially decreasing in " in a distributed setting. We discuss this further in Section 5.3.

The rest of this chapter is organized as follows: we rst give an informal de nition of the Arete distribution along with the de nition of the Arete mechanism in De nition 5.2 and our main results: Lemma 5.2 and Corollary 5.1. In Section 5.2 we discuss how our results relate to previous work and brie y mention how to extend it to multiple dimensions. In Section 5.3, we discuss how to apply the Arete distribution to achieve

[1]The name *Arete* is inspired by the word arête (pronounced "ah-ray't"), which is a sharp-crested mountain ridge, while also a concept from Greek mythology, Arete (pronounced "ah-reh-'tay") referring to moral virtue and excellence: the notion of the fulfillment of purpose or function and the act of living up to one's full potential [142].



(a) Laplace distribution

(b) Staircase distribution

(c) Arete distribution

Figure 5.1: Figure 2.2 repeated here for convenience. Illustration of the density functions for the Laplace, Staircase and Arete distributions. The purpose of this figure is to give an *intuitive* idea of the shapes. See Figure 5.2 for plots showing the densities of the Arete vs. Laplace distributions for certain parameter settings.

"-di erentially private protocols with low error in a distributed setting. Section 5.4 presents the technical preliminaries and gives a formal de nition of the Arete distribution, as well as the Laplace and Staircase mechanisms. Finally, Section 5.6.3 contains the proof of our main result, Lemma 5.2.

The Arete Distribution

We here give an informal introduction to the Arete distribution and refer to Section 5.4 for a formal de nition and a recap of the and Laplace distributions. The goal is to approximate the staircase distribution with an in nitely divisible distribution, so it is instructive to understand the essential properties of the staircase distribution: Only probability mass $e^{(")}$ is placed in the tails, which can be seen as a piece-wise uniform version of a scaled Laplace distribution. The majority of the probability mass is placed in a uniform distribution on an interval around zero of length $e^{(")}$.

Definition 5.1 (Arete distribution, informal)**.** *Let independent random variables $X_1; X_2$ (;) and Y Laplace( ). Then $Z := X_1 X_2 + Y$ has Arete distribution with parameters ; and , denoted Arete(; ;). When the parameters , and are understood from the context, we use $f_A(t)$, t 2 R, to denote the density function of Z.*

Since the and Laplace distributions are continuous, symmetric, and in nitely divisible it follows that the Arete distribution also has these properties. In Section 5.5.2 we show:

Lemma 5.1. *For any choice of parameters ; ; > 0, the Arete(; ;) distribution is in nitely divisible and has density $f_A(t)$ that is continuous, symmetric around 0, and monotonely decreasing for $t > 0$.*

Next, we discuss the intuition behind the noise and privacy properties of the Arete distribution: For privacy parameter " > 0 and sensitivity > 0 we concern ourselves with distributions D with support S

and density function f_D satisfying

$$e^{-\varepsilon} \leq \frac{f_D(t)}{f_D(t+a)} \leq e^{\varepsilon}, \qquad \forall t, a \in \mathbb{R}, |a| \leq \Delta \tag{5.1}$$

as this property is sufficient to ensure differential privacy, which is our main goal. We will refer to the property (5.1) as the *differential privacy constraint*. In order to minimize the magnitude of the noise, the goal is to find a distribution with minimal expected (absolute) value while satisfying (5.1).

The difference of two distributed random variables can be parameterized to have similar tails and to "peak" in an interval around zero of the same width as the staircase distribution. However, this does not provide differential privacy since the density function has a singularity at zero. To achieve differential privacy, we add a small amount of Laplace noise that "smooths out" the singularity. In more detail, the (α, β)-distribution (see Definition 5.4) with shape $\alpha < 1$, has most of its probability mass on an interval $(0, O(\beta))$. The difference of two distributions does not satisfy (5.1) for any choice of $\alpha < 1$, as the density tends to infinity for values going to zero. To fix this we need to "flatten the curve" of the density function in the neighborhood of 0. Consider $Z' := X + Y$ for independent $X \sim (\alpha, \beta)$ and $Y \sim Exp(\lambda)$. The Exponential distribution, with a suitable choice of parameter λ, is used to flatten the density function of the distribution close to 0. In order to get a noise distribution that is symmetric around zero, we further consider $Z = X_1 + Y_1 - (X_2 + Y_2)$ for $X_1, X_2 \sim (\alpha, \beta)$ and $Y_1, Y_2 \sim Exp(\lambda)$. Our definition of the Arete distribution follows from the fact that if $Y_1, Y_2 \sim Exp(\lambda)$, then $Y = Y_1 - Y_2 \sim Laplace(\lambda)$. We provide an explicit setting for the parameters α, β, λ in Lemma 5.2.

Main Results

Let the Arete distribution $Arete(\alpha, \beta, \lambda)$ be as in Definition 5.1 (and formally, Definition 5.7) with density function f_A. In Section 5.6.3 we show the following result:

Lemma 5.2. *For every choice of $\Delta > 2/e$ and $\varepsilon \geq 20 + 4\ln(1/\Delta)$ there exist parameters $\alpha, \beta, \lambda > 0$ such that:*

- *For every choice of $t, a \in \mathbb{R}$ with $|a| \leq \Delta$, $e^{-\varepsilon} \leq \frac{f_A(t)}{f_A(t+a)} \leq e^{\varepsilon}$:*
- *For $Z \sim Arete(\alpha, \beta, \lambda)$, $\mathrm{E}[|Z|] = O(\Delta e^{\varepsilon/4})$ and $\mathrm{Var}[Z] = O(\Delta^2 e^{\varepsilon/4})$.*

Parameters $\alpha = e^{-\varepsilon/4}$, $\beta = \frac{4}{\varepsilon\Delta}$ and $\lambda = e^{-\varepsilon/4}$ suffice.

The following corollary shows that adding noise from the Arete distribution gives an ε-differentially private mechanism.

Definition 5.2 (The Arete mechanism)**.** *Let $x \in X^d$ be an input and $q : X^d \to \mathbb{R}$ a query with sensitivity bounded by $\Delta \geq 2/e$. Given parameters α, β, λ, the* Arete mechanism *$M_{Arete}(x)$ samples $Z \sim Arete(\alpha, \beta, \lambda)$ and returns $q(x) + Z$.*

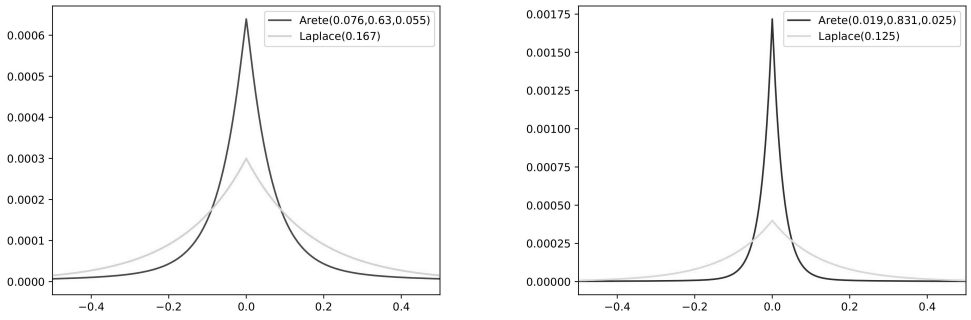
Corollary 5.1. *The Arete mechanism M_{Arete} with parameters as specified in Lemma 5.2 has expected error $O(\Delta e^{\varepsilon/4})$ and is ε-differentially private.*

Proof. We refer to Section 5.5.2 for the proof. □

Discussion of Large Values of ε

Values of ε larger than one are often used in practice – a few examples of deployments using large values of ε include Google's RAPPOR with ε up to 9 and Apple's MacOS with $\varepsilon = 6$ and iOS10 with $\varepsilon = 14$ [79] and US Census Bureau with ε up to 19.6 [121].

We acknowledge that the required lower bound on ε in Lemma 5.2 is high, but note that we have not optimized for constants in our analysis, and in practice, we may achieve differential privacy for significantly lower ε – that is, bullet 3 in Lemma 5.2 may be satisfied for lower values of ε than $20 + 4\ln(1/\Delta)$ (this claim can be supported by empirical analysis – see Figure 5.2). In order to get a mechanism for any $\varepsilon > 0$, one can add noise from the Arete distribution for $\varepsilon \geq 20 + 4\ln(1/\Delta)$ and apply the Laplace mechanism for smaller values of ε. For a thorough discussion of small versus large ε, we refer to [58, 63].



(a) Densities for $\varepsilon = 6$

(b) Densities for $\varepsilon = 8$

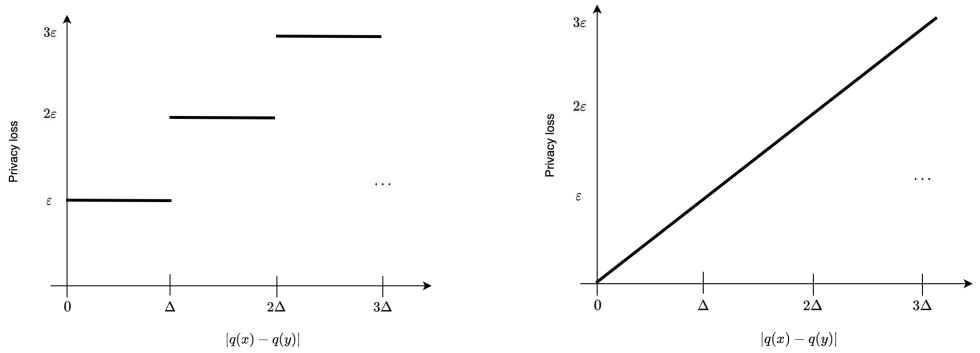
Figure 5.2: Density functions for Arete distributions that empirically yield ε-differential privacy with $\varepsilon = 6$ and $\varepsilon = 8$, respectively. The density functions were approximated by rounding the constituent Γ and Laplace distribution values to a multiple of 0.001 and computing the discrete convolution. Parameters were found using a local search heuristic. For comparison, Laplace distributions with the same privacy guarantee have been included and are clearly less concentrated around zero.

5.2 Related Work

A fundamental question that presents itself is what we can say about the tradeoff between error and privacy. Hardt & Talwar [82] study this tradeoff for linear queries, showing a lower bound of $\Omega(1/\varepsilon)$ for worst-case expected ℓ_2-norm of noise (std. deviation) under the constraint of ε-differential privacy for small ε. Nikolov et al. [118] extend the work [82] to the tradeoff between error and (ε, δ)-differential privacy. For error that can be a general function of the added noise, Ghosh, Roughgarden, Sundararajan & Gupte [75, 80] introduced the Geometric Mechanism for *counting queries* (integer valued) with sensitivity 1, showing that the optimal noise has a (symmetric) Geometric distribution with error (std. deviation) $\Theta(e^{\varepsilon/2})$. Brenner & Nissim [23] extend [75, 80] showing that for general queries there is no optimal mechanism for ε-differential privacy, while in the high privacy regime, Geng & Viswanath [71] present a (near) optimal mechanism for integer-valued vector queries for (ε, δ)-differential privacy, achieving error (for single-dimensional queries) $\Theta(\min\{1/\varepsilon, 1/\delta\})$ for small ε and δ. Though the geometric mechanism yields optimal error in the *discrete* setting and is infinitely divisible [76], it does not seem to generalize to a differentially private, infinitely divisible noise distribution in the real-valued setting.

Generalizing to *real-valued* one-dimensional queries with arbitrary sensitivity, Geng & Viswanath [70] introduced the ε-differentially private Staircase mechanism (see Lemma 5.4), which adds noise from the Staircase distribution – a geometric mixture of uniform distributions. The density function of the Staircase distribution, f_{SC}, is a piece-wise continuous step (or "staircase-shaped") function, symmetric around zero, monotonically decreasing, and geometrically decaying. Geng & Viswanath [70] prove that the optimal ε-differentially private mechanism for single real-valued queries, measuring error as expected magnitude or variance of the noise, is not Laplace but rather Staircase distributed: while the Laplace mechanism is asymptotically optimal as $\varepsilon \to 0$, the Staircase mechanism performs better in the low privacy regime (i.e., for large ε), as the expected magnitude of the noise is exponentially decreasing in ε. Specifically, for sensitivity Δ and for the parameter setting of δ optimizing for expected noise magnitude, the Staircase mechanism achieves error $\Theta(\Delta e^{-\varepsilon/2})$. For the choice of δ optimizing for variance, the Staircase mechanism ensures variance of the noise $\Theta(\Delta^2 e^{-2\varepsilon/3})$. We remark that the δ optimizing for noise magnitude is not generally the same a for optimizing for variance. The Laplace distribution has expected noise magnitude $\Theta(\Delta/\varepsilon)$ and variance $\Theta(\Delta^2/\varepsilon^2)$. In comparison, the expected noise magnitude and variance of the Arete distribution are also exponentially decreasing in ε, specifically $O(\Delta e^{-\varepsilon/4})$ and $O(\Delta^2 e^{-\varepsilon/4})$, respectively, for our choice of parameters. The expected error and variance mentioned here are for a single parameter setting for both Laplace and Arete mechanisms.

As we want a noise distribution that is implementable in a distributed setting, we limit our interest to



(a) Staircase mechanism (b) Arete mechanism

Figure 5.3: Figure 2.3 repeated here for convenience. Illustrations of the worst-case privacy loss of the Staircase and Arete mechanisms depending on the difference between query outputs. As we have no closed form for the density of the Arete distribution (see discussion in Note 5.1), we cannot explicitly determine the privacy loss, and so the graph given here is an approximation.

noise distributions that are oblivious of the input data and the query output. An important property of the Arete distribution is that the density function is continuous, and so we get a more graceful decrease in privacy than the Staircase mechanism for inputs that are not quite neighboring. For such inputs, the query outputs may di er by more than the sensitivity, and the di erential privacy requirement (as mentioned in (5.1)) no longer applies. It is, however, still interesting to study how the level of privacy decreases for inputs that are almost neighbors. The Staircase mechanism is *exactly* tted to the sensitivity of the query such that di erential privacy is guaranteed for neighboring inputs, but for query outputs that di er by more than the sensitivity, the level of privacy is immediately halved. The privacy level decreases in a smoother fashion when applying the Arete distribution due to the continuity of the density function (See Figure 5.3). Geng et al. [69] extend the Staircase mechanism from [70] to queries in multiple dimensions.

Multiple Dimensions

We will limit ourselves to single dimensional queries, but here brie y touch upon two techniques for extending to multiple dimensions. In order to generalize to d-dimensional queries, we may simply add independent noise from the Arete distribution to each coordinate of the query output, exactly as we usually do with Laplace noise (as was also the technique applied in Chapter 4). This strategy results in noise growing with the number of dimensions, d leading to expected absolute noise magnitude $\mathrm{E}[k'k_1] = O(d\ \mathrm{E}['_j]) = O(d\ e\ ^{"=4})$ and expected (squared) $`_2$-norm of the noise (i.e., variance) $\mathrm{E}[k'k_2^2] = O(d\ \mathrm{Var}['_j]) = O(d\ ^2e\ ^{"=4})$ for $'_j\ \ Arete(\ ;\ ;\ )$ with parameters $\ ;\ ;\ $ as speci ed in Lemma 5.2. Geng et al. [69] suggested an approach to extend the Staircase mechanism to multiple dimensions, where the sampling probability for d-dimensional noise vector $'$ depends on the $`_1$-norm of $'$:

$$f_{SC_{mult}}('') = \begin{cases} a(\ )e\ ^{k"}; & k'k_1\ 2\ [k\ \ ;(k+\ )\ \ ) \\ a(\ )e\ ^{(k+1)"}; & k'k_1\ 2\ [(k+\ )\ \ ;(k+1)\ \ ) \end{cases} ; \qquad k\ 2\ Z_{\ 0}:$$

Naturally, the normalizing factor $a(\ )$ is adapted to ensure that $f_{SC_{mult}}$ is still a probability measure. Geng et al. do not explicitly state the error incurred by this noise addition for general d or how it relates to the error obtained by adding independent Staircase distributed noise to each coordinate of the query output.

5.3 Applications

As discussed in Chapter 2, two models are prevalent in di erential privacy: the central model and the local model. In the *central* model of di erential privacy [57], all data is held by a single trusted unit which makes the result of a query di erentially private before releasing it. This is often done by adding noise to the query result. The central model usually has a very high level of accuracy but requires a high level of trust. Often, data is split among many curators, and a single trusted curator is not available. This setting is commonly known as the *local* model of di erential privacy [54, 95]. In this model, each curator must ensure privacy for their own data and so applies a di erentially private mechanism locally, which is then forwarded to an analyst who combines all reports to compute an approximate answer to the query. For many queries, the overall error in the local model grows rather quickly as a function of the number of players, signi cantly limiting utility. For example, while we can achieve error constant error in the central model [60], a count query requires $O(\sqrt{n})$ error for the same level of privacy as in the central model, where n is the number of players [33]. The local model is often attractive for data collection as the collecting organizations are not liable for storing sensitive user data in this model { a few examples of deployment include Google's RAPPOR [65], Apple (several features such as Lookup Hints, Emoji suggestion, etc.) [8] and Microsoft Telemetry [48].

In order to bridge this trust/utility gap, we may imitate the trusted unit from the centralized setting with cryptographic primitives [139], allowing for di erentially private implementations with better utility than in the local model while having lower trust assumptions than in the centralized model. Cryptographic primitives ensure that all parties learn only the output of the computation, while di erential privacy further bounds the information leakage from this output, so the combination gives powerful guarantees. We limit our discussion to the problem of computing the sum of real inputs, which is a basic building block in many other applications. If we can divide the noise among all players, we can obtain the same accuracy in a distributed setting as in the central model without assuming access to a trusted aggregator. Luckily, we can divide the noise between the players if the noise distribution D is in nitely divisible. Thus, the Arete distribution can be applied in this model.

We discuss di erential privacy implementations with two cryptographic primitives: Secure Multiparty Aggregation and Anonymous Communication but note that such implementations come with assumptions about the computational power of the analyst, which are accepted by the security community, but limit the privacy guarantee to computational di erential privacy [139].

Secure Multiparty Aggregation

The cryptographic primitive secure multiparty Aggregation, rooted in the work of Yao [149], has often been combined with di erential privacy to solve the problem of private real summation; see for example [21, 30, 125]. Goryczka et al. [76] give a comparative study of several protocols for private summation in a distributed setting. These protocols combine common approaches for achieving security (secret sharing, homomorphic encryption, and perturbation-based) while each party adds noise shares whose sum follows the Laplace distribution before sharing their data to ensure di erential privacy. Continuing their line of work, we may exchange the Laplace noise in [76] with Arete distributed noise to achieve $"$-di erentially private protocols with error exponentially small in $"$.

Anonymous Communication

Another line of work that has received much attention over the past few years is the *shu e* model of di erential privacy [19, 33, 64]. Along with Google's Prochlo framework, Bittau et al. [19] introduced the ESA (Encode Shu e Analyze) framework where each curator encodes their data before releasing it to a shu er. The shu er randomly permutes the encoded inputs and releases the (private) permuted set of data to an (untrusted) analyst, who then performs statistical analysis on the encoded, shu ed data. For recent work on the problem of summation in the shu e model and a discussion of error/privacy-tradeo , we refer to, for example, [11, 12, 72, 73, 74]. Ghazi et al. [73] propose an $(";)$-di erentially private protocol for summation in the shu e model for summing reals or integers where each user sends expected $1 + o(1)$ messages. The protocol adds discrete Laplace noise (also sometimes called Geometrically distributed noise)

and achieves error arbitrarily close to that of the Laplace mechanism (applied in the central model) while they leave open the problem of achieving error exponentially decreasing in " in the shu e model. The Arete distribution solves this open problem: as the Arete distribution is in nitely divisible, simply exchange the discrete Laplace noise (the \central" noise distribution in the protocol) with the Arete distribution.

Note. By the post-processing property of di erential privacy (Lemma 2.1), we still achieve di erential privacy if more than n players participate, and so we only need to choose the noise shares based on a lower bound on the number of players in order to ensure di erential privacy. Hence, it is not strictly necessary to know the exact number of players in advance.

5.4 Preliminaries

5.4.1 Probability Distributions

This section states the de nitions and basic facts that we need to analyze the Arete distribution. References to further information can be found in [76].

Definition 5.3 (In nite Divisibility)**.** *A distribution D is in nitely divisible if, for any random variable X with distribution D, then for every positive integer n there exist n i.i.d. random variables $X_1; :::; X_n$ such that $\sum_{i=1}^{n} X_i$ has the same distribution as X. The random variables X_i need not have distribution D.*

We recall the de nitions of the distributions that we use to de ne the Arete distribution and give a formal de nition of the latter. Whenever the parameters are implicit, we leave them out and simply write f , f_L, f and f_A for the densities of the , Laplace, and Arete distributions, resp.

Definition 5.4 (The Distribution)**.** *A random variable X has Gamma distribution with shape parameter > 0 and scale parameter > 0, denoted X (;), if its density function is*

$$f_{(;)}(t) = \frac{e^{t=} t^{1}}{( )}; \quad t > 0:$$

In the special case = 1, the random variable X has Exponential distribution with parameter .

The -distribution is in nitely divisible: For n independent random variables X_i ($_i$;), we have $X = \sum_{i=1}^{n} X_i$ ($\sum_{i=1}^{n}$ $_i$;). Furthermore, for X (;) we have E[X] = and Var[X] = ^{2}.

Definition 5.5 (The Laplace Distribution)**.** *A random variable X has Laplace distribution with location parameter and scale parameter > 0, denoted X Laplace(;), if its density function is*

$$f_{L(;)}(t) = \frac{e^{jt \quad j=}}{2}; \quad t \in \mathbb{R}:$$

If = 0 we just write Laplace( ).

If X *Laplace*( ), then jXj *Exp*( ) and E[X] = 0 while E[jXj] = . Similarly, Var[X] = 2 ^{2} while Var[jXj] = ^{2}. The Laplace distribution is in nitely divisible: For $2n$ independent random variables $X_i; Y_i$ (1=n;), we have $X = \sum_{i=1}^{n}(=n + X_i \quad Y_i)$ *Laplace* (;).

5.4.2 Differentially Private Mechanisms

Informally, di erential privacy promises that an analyst cannot, given a query answer, decide whether the underlying data contains a speci c data record or not. Therefore, di erential privacy relies on the notion of neighboring inputs, i.e., datasets $x; y \in X^d$ that di er by one data record. The sensitivity of a query quanti es how much the query output can di er for neighboring inputs and so describes how much di erence the added noise needs to hide. We refer to Section 2.1.2 for the formal de nitions of neighboring inputs, the sensitivity of a query, and di erential privacy.

We remind the reader of the Laplace mechanism (Lemma 2.3), here stated for single-dimensional queries.

Lemma 5.3 (The Laplace Mechanism [60]). *For real-valued query $q : X^d \to \mathbb{R}$ and input $x \in X^d$, the Laplace mechanism outputs $q(x) + X$ where $X \sim Lap(\ )$. If is the sensitivity of q, the Laplace mechanism with parameter $= =$" is "-di erentially private.*

Lemma 5.4 (The Staircase Mechanism [70]). *Let $q : \mathbb{R} \to \mathbb{R}$ be a real-valued query with sensitivity . Let random variable $X \sim SC(\ ;\ )$ have Staircase distribution with parameters $\in [0;1]$ and > 0 such that the density of X is*

$$f_{SC}(t) = \begin{cases} a(\ ); & t \in [0; \quad) \\ e^{\ "}a(\ ); & t \in [\quad ;\ ) \\ e^{\ k"}f_{SC}(t \quad k\ ); & t \in [k\ ;(k+1)\ );\ k \in \mathbb{N} \\ f_{SC}(\ t); & t < 0 \end{cases}$$

where $a(\ ) = \frac{1\ e^{\ "}}{2\ (\ +e^{\ "(1\ \ )})}$ is a normalization factor. Then for input $x \in \mathbb{R}$, the Staircase mechanism which outputs $q(x) + X$ where $X \sim SC(\ ;\ )$ is "-di erentially private.

For optimal parameter , the Staircase mechanism achieves expected absolute error $(\ e^{\ "=2})$ and variance $(\ ^2e^{\ 2"=3})$. We remark that the optimizing for expected magnitude of the noise is not the same as the optimizing for variance.

5.5 The Arete Distribution

We now turn to the Arete distribution. The following lemma is well-known from the probability theory literature:

Lemma 5.5. *If X and Y are independent, continuous random variables with density functions f_X and f_Y, then $Z = X + Y$ is a continuous random variable where the density is the convolution*

$$f_Z(z) = \int_{\ 1}^{\ 1} f_X(x)f_Y(z \quad x)dx = \int_{\ 1}^{\ 1} f_X(z \quad x)f_Y(x)dx:$$

The following distribution will be useful in de ning the Arete distribution:

Definition 5.6 (The Distribution). *Let $X_1; X_2 \sim (\ ;\ )$ be independent and de ne $X := X_1 \quad X_2$. We say that X has the $(\ ;\ )$ distribution and the density of X is*

$$f_{\quad (\ ;\ )}(t) = \int_0^{\ 1} f_{\ (\ ;\ )}(t+x)f_{\ (\ ;\ )}(x)dx = \begin{cases} \int_0^{\ 1} f\ (t+x)f\ (x)dx; & t \quad 0 \\ \int_{jtj}^{\ 1} f\ (t+x)f\ (x)dx; & t < 0: \end{cases}$$

where the integrals are reduced to the intervals where $f\ (t+x)f\ (x)$ is non-zero.

Lemma 5.6. *The distribution is in nitely divisible: For $2n$ independent random variables $X_i; Y_i \sim (\ =n;\ )$, we have $X = \sum_{i=1}^n (X_i \quad Y_i) \sim (\ ;\ )$.*

Proof. The result follows immediately from in nite divisibility of the -distribution. □

Definition 5.7 (The Arete distribution). *Let $X \sim (\ ;\ )$ and $Y \sim Laplace(\ )$ be independent. De ne $Z := X + Y$, then $Z \sim Arete(\ ;\ ;\ )$ for $\ ;\ ; > 0$. The density of Z is*

$$f_{A(\ ;\ ;\ )}(t) = \int_{\ 1}^{\ 1} f_{\quad (\ ;\ )}(t \quad x)f_{L(\ )}(x)dx = \int_{\ 1}^{\ 1} f_{L(\ )}(t \quad x)f_{\quad (\ ;\ )}(x)dx; \quad t \in \mathbb{R}:$$

Lemma 5.7. *The Arete distribution is in nitely divisible: For $4n$ independent random variables $X_{1i}; X_{2i} \sim (\ =n;\ )$ and $Y_{1i}; Y_{2i} \sim (1=n;\ )$, we have $X = \sum_{i=1}^n (X_{1i} \quad X_{2i} + (Y_{1i} \quad Y_{2i})) \sim Arete(\ ;\ ;\ )$.*

Proof. The result follows immediately from in nite divisibility of the Laplace distribution and Lemma 5.6. □

Note 5.1. *We remark that we are only interested in* 0 < < 1*. Furthermore, we do not explicitly state the density of the Arete distribution, as there is no simple closed form for the density of the distribution. (It can, however, be expressed in terms of Bessel functions { see [102].) A similar intuitive way of de ning our distribution would be to use a symmetric version of the -distribution (two halved -distributions put back-to-back at zero) instead of the -distribution. An essential property of our distribution is in nite divisibility such that we can draw independent noise shares that sum to a random variable following the Arete distribution. As opposed to our distribution, it is not clear whether a symmetric -distribution is in nitely divisible.*

5.5.1 Symmetric Density Functions

We observe some simple properties of the Arete distribution, see Section 5.8.1 for the omitted, elementary proofs.

Lemma 5.8. *For* $f; g : \mathrm{R} \to \mathrm{R}$*, that are symmetric around 0, i.e.,* $f(x) = f(-x)$ *and* $g(x) = g(-x)$*, we have for any* $t \in \mathrm{R}$

$$\int_{-1}^{1} f(x)g(t-x)dx = \int_{-1}^{1} f(x)g(|t|-x)dx.$$

In particular, the convolution $f * g$ *is symmetric around 0.*

Lemma 5.9. f *is symmetric around 0.*

Corollary 5.2. f_A *is symmetric around 0.*

Proof. The result follows directly from symmetry of the density of the Laplace distribution, f_L, and Lemmas 5.8 and 5.9. □

5.5.2 Properties of the Arete Distribution

We restate the lemma here for convenience

Lemma 5.1. *For any choice of parameters* $; ; > 0$*, the* $Arete(; ;)$ *distribution is in nitely divisible and has density* $f_A(t)$ *that is continuous, symmetric around 0, and monotonely decreasing for* $t > 0$*.*

Proof. Symmetry of the density function f_A is proven in Corollary 5.2 and in nite divisibility in Lemma 5.7. Since f and f_L are continuous, f and f_A are also continuous by Lemma 5.5. We prove that f_A is monotonely decreasing, i.e., for $|t| \leq |t'|$ we have $f_A(t) \geq f_A(t')$. First, we argue that f is monotonely decreasing. Recall De nition 5.6 and observe

$$f(t) = \int_{0}^{1} f(|t| + x)f(x)dx; \qquad \forall t \in \mathrm{R}$$

which is immediate for $t \geq 0$ while for $t < 0$

$$f(t) = \int_{|t|}^{1} f(-|t| + x)f(x)dx = \int_{0}^{1} f(x')f(x' + |t|)dx'$$

where we substituted $x' := x - |t|$. So assume $|t| \leq |t'|$. Then, since f is monotonely decreasing

$$f(t) = \int_{0}^{1} f(|t| + x)f(x)dx \geq \int_{0}^{1} f(|t'| + x)f(x)dx = f(t').$$

We prove that f_A is also monotonely decreasing: Assuming that jtj jt^0j we prove that $f_A(t)$ $f_A(t^0)$. Recall De nition 5.7 and observe

$$f_A(t) = \int_{1}^{1} f \quad (jtj \quad x) f_L(x)dx = \int_{1}^{1} f \quad (x) f_L(jtj \quad x)dx$$

which is obvious for t 0 and since for $t < 0$:

$$f_A(t) = \int_{1}^{1} f \quad (t \quad x) f_L(x)dx = \int_{1}^{1} f \quad (jtj + x) f_L(x)dx = \int_{1}^{1} f \quad (x^0) f_L(jtj \quad x^0)dx^0$$

using that f and f_L are symmetric and a substitution with $x^0 := jtj + x$. A similar argument can be made if the convolution is ipped. We conclude that

$$f_A(t) = \int_{1}^{1} f \quad (jtj \quad x) f_L(x)dx \quad \int_{1}^{1} f \quad (jt^0j \quad x) f_L(x)dx = f_A(t^0)$$

using that f is monotonely decreasing. □

We nally assume Lemma 5.2 and prove Corollary 5.1, restated here for convenience. The proof of Lemma 5.2 is given in Section 5.6.

Corollary 5.1. *The Arete mechanism M_{Arete} with parameters as speci ed in Lemma 5.2 has expected error $O(\ e^{\ "=4})$ and is "-di erentially private.*

Proof. The expected error bound follows directly from the bound on E[jZj] in Lemma 5.2. For the claim of di erential privacy, let $x; x^0 \ 2$ R with $jx \quad x^0j \quad 1$. We show that for any subset S R

$$\Pr[M_{Arete}(x) \ 2 \ S] \quad e^{"} \Pr[M_{Arete}(x^0) \ 2 \ S]: \tag{5.2}$$

Let noise Z $Arete(\ ; \ ; \ )$ for parameters $\ ; \ ;$ as in Lemma 5.2. De ne $S^0 := S \quad q(x) = fs \quad q(x) : s \ 2 \ Sg$, then:

$$\frac{\Pr[M_{Arete}(x) \ 2 \ S]}{\Pr[M_{Arete}(x^0) \ 2 \ S]} = \frac{\int_{S^0} f_A(z)dz}{\int_{S^0} f_A(z + q(x^0) \quad q(x))dz} \quad \frac{\int_{S^0} f_A(jzj)dz}{\int_{S^0} f_A(jzj + jq(x^0) \quad q(x)j)dz}$$

where we used symmetry of f_A, the triangle inequality, and the fact that $f_A(t)$ is decreasing for $t > 0$. By assumption $jq(x) \quad q(x^0)j$. Lemma 5.2 says that $f_A(t)=f_A(t + a) \quad e^{"}$ for all $t \ 2$ R and a , and so we get

$$\frac{\int_{S^0} f_A(jzj)dz}{\int_{S^0} f_A(jzj + jq(x^0) \quad q(x)j)dz} \quad \frac{\int_{S^0} e^{"} f_A(jzj + \quad)dz}{\int_{S^0} f_A(jzj + jq(x^0) \quad q(x)j)dz} \quad e^{"}:$$

That we also have $\Pr[M_{Arete}(x^0) \ 2 \ S] \quad e^{"} \Pr[M_{Arete}(x) \ 2 \ S]$ follows by symmetry. □

5.6 Proof of Main Lemma

In the remaining part of this section, we prove a number of theoretical lemmas that will help prove our main result, Lemma 5.2. The bulk of the analysis is the proof of the rst bullet point of Lemma 5.2, showing that the given parameters $\ ; \ ; \ > 0$ su ce to bound $f_A(t)=f_A(t + a)$ for all $t; a \ 2$ R, jaj . We break this part of the analysis down in this section. The intuition behind the structure is as follows: We rst remark that (to the best of our knowledge) there is no simple expression for the density of the distribution (see Note 5.1). Hence, we will show upper and lower bounds for f and use these to bound the ratio $f_A(t)=f_A(t + a)$. As discussed earlier, we have not optimized for constants, and as our proof includes *several* steps of bounding, our analysis may not be tight, thus leading to the high value of " required in Lemma 5.2. A tighter analysis will likely allow for a better setting of parameters $\ ; \ ;$, and a smaller ". We give the proof of Lemma 5.2 in Section 5.6.3.

5.6.1 Bounds on Density of Γ Γ Distribution

We rst derive upper and lower bounds on the density function of the distribution (see Section 5.4 for de nitions).

Lemma 5.10. *For any t 2 R and any > 0*

$$f\ (jtj +)c \quad f \quad (t) \quad f\ (jtj) \quad \text{where} \quad c := \int_0 \quad f\ (x)dx :$$

Proof. Recall De nition 5.6 and Lemma 5.9 and let t 2 R. For the upper bound, we have

$$f \quad (t) = \int_0^{1} f\ (jtj + x)f\ (x)dx < f\ (jtj) \int_0^{1} f\ (x)dx = f\ (jtj):$$

For the lower bound, we have for any > 0

$$f \quad (t) = \int_0^{1} f\ (jtj + x)f\ (x)dx \quad \int_0 \quad f\ (jtj + x)f\ (x)dx \quad f\ (jtj + \ ) \int_0 \quad f\ (x)dx :$$

□

5.6.2 Bounds on Density of Arete Distribution

In this section we show that for > 0 and setting of parameters ; ; and for large enough ":

$$e^{\ "} \quad f_A(t)=f_A(t + \ ) \quad e^{"}; \quad 8t\ 2\ \mathrm{R}: \tag{5.3}$$

We remark that by monotonicity of the density of the Arete distribution, it su ces to show (5.3) to prove that f_A satis es (5.1): Take any a 2 R such that jaj and suppose without loss of generality that $f(t)$ $f(t+a)$ (if this is not the case, substitute $t^0 := jtj$ a, such that $f(t^0)$ $f(t^0 + a)$). Then $e^{\ "} \frac{f(t)}{f(t+a)}$. We prove that $f(t+a)$ $f(t+\ )$ ensuring $\frac{f(t)}{f(t+a)}$ $\frac{f(t)}{f(t+\ )}$ $e^{"}$, which nishes the argument: by assumption $f(t)$ $f(t+a)$ and so jtj $jt + aj$, further implying that t a=2. Hence, as jaj we have $jt + \ j$ $jt + aj$ and so we conclude that $f(t + a)$ $f(t + \ )$ as wanted.

Throughout the section we assume that jtj $jt + \ j$ (and so t =2). For such t, $f_A(t)$ $f_A(t + \ )$ and so the rst inequality in (5.3) is immediate. Hence, we put our focus toward proving the latter inequality. If $jt + \ j$ jtj, the result follows by symmetry of f_A (Corollary 5.2).

We start with the following lower bound on the density f_A:

Lemma 5.11. *Let and c be as in Lemma 5.10 and assume* =ln(2) *for* > 0. *For* =2 *t 2* R

$$f_A(t + \ ) \quad f\ (jt + \ j + \ )c \quad c_L \quad \text{where} \quad c_L := 1=4:$$

Proof. By De nition 5.7 and Lemma 5.10 we have

$$\begin{aligned} f_A(t + \ ) &= \int_1^{1} f \quad (t + \quad x)f_L(x)dx \quad \int_1^{1} f\ (jt + \quad xj + \ )c \quad f_L(x)dx \\ &\quad c \quad \int_0^{2(t+\ )} f\ (jt + \quad xj + \ )f_L(x)dx \quad c \quad f\ (t + \ + \ ) \int_0^{2(t+\ )} f_L(x)dx; \end{aligned}$$

where we used that $f\ (jt + \quad xj + \ )$ $f\ (jt + \ j + \ )$ for x 2 $(0; 2(t + \ ))$ and that by assumption $t + \ $ =2 allowing us to remove the absolute value signs. Again using that $t + \ $ =2

$$\int_0^{2(t+\ )} f_L(x)dx \quad \int_0 \quad f_{L(\ )}(x)dx = \frac{1}{2} \int_0 \quad f_{Exp(\ )}(x)dx \quad \frac{1}{4}; \qquad < \quad = \ln(2)$$

where we noticed that on the positive reals, the density function of the Laplace distribution is 1=2 times the density function of the Exponential distribution, and used that the median of the latter is ln(2) , so that the last inequality is valid as long as ln(2) . Hence,

$$f_A(t + \quad) \quad c \quad f \ (t + \quad + \quad)1=4:$$

De ning $c_L := 1=4$ nishes the proof. □

The following three lemmas are highly technical and give upper bounds for the ratio $f_A(t)=f_A(t + \quad)$; rst for large and small jtj separately in Lemmas 5.12 and 5.13 (i.e., for t close to and far from 0, where "close to/far from" is quanti ed by a parameter $_u$, which we will set in Lemma 5.15). We combine these results to an upper bound for general t in Lemma 5.14 (still assuming t is s.t. $f_A(t) \quad f_A(t + \quad)$) and nally choose parameters to ensure an upper bound of $e^{"}$ in Lemma 5.15, thus satisfying the second inequality of (5.3). Throughout the next three lemmas we make use the variables $\quad ; c \quad$ (from Lemma 5.10) and c_L (from Lemma 5.11), all of which will be handled in the proof of Lemma 5.15.

Lemma 5.12. *Let* $_u > 0$ *be given and assume* $0 < \quad 1$. *Let* $\quad ; c \quad$ *be as in Lemma 5.10 and* c_L *as in Lemma 5.11. Assume* $1= \quad 1= \quad 1=(\ _u + \quad + \quad)$ *and* $\quad = \ln(2)$ *for* $\quad > 0$. *For* $\quad =2 < t \ 2 \ \mathbb{R}$ *with* $jtj \quad _u$ *we have*

$$\frac{f_A(t)}{f_A(t + \quad)} \quad \frac{e^{(\quad + \quad)=}}{c} \left(\left(1 + \frac{+}{_u} \right) + \frac{c_{\ u} \ (\ )}{2 \ c_L} e^{\ u=} \ (\quad + \quad + \ _u)^1 \right)$$

where $c_{\ u} := 2 \int_0^{\ u} f \ (x)dx$.

Proof. The proof can be found in Section 5.8.2. □

Lemma 5.13. *Let* $_u > 0$ *be given and assume* $0 < \quad 1$. *Let* $\quad ; c \quad$ *be as in Lemma 5.10 and* c_L *as in Lemma 5.11. Assume* $1= \quad 1= \quad 1=(\quad + \quad)$ *and* $\quad = \ln(2)$ *for* $\quad > 0$. *For* $\quad =2 < t \ 2 \ \mathbb{R}$ *such that* $jtj \quad _u$ *we have*

$$\frac{f_A(t)}{f_A(t + \quad)} \quad \frac{e^{(\quad + \quad)=}}{c_L c} \left(\frac{_u + \quad +}{} + \ (\ ) \ (\quad + \ )^1 \right)$$

Proof. The proof can be found in Section 5.8.2. □

The following lemma combines Lemmas 5.12 and 5.13 to give an upper bound for general $t > \quad =2$:

Lemma 5.14. *Let* $\quad ; c \quad$ *be as in Lemma 5.10 and* c_L *as in Lemma 5.11. Assume* $0 < \quad 1, \quad +1,$ $\min f \ =2; \ = \ln(2) g$ *and* $(\ ) \quad 1= \quad$ *for* $\quad > 0$. *For* $\quad =2 < t \ 2 \ \mathbb{R}$

$$\frac{f_A(t)}{f_A(t + \quad)} \quad \frac{2e^{(\quad + \quad)=} \ e \ (\quad + \quad + \quad)}{c \quad c_L}$$

Proof. The proof can be found in Section 5.8.2. □

Note 5.2. *The fact that* $(\ ) \quad 1= \quad$ *for* $0 < \quad 1$ *follows from Euler's de nition of the Gamma function,*

$$(\ ) = \frac{1}{} \prod_{n=1}^{1} \frac{(1 + \frac{1}{n})}{1 + \frac{}{n}};$$

since $(1 + \frac{1}{n}) \quad 1 + \frac{}{n}$ *for any* $n > 0$ *and* $0 < \quad 1$.

We nally choose parameters $\quad ; \quad ;$ ensuring that the ratio $e^{\ "} \quad f(t)=f(t + \quad) \quad e^{"}$ for $"$ large enough:

Lemma 5.15. *Suppose* $" \quad 20 + 4 \ln(\ )$ *for* $\quad 2=e$. *Let* $\quad = e^{\ "=4}; \ = \frac{4}{"}$ *and* $\quad = e^{\ "=4}$. *Then for* $t \ 2 \ \mathbb{R}$

$$e^{\ "} \quad \frac{f_A(t)}{f_A(t + \quad)} \quad e^{"}:$$

Proof. The proof can be found in Section 5.8.2. □

5.6.3 Putting Things Together

We restate the lemma here for convenience:

Lemma 5.2. *For every choice of* $2=e$ *and* $"$ $20 + 4\ln(\ )$ *there exist parameters* $\ ;\ ;\ > 0$ *such that:*

- *For every choice of* $t; a\ 2$ R *with* jaj $, e^{\ "} \ \frac{f_A(t)}{f_A(t+a)}\ e^{"}$*:*
- *For* Z $Arete(\ ;\ ;\ )$*,* $\mathrm{E}[jZj] = O(\ e^{\ "=4})$ *and* $\mathrm{Var}[Z] = O(\ ^2e^{\ "=4})$*.*

Parameters $= e^{\ "=4}$*;* $= \frac{4}{"}$ *and* $= e^{\ "=4}$ *su ce.*

Proof. The rst bullet with the choice of parameters $= e^{\ "=4}$*;* $= 4\ =$" and $= e^{\ "=4}$ follow from Lemma 5.15 and monotonicity of f_A (Lemma 5.1), as described at the beginning of Section 5.6.2. The second bullet also follows from Lemma 5.15, as the expected error of a random variable $Z = X + Y$, where X $(\ ;\ )$ and Y $Laplace(\ )$, i.e., Z $Arete(\ ;\ ;\ )$, is

$$\mathrm{E}[jZj] = \mathrm{E}[jX + Yj] = \mathrm{E}[jX_1\ \ X_2 + Yj]\ \ 2\,\mathrm{E}[X_1] + \mathrm{E}[jYj] = 2\ \ +\ \ = \frac{8\ \ e^{\ "=4}}{"} + e^{\ "=4} = O\left(\frac{}{"}e^{\ "=4}\right)$$

where $X_1; X_2$ $(\ ;\ )$, and similarly, by independence

$$\mathrm{Var}[Z] = \mathrm{Var}[X + Y] = \mathrm{Var}[X_1\ \ X_2 + Y] = 2\,\mathrm{Var}[X_1] + \mathrm{Var}[Y] = 2\ \ ^2 + 2\ ^2 = 2\frac{16\ \ ^2e^{\ "=4}}{"^2} + 2e^{\ "=2}$$
$$= O\left(\frac{^2}{"^2}e^{\ "=4}\right)$$

with our choice of parameters. Finally, for $"$ $1=\sqrt{2}$ (which is signi cantly smaller than the values of $"$ that we are interested in), we may simplify to

$$\mathrm{E}[jZj] = O\left(\ e^{\ "=4}\right);\qquad \mathrm{Var}[Z] = O\left(\ ^2e^{\ "=4}\right)$$

thus nishing the proof. □

5.7 Open Problems

Our analysis of the privacy from this noise distribution involves a sequence of bounds, and so one could attempt a tighter analysis of the privacy guarantees ensured by the Arete mechanism. Especially we may ask for the optimal parameters such that the Arete mechanism ensures di erential privacy and whether we can get rid of the assumption on the size of $"$ (simulations suggest that the constant factors of the Arete mechanism, with parameters chosen as we have described, can be improved { see Figure 5.2). We remark that one could also have used other distributions to atten the -distribution { a Gaussian distribution might be a natural choice. We leave open the question of what privacy and accuracy guarantees can be achieved with other choices of distributions. Generally, we leave open the question of nding a noise distribution that is in nitely divisible and has a continuous density function while permitting a di erentially private mechanism matching the (optimal) error of $(\ e^{\ "=2})$ from the Staircase mechanism.

5.8 Technical Details

In this section we give the technical details and proofs omitted in the previous sections.

5.8.1 Omitted Proofs for Symmetric Density Functions

Lemma 5.8. *For $f;g : \mathbb{R} \to \mathbb{R}$, that are symmetric around 0, i.e., $f(x) = f(-x)$ and $g(x) = g(-x)$, we have for any $t \in \mathbb{R}$*

$$\int_{-1}^{1} f(x)g(t-x)dx = \int_{-1}^{1} f(x)g(|t|-x)dx.$$

In particular, the convolution $f * g$ is symmetric around 0.

Proof. The statement is immediate for $t \geq 0$, so suppose $t < 0$. Then for any $a;b \in \mathbb{R}$

$$\int_{a}^{b} f(x)g(t-x)dx = \int_{a}^{b} f(x)g(|t|+x)dx = \int_{b}^{a} f(-x)g(|t|-x)dx = \int_{b}^{a} f(x)g(|t|-x)dx$$

where the first step is by symmetry of g, the second step follows from integration by substitution, and the last step is by symmetry of f. In particular, we may let a and b be ∓ 1. □

Lemma 5.9. *f_{*} is symmetric around 0.*

Proof. We prove that $f_{*}(t) = f_{*}(|t|)$ for all $t \in \mathbb{R}$. Clearly, this is the case if $t \geq 0$, so suppose $t < 0$. By Definition 5.6

$$f_{*}(t) = \int_{|t|}^{1} f_{-}(t+x)f_{+}(x)dx = \int_{|t|}^{1} f_{-}(x-|t|)f_{+}(x)dx = \int_{0}^{1} f_{-}(x)f_{+}(|t|+x)dx = f_{*}(|t|)$$

where the penultimate step follows from integration by substitution with $x - |t|$. □

5.8.2 Omitted Proofs for Bounds on Density of Arete Distribution

Supporting lemmas

Lemma 5.16. *Let α be as in Lemma 5.10 and $\beta > 0$. Assume $0 < \gamma < 1$, $1/\alpha + 1/\beta = 1/\gamma$. Then $\forall t \geq 0$*

$$\frac{(\alpha+\beta+t)^{\alpha-1}}{e^{t(1/\alpha-1/\gamma)}} \leq \frac{(\alpha+\beta)^{\alpha-1}}{e^{-t/\gamma}}.$$

Proof. The function

$$g(t) = \frac{(\alpha+\beta+t)^{\alpha-1}}{e^{t(1/\alpha-1/\gamma)}}$$

maximized for

$$t^{*} = \frac{1}{1/\alpha-1/\gamma}(1/\alpha-1/\gamma)(\alpha+\beta) = \frac{1}{1/\alpha-1/\gamma} - (\alpha+\beta); \quad 1/\alpha-1/\gamma > 0; \quad 0 < \gamma < 1;$$

and monotonely decreasing for $t \geq t^{*}$. By assumption

$$\frac{1}{1/\alpha-1/\gamma} \leq (\alpha+\beta) \leq t:$$

and so $g(t) \leq g(t^{*})$. Furthermore, for all t

$$g(t) = \frac{(\alpha+\beta+t)^{\alpha-1}}{e^{t(1/\alpha-1/\gamma)}} \leq \frac{(\alpha+\beta)^{\alpha-1}}{e^{-t/\gamma}} = g(0):$$

□

Proof of Lemma 5.12

Lemma 5.12. *Let* $_u > 0$ *be given and assume* $0 <$ 1. *Let* $;c$ *be as in Lemma 5.10 and* c_L *as in Lemma 5.11. Assume* $1= \quad 1= \quad 1=(\ _u + \quad + \quad)$ *and* $\quad = \ln(2)$ *for* $\quad > 0$. *For* $\quad =2 < t\ 2$ R *with* $jtj \quad _u$ *we have*

$$\frac{f_A(t)}{f_A(t+\quad)} \quad \frac{e^{(\quad + \quad)=}}{c}\left(\left(1+\frac{+}{_u}\right)+\frac{c_{\ u}(\quad)}{2\ c_L}e^{\ u=(\quad+\quad+\ _u)^1}\right)$$

where $c_{\ u} := 2\int_0^{\ u} f\ (x)dx$.

Proof of Lemma 5.12. Suppose $_u \quad jtj$. By Lemma 5.10

$$\frac{f_A(t)}{f_A(t+\quad)} = \frac{\int_{\ 1}^{\ 1} f \quad (t \quad x)f_L(x)dx}{\int_{\ 1}^{\ 1} f \quad (t+\quad\ x)f_L(x)dx} \quad \frac{\int_{\ 1}^{\ 1} f\ (jt \quad xj)f_L(x)dx}{\int_{\ 1}^{\ 1} f\ (jt+\quad\ xj+\quad)c \quad f_L(x)dx}: \tag{5.4}$$

Note that $jt+\quad xj \quad jt \quad xj+\quad$ and

$$jt \quad xj\ ^1 = (jt \quad xj+\quad+\quad)\ ^1\left(1+\frac{+}{jt \quad xj}\right)^1 \quad : \tag{5.5}$$

So lling in the density f and applying (5.5), we can write (5.4) as

$$\frac{f_A(t)}{f_A(t+\quad)} \quad \frac{\int_{\ 1}^{\ 1} \frac{1}{(\quad)}e^{\ jt \quad xj=}(jt \quad xj+\quad+\quad)\ ^1\left(1+\frac{+}{jt \quad xj}\right)^1 \quad f_L(x)dx}{c\ \int_{\ 1}^{\ 1} f\ (jt \quad xj+\quad+\quad)f_L(x)dx}:$$

Since $\left(1+\frac{+}{jt \quad xj}\right)^1$ is maximized for $x\ !\ t$, we can bound this term as long as x is not too close to t. Hence, rewind to equation (5.4) and treat the cases where x is far from t and x is close to t separately by splitting the numerator from (5.4) at the intervals $x\ 2\ (\ 1;t \quad _u)\ [\ (t+\ _u;1)$ and $x\ 2\ (t \quad _u;t+\ _u)$ (these intervals are well-de ned since $_u > 0$):

$$\begin{aligned}\frac{f_A(t)}{f_A(t+\quad)} &\quad \frac{\int_{\ 1}^{\ t\ u} f\ (jt \quad xj)f_L(x)dx+\int_{t+\ _u}^{\ 1} f\ (jt \quad xj)f_L(x)dx}{c\ \int_{\ 1}^{\ 1} f\ (jt \quad xj+\quad+\quad)f_L(x)dx}+\frac{\int_{t\ \ _u}^{t+\ _u} f\ (jt \quad xj)f_L(x)dx}{c\ \int_{\ 1}^{\ 1} f\ (jt+\quad\ xj+\quad)f_L(x)dx}\\ &= \frac{\int_{\ 1}^{\ t\ u} e^{\ jt \quad xj=}(jt \quad xj+\quad+\quad)\ ^1\left(1+\frac{+}{jt \quad xj}\right)^1 \quad f_L(x)dx}{c\ \int_{\ 1}^{\ 1} e^{\ (jt \quad xj+\quad+\quad)=}(jt \quad xj+\quad+\quad)\ ^1 f_L(x)dx}\\ &\quad + \frac{\int_{t+\ _u}^{\ 1} e^{\ jt \quad xj=}(jt \quad xj+\quad+\quad)\ ^1\left(1+\frac{+}{jt \quad xj}\right)^1 \quad f_L(x)dx}{c\ \int_{\ 1}^{\ 1} e^{\ (jt \quad xj+\quad+\quad)=}(jt \quad xj+\quad+\quad)\ ^1 f_L(x)dx}\\ &\quad + \frac{\int_{t\ \ _u}^{t+\ _u} f\ (jt \quad xj)f_L(x)dx}{c\ \int_{\ 1}^{\ 1} f\ (jt+\quad\ xj+\quad)f_L(x)dx}\end{aligned}$$

where we in the last step again lled in the density function f and applied (5.5) in the rst two terms and left the last term as it was. Note that the constant $(\quad)$ from the density f cancels out in the fraction.

Now (still leaving the last term alone), for the rst two terms upper bound the factor

$$\left(1+\frac{+}{jt \quad xj}\right)^1 \qquad \left(1+\frac{+}{_u}\right)^1$$

and pull out $e^{(\quad+\quad)=}$ from the denominator, to see that

$$\frac{f_A(t)}{f_A(t+\quad)}$$

$$e^{(\quad+\quad)=}\left(1+\frac{+}{u}\right)^{1}\quad\frac{\left(\int_{1}^{t\ u}\frac{(jt\quad xj+\quad+\quad)\quad^{1}}{e^{jt\quad xj=}}f_L(x)dx+\int_{t+\ u}^{1}\frac{(jt\quad xj+\quad+\quad)\quad^{1}}{e^{jt\quad xj=}}f_L(x)dx\right)}{c\quad\int_{1}^{1}\frac{(jt\quad xj+\quad+\quad)\quad^{1}}{e^{jt\quad xj=}}f_L(x)dx}$$

$$+\frac{\int_{t\ u}^{t+\ u}f\ (jt\quad xj)f_L(x)dx}{c\quad\int_{1}^{1}f\ (jt+\quad\quad xj+\quad)f_L(x)dx}$$

$$\frac{1}{c}\left(e^{(\quad+\quad)=}\left(1+\frac{+}{u}\right)+\frac{\int_{t\ u}^{t+\ u}f\ (jt\quad xj)f_L(x)dx}{\int_{1}^{1}f\ (jt+\quad\quad xj+\quad)f_L(x)dx}\right)$$

where in the last step, we upper-bounded the fraction in the first term by $1=c$ and rounded the exponent $(1\quad)$ up to 1 for simpler notation. The following claim handles the last term and finishes the proof:

Claim 5.1. *Let $_u > 0$ be given. Let $;c$ be as in Lemma 5.10 and c_L as in Lemma 5.11. Assume $jtj\quad_u$, $jtj\quad jt+\quad j$, $\quad=\ln(2)$ and $1=\quad 1=\quad 1=(\quad+\quad+\ _u)$. Then*

$$\frac{\int_{t\ u}^{t+\ u}f\ (jt\quad xj)f_L(x)dx}{\int_{1}^{1}f\ (\quad+jt+1\quad xj)f_L(x)dx}\quad\frac{c_{\ u}(\quad)}{2\quad c_L}e^{(\quad+\ _u+\quad)=}(\quad+\quad+\ _u)^{1}$$

where

$$c_{\ u}:=2\int_{0}^{\ u}f\ (x)dx:$$

□

Proof of Claim. As $jtj > \ _u$, $f0g\ni(t\quad\ _u;t+\ _u)$ and so $f_L(x)$ is maximal at $x=\min\{jt\quad\ _uj;jt+\ _ujg=\min\{jtj+\ _u;jtj\quad\ _ug=jtj\quad\ _u$. Hence

$$\int_{t\quad\ u}^{t+\ u}f\ (jt\quad xj)f_L(x)dx\quad f_L(jtj\quad\ _u)\int_{t\quad\ u}^{t+\ u}f\ (jt\quad xj)dx=f_L(jtj\quad\ _u)c_{\ u}\qquad(5.6)$$

where we define

$$c_{\ u}:=\int_{t\quad\ u}^{t+\ u}f\ (jt\quad xj)dx=\int_{\ u}^{\ u}f\ (jxj)dx=2\int_{0}^{\ u}f\ (x)dx:$$

Now, consider

$$\frac{\int_{t\ u}^{t+\ u}f\ (jt\quad xj)f_L(x)dx}{\int_{1}^{1}f\ (\quad+jt+\quad\quad xj)f_L(x)dx}:$$

Recalling the assumptions $jtj\quad jt+\quad j$ and $\quad=\ln(2)$, apply Lemma 5.11 in the denominator and (5.6) in the numerator, we get

$$\frac{\int_{t\ u}^{t+\ u}f\ (jt\quad xj)f_L(x)dx}{\int_{1}^{1}f\ (\quad+jt+\quad\quad xj)f_L(x)dx}\quad\frac{f_L(jtj\quad\ _u)c_{\ u}}{f\ (\quad+jt+\quad j)c_L}\quad\frac{c_{\ u}e^{\ jjtj\quad\ _uj=}(\quad)}{2\quad c_Le^{\ (\quad+\quad+jtj)=}(\quad+\quad+jtj)^{\quad1}}$$

$$\frac{c_{\ u}(\quad)}{2\quad c_L}e^{\ u=}e^{(\quad+\quad)=}\frac{(\quad+\quad+jtj)^{1}}{e^{jtj(1=\quad1=\ )}}$$

where we at ( ) lled in the density functions f and f_L and used that $jt + \ j \ jtj + \ $. In the last step, recall $jtj > \ _u$, so $jtj \ \ _u > 0$. Applying Lemma 5.16 (recall $0 < \ < 1$ and the assumption $1= \ 1= \ 1=(\ + \ + \ _u)$) with $= \ _u$, we get

$$\frac{\int_{t \ _u}^{t+ \ _u} f \ (jt \ \ xj) f_L(x)dx}{\int_{1}^{1} f \ (\ +jt + \ \ xj) f_L(x)dx} \ \ \frac{c_{\ _u}(\ )}{2 \ c_L} e^{\ _u=} e^{(\ \ + \ )=} \frac{(\ \ + \ \ + \ _u)^{1}}{e^{\ _u(1= \ \ 1=\ )}}$$
$$= \frac{c_{\ _u}(\ )}{2 \ c_L} e^{(\ \ + \ _u + \ )=} (\ \ + \ \ + \ _u)^{1} \ :$$

□

Proof of Lemma 5.13

Lemma 5.13. *Let* $_u > 0$ *be given and assume* $0 < \ \ 1$. *Let* $\ ; c \ $ *be as in Lemma 5.10 and* c_L *as in Lemma 5.11. Assume* $1= \ \ 1= \ \ 1=(\ + \ )$ *and* $\ = \ln(2)$ *for* $\ > 0$. *For* $\ =2 < t \ 2$ R *such that* $jtj \ \ _u$ *we have*

$$\frac{f_A(t)}{f_A(t + \ )} \ \ \frac{e^{(\ \ + \ )=}}{c_L c} \left(\frac{_u + \ \ + \ }{} + (\ ) \ (\ \ + \ )^{1} \right)$$

Proof of Lemma 5.13. Suppose $jtj < \ _u$. By Lemmas 5.8 and 5.10 we have

$$f_A(t) = f_A(jtj) = \int_{1}^{1} f \ \ (x) f_L(jtj \ \ x)dx \ \ \int_{1}^{1} f \ (jxj) f_L(jtj \ \ x)dx:$$

Note that $f_L(jtj \ \ x) \ \ f_L(t)$ when $jjtj \ \ xj \ \ jtj$, which is satis ed whenever $x \ 2$ (0; 2jtj).

$$f_A(t) = f_A(jtj) = \int_{1}^{1} f \ \ (x) f_L(jtj \ \ x)dx \ \ \int_{1}^{1} f \ (jxj) f_L(jtj \ \ x)dx$$
$$= \int_{1}^{0} f \ (jxj) f_L(jtj \ \ x)dx + \int_{0}^{2jtj} f \ (jxj) f_L(jtj \ \ x)dx + \int_{2jtj}^{1} f \ (jxj) f_L(jtj \ \ x)dx$$
$$f_L(t) \left(\int_{1}^{0} f \ (jxj)dx + \int_{2jtj}^{1} f \ (jxj)dx \right) + \int_{0}^{2jtj} f \ (jxj) f_L(jtj \ \ x)dx$$
$$^{(\ )} \ 2f_L(t) + 2 \int_{0}^{jtj} f \ (jxj) f_L(jtj \ \ x)dx$$
$$= 2f_L(t) + \frac{2}{(\ ) \ \ 2} \int_{0}^{jtj} e^{\ x=} x \ \ ^{1} e^{\ jjtj \ \ xj=} dx$$
$$= 2f_L(t) + \frac{1}{(\ )} e^{\ jtj=} \int_{0}^{jtj} e^{x(1= \ \ 1=\ )} x \ \ ^{1} dx$$
$$2f_L(t) + \frac{1}{(\ )} e^{\ jtj=} \int_{0}^{jtj} x \ \ ^{1} dx$$
$$= 2f_L(t) + \frac{1}{(\ )} e^{\ jtj=} \frac{jtj}{}$$

At ( ) we use that $f \ (jxj)$ is smaller on $(t; 2t)$ than on $(0; t)$. In the last step we used that

$$\int_{0} x^{n} dx = \frac{n+1}{n+1}; \ \ \ \ n \ 6= \ 1:$$

Recalling the assumptions $jtj \quad jt + \quad j$ and $\quad = \ln(2)$, apply Lemma 5.11 to get

$$\frac{f_A(t)}{f_A(t+\quad)} \quad \frac{2f_L(t) + \frac{1}{(\quad)} e^{\quad jtj=\quad} \underline{jtj}}{c_L c \quad f(\quad + jt + \quad j)} = \frac{1}{c_L c}\left(\frac{2f_L(t)}{f(\quad + jt + \quad j)} + \frac{1}{(\quad)} \frac{e^{\quad jtj=\quad} \underline{jtj}}{f(\quad + jt + \quad j)}\right)$$

$$\overset{(\quad)}{=} \frac{1}{c_L c}\left(\frac{2e^{\quad jtj=\quad}(\quad)}{2\, e^{(\quad + jt + \quad j)=(\quad + jt + \quad j)\quad 1}} + \frac{(\quad)}{(\quad)} \frac{e^{\quad jtj=\quad} \underline{jtj}}{e^{(\quad + jt + \quad j)=(\quad + jt + \quad j)\quad 1}}\right)$$

$$\overset{(\quad)}{=} \frac{1}{c_L c}\left(\frac{e^{\quad jtj=\quad}(\quad)}{e^{(\quad + jtj + \quad)=(\quad + jtj + \quad)\quad 1}} + \frac{e^{\quad jtj=\quad} \underline{jtj}}{e^{(\quad + jtj + \quad)=(\quad + jtj + \quad)\quad 1}}\right)$$

$$= \frac{e^{(\quad + \quad)=}}{c_L c}\left(\frac{e^{\quad jtj=\quad}(\quad)}{e^{\quad jtj=(\quad + jtj + \quad)\quad 1}} + \frac{e^{\quad jtj=\quad} \underline{jtj}}{e^{\quad jtj=(\quad + jtj + \quad)\quad 1}}\right)$$

$$= \frac{e^{(\quad + \quad)=}}{c_L c}\left((\quad)\ \frac{(\quad + jtj + \quad)^{1}}{e^{jtj(1=\quad 1=\quad)}} + (\quad + jtj + \quad)^{1}\ \underline{jtj}\right):$$

where we at $(\quad)$ lled in the density functions and at $(\quad)$ used that $jt + \quad j \quad jtj + \quad$.

Using the identity $jtj \quad = jtj=jtj^{1}$ we see that

$$\frac{f_A(t)}{f_A(t+\quad)} \quad \frac{e^{(\quad + \quad)=}}{c_L c}\left((\quad)\ \frac{(\quad + jtj + \quad)^{1}}{e^{jtj(1=\quad 1=\quad)}} + \underline{jtj}\left(\frac{\quad + jtj + \quad}{jtj}\right)^{1}\right)$$

$$\frac{e^{(\quad + \quad)=}}{c_L c}\left((\quad)\ \frac{(\quad + jtj + \quad)^{1}}{e^{jtj(1=\quad 1=\quad)}} + \frac{\quad + jtj + \quad}{}\right)$$

$$\frac{e^{(\quad + \quad)=}}{c_L c}\left((\quad)\ \frac{(\quad + jtj + \quad)^{1}}{e^{jtj(1=\quad 1=\quad)}} + \frac{\quad + \quad_u + \quad}{}\right);$$

where in the last two steps, we removed the $(1 \quad)$-exponent on the last term and used that $jtj \quad _u$. Finally applying Lemma 5.16 (recall $0 < \quad < 1$ and the assumption $1= \quad 1= \quad 1=(\quad + \quad)$) with $\quad = 0$ nishes the proof:

$$\frac{f_A(t)}{f_A(t+\quad)} \quad \frac{e^{(\quad + \quad)=}}{c_L c}\left((\quad)\ (\quad + \quad)^{1} \quad + \frac{\quad + \quad_u + \quad}{}\right):$$

□

Proof of Lemma 5.14

Lemma 5.14. *Let* $\quad ; c \quad$ *be as in Lemma 5.10 and* c_L *as in Lemma 5.11. Assume* $0 < \quad 1, \quad +1,$ $\min f \quad =2; \quad = \ln(2) g$ *and* $(\quad) \quad 1= \quad$ *for* $\quad > 0$. *For* $\quad =2 < t \ 2 \ \mathbb{R}$

$$\frac{f_A(t)}{f_A(t+\quad)} \quad \frac{2e^{(\quad + \quad)=} e\ (\quad + \quad + \quad)}{c \quad c_L}$$

Proof. We rst give the intuition behind the proof: Lemmas 5.12 and 5.13 give upper bounds on the ratio for certain values of t, assuming $1= \quad 1= \quad \max f 1=(\ _u + \quad + \quad); 1=(\quad + \quad) g = 1=(\quad + \quad)$. An upper bound on both of these bounds simultaneously gives us a bound on the ratio, which holds for general $t > \quad =2$. We note that

$$1= \quad 1= \quad 1=(\quad + \quad) \quad , \quad \frac{\quad}{\quad + 1};$$

so if $\quad +$, then $\quad =2$ su ces. Hence, our assumptions $\quad =2, \quad = \ln(2), \quad +$ and $t > \quad =2$ ensure that we can use Lemmas 5.12 and 5.13.

So, by Lemmas 5.12 and 5.13 we have for $=2 < t\ 2$ R

$$\frac{f_A(t)}{f_A(t+\ )} \quad \frac{e^{(\ +\ )=}(\ _u+\ +\ )}{c}\max\left\{\frac{1}{c_L};\frac{1}{\ _u}\right\}$$
$$+\frac{e^{(\ +\ )=}}{c}\frac{(\ )}{c_L}\max\left\{(\ +\ )^1\ ;\frac{c\ _u e^{\ u=}}{2}(\ +\ +\ _u)^1\ \right\}:$$

As, by assumption, $+\ $, we see

$$(\ +\ )^1 \ < \ (\ _u+\ +\ )^1 \ < \ _u+\ +$$

and so we may simplify to

$$\frac{f_A(t)}{f_A(t+\ )} \quad \frac{e^{(\ +\ )=}}{c}\left(\max\left\{\frac{1}{c_L};\frac{1}{\ _u}\right\}(\ _u+\ +\ )+\frac{(\ )(\ +\ +\ _u)}{c_L}\max\left\{1;\frac{c\ _u e^{\ u=}}{2}\right\}\right)$$
$$=\frac{e^{(\ +\ )=}(\ _u+\ +\ )}{c}\left(\max\left\{\frac{1}{c_L};\frac{1}{\ _u}\right\}+\frac{(\ )}{c_L}\max\left\{1;\frac{c\ _u e^{\ u=}}{2}\right\}\right):$$

Let $_u =$ (i.e., the mean of the -distribution). Recalling that by de nition $c_L = 1=4$ and by assumption $=2$, so $c_L \quad =8 < \ = \ _u$:

$$\frac{f_A(t)}{f_A(t+\ )} \quad \frac{e^{(\ +\ )=}(\ +\ +\ )}{c}\left(\frac{1}{c_L}+\frac{(\ )}{c_L}\max\left\{1;\frac{c\ _u e}{2}\right\}\right)$$
$$\frac{e^{(\ +\ )=}(\ +\ +\ )}{c\ \ c_L}\left(\frac{1}{-}+(\ )e\right);$$

where the last step follows from the observation that $1 \quad c\ _u \quad 2$ (recall $c\ _u$ was de ned in Lemma 5.12) and $e\ > 1$ for > 0.

By assumption $(\ ) < 1=\ $, then

$$1=\ +(\ )e \quad \frac{2e}{-}$$

and so we conclude

$$\frac{f_A(t)}{f_A(t+\ )} \quad \frac{2e^{(\ +\ )=}e\ (\ +\ +\ )}{c\ \ c_L}:$$

□

Proof of Lemma 5.15

Lemma 5.15. *Suppose* $" \quad 20+4\ln(\ )$ *for* $\quad 2=e$. *Let* $= e^{\ "=4}$; $= \frac{4}{"}$ *and* $= e^{\ "=4}$. *Then for* $t\ 2$ R

$$e^{\ "} \quad \frac{f_A(t)}{f_A(t+\ )} \quad e^{"}:$$

Proof. Suppose $jtj \quad jt+\ j$. The rst inequality is satis ed as $f_A(t) \quad f_A(t+\ )$. Let be as in Lemma 5.10. We turn to prove the latter inequality: In order to apply Lemma 5.14 we make the following assumptions:

$$+\ ; \qquad =2; \qquad =\ln(2) \quad \text{and} \quad (\ ) \quad 1=\ : \tag{5.7}$$

We choose parameters satisfying these assumptions towards the end of the proof.

If is at least the median of the -distribution then $c \quad 1=2$. So let $= \quad$ be the mean of the -distribution (the mean is an upper bound on the median of the -distribution [31]), to see

$$\frac{f_A(t)}{f_A(t+\ )} \overset{\text{(Lemma 5.14)}}{\quad} \frac{2e^{(\ +\ )=}e\ (\ +\ +\ )}{c\ \ c_L} \quad \frac{2\ (2\ +\ )}{1=2\ \ 1=4}\frac{e\ ^{=}e^2}{-}: \tag{5.8}$$

Suppose 1=2 and recall by assumption + = + , so =(1). Then =(1) = and so . We revise our set of assumptions, to also ensure that , and so our set of assumptions is:

$$\frac{\quad}{1};\qquad \min f\ =2;\ \ =\ln(2)g;\qquad 1=2\quad\text{and}\quad(\ )\quad 1=\ :\tag{5.9}$$

Under these assumptions we have 2 + 3 and inserting into (5.8), we conclude

$$\frac{f_A(t)}{f_A(t+\ \ )}\quad 48\ \ \frac{e^{\ =}\ e^2}{\ \ }:$$

Now de ne

$$=1=e^{"=k}\ ;\qquad =\frac{k}{"}\quad\text{and}\qquad =1=e^{"=k}$$

Observing 2 1 and $\ln(48e) < 4{:}9$

$$\frac{f_A(t)}{f_A(t+\ \ )}\quad 48e^2\quad e^{"(1=k\ +1=k\ +1=k\ )}<e^{"(1=k\ +1=k\ +1=k\ )+4:9+\ln(\ \ )}:$$

Hence, we ensure that

$$\frac{f_A(t)}{f_A(t+\ \ )}\quad e^{"}$$

when the assumptions in (5.9) are satis ed and

$$"(1=k\ +1=k\ +1=k\ )+5+\ln(\ \ )\quad "\ ,\quad 1=k\ +1=k\ +1=k\quad 1\ \ \frac{5+\ln(\ \ )}{"}:\tag{5.10}$$

Let $k = k = k = 4$. It is easy to check that the assumptions on and in (5.9) are satis ed simultaneously for, " 4 ln(2) (and we can check that ( ) 1= numerically). Furthermore, for " 4 ln(2), we require $\min f2="; 1=\ln(2)g = 2\ =$" and so the assumption on is satis ed when " =2 = $"e^{\ "=4}=2$. Observing that $2=e\quad "e^{\ "=4}=2$, we conclude that the assumptions in (5.9) are satis ed for " 4 ln(2) and 2=e. The inequality in (5.10) is satis ed for

$$3=4\quad 1\ \ \frac{5+\ln(\ \ )}{"}\quad ,\quad 20+4\ln(\ \ )\quad ":$$

Finally, observe that if $jtj\quad jt+\ \ j$, the result follows by symmetry of f_A (Corollary 5.2). □

Chapter 6

Conclusion and Open Problems

> "Can anything harm us, mother, after the night-lights are lit?"
> "Nothing, precious," she said; "they are the eyes a mother leaves behind her to guard her children."
>
> J. M. Barrie
> Peter Pan

As data collection and analysis escalates, it becomes increasingly relevant to understand how to perform accurate and private analysis over distributed data. Private data analysis allows us to gain valuable knowledge without leaking sensitive information. In this thesis, we have seen two new di erentially private sketches for e ciently and accurately estimating Euclidean distance and (weighted) cardinality for the symmetric di erence between two sets. We have seen formal proofs for the privacy, accuracy, and e ciency guarantees (which are comparable to known lower bounds) of the sketches and corresponding estimators. Moreover, we have seen a new noise distribution, the Arete distribution, which is symmetric around zero and monotonically decreasing for $t \quad 0$. We have described a new mechanism, the Arete mechanism, for real-valued queries adding noise from this distribution and provided a parameter setting ensuring that this mechanism is $"$-di erentially private. We have proven that the Arete mechanism achieves error comparable to the (optimal) Staircase mechanism. Simultaneously, the Arete distribution stands out compared to the Staircase distribution due to two desirable properties: (1) It has a continuous density function, thus ensuring a smooth decrease in privacy for inputs that are not quite neighboring, and (2) it is in nitely divisible, and so allows for combination with cryptographic techniques to permit accurate statistical analysis when input data is distributed among many participants.

Concluding each chapter, we have explored open problems related to the contributions presented. A key question concerning di erentially private analysis over distributed data (mentioned in Chapter 3) is how to combine di erentially private linear sketches such that one can analyze (functions of) the whole dataset without a blowup in the noise level. That is, is there a way to combine private sketches while (almost) preserving the noise level and hence the accuracy guarantees? This question is particularly interesting if we wish to combine an unknown number of sketches. A closely related question with a whole di erent issue is how to continuously release sketches or statistics over the *same* (updated) dataset while preserving privacy. This problem, often referred to as *continual observation*, was proposed by Dwork et al. [61] and is particularly interesting for streaming data. Continual observation under di erential privacy has a vast number of applications in practice including analysis of user behavior on social media, tra c analysis or monitoring the spread of a pandemic. Whereas the noise level (and so also the privacy level) increases when we combine private sketches for distributed datasets, the privacy level deteriorates over time when releasing several statistics over the same dataset (recall The Fundamental Law of Information Recovery). Hence,

continuous releases may result in serious data leaks. Several large organizations have faced massive critique of their data collection for applying privacy techniques developed for single-time use and so not handling this particular problem su ciently well { see for example [13, 26, 42, 50, 79]. Di erential privacy under continual observation has gained some attention, but previous works all have rather strict assumptions. Dwork et al. [61] (and extensions [29, 88]) show how to keep a counter under continual observation in the central model by adding carefully correlated noise across updates, thereby achieving error polylog in the number of time steps (the number of times the counter is released). Erlingsson et al. [64] combined the technique from [61] with shu ing to allow for continual observation in the local model of di erential privacy. These works all require that the number of time steps is known in advance, and the error scales with this number. Joseph et al. [91] studied a technique where the accuracy guarantees degrade with the number of times the *query result* changes "signi cantly" and so relies on the assumption that user inputs change infrequently. In case of more frequent updates, one can use memoization to answer identical queries consistently. This technique was used in the deployments by Google [65] and Microsoft [48], but has the limitation that memoization is vulnerable to adversaries asking several correlated queries. So, we ask the question, which may well be one of the most relevant privacy questions at the moment: *How can we continuously release accurate statistics over streaming data while preserving di erential privacy?*

Appendix A

Notation

General	
$a; b$	lower case letters refer to vectors and numbers
$A; B$	upper case letters refer to sets, matrices and random variables
$A; B$	upper case curly letters refer to distributions, mechanisms and domains
$\hat{A}; \hat{B}$	hat for estimates
Specific Variables	
$"$;	privacy parameters
	accuracy parameter
$'$;	noise vectors
p	a probability
U	the universe
u	the size of the universe and input dimension
S	sketch matrix
s	sparsity of matrix
	variance parameter for Normal distribution
	scale parameter for Laplace/Exponential distribution
;	shape and scale parameters for -distribution, resp.
	query sensitivity
Functions and set notation	
$[n]$	$f1; 2; :::; ng$
$1[P]$	indicator function for predicate P
$j \ j$	absolute value
$k \ k_p$	$`_p$-norm
$h \ ; \ i$	inner product
$A \ 4 \ B$	symmetric di erence between sets A and B
Distributions	
$X \quad D$	Random variable X with distribution D
D	(Typically noise) distribution
$N(\ ; \ ^2)$	Normal distribution with mean and variance parameters and $^2 > 0$, resp.
$Lap(\ ; \ )$	Laplace distribution with mean and scale parameters and > 0, resp.
$(\ ; \ )$	Gamma distribution with shape and scale parameters > 0 and > 0, resp.
$Exp(\ )$	Exponential distribution with scale parameter > 0.
$SC(\ ; \ )$	Staircase distribution with parameters $0 < < 1, > 0$
$Arete(\ ; \ ; \ )$	Arete distribution with parameters $0 < < 1, > 0, > 0$

References

[1] Achlioptas, D. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci. 66*, 4 (2003), 671{687.

[2] Ács, G., and Castelluccia, C. I have a dream! (Di eRentially privatE smArt Metering). In *Information Hiding, IH* (2011), vol. 6958 of *Lecture Notes in Computer Science*, pp. 118{132.

[3] Ailon, N., and Chazelle, B. The fast johnson{lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput. 39*, 1 (2009), 302{322.

[4] Alaggan, M., Gambs, S., and Kermarrec, A. BLIP: non-interactive di erentially-private similarity computation on bloom lters. In *Stabilization, Safety, and Security of Distributed Systems - 14th International Symposium, SSS* (2012), pp. 202{216.

[5] Alaggan, M., Gambs, S., Matwin, S., and Tuhin, M. Sanitization of call detail records via di erentially-private bloom lters. In *Data and Applications Security and Privacy XXIX - 29th Annual IFIP WG 11.3 Working Conference, DBSec 2015* (2015), pp. 223{230.

[6] Alon, N., Gibbons, P. B., Matias, Y., and Szegedy, M. Tracking join and self-join sizes in limited storage. *Journal of Computer and System Sciences 64*, 3 (2002), 719{747.

[7] Alon, N., Matias, Y., and Szegedy, M. The space complexity of approximating the frequency moments. In *Symposium on the Theory of Computing* (1996), pp. 20{29.

[8] Apple. Apple di erential privacy technical overview. `https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf`.

[9] Awasthi, P., Balcan, M.-F., Haghtalab, N., and Zhang, H. Learning and 1-bit compressed sensing under asymmetric noise. In *Conference on Learning Theory* (2016), pp. 152{192.

[10] Balle, B., Barthe, G., and Gaboardi, M. Privacy ampli cation by subsampling: Tight analyses via couplings and divergences. In *Neural Information Processing Systems, NeurIPS* (2018), pp. 6280{6290.

[11] Balle, B., Bell, J., Gascón, A., and Nissim, K. The privacy blanket of the shu e model. In *Advances in Cryptology - CRYPTO* (2019), vol. 11693 of *Lecture Notes in Computer Science*, pp. 638{667.

[12] Balle, B., Bell, J., Gascón, A., and Nissim, K. Private summation in the multi-message shu e model. In *Conference on Computer and Communications Security, CCS* (New York, NY, USA, 2020), CCS '20, p. 657{676.

[13] Bambauer, J., Muralidhar, K., and Sarathy, R. Fool's gold: an illustrated critique of di erential privacy. *Vand. J. Ent. & Tech. L. 16* (2013), 701.

[14] Bar-Yossef, Z., Jayram, T., Kumar, R., Sivakumar, D., and Trevisan, L. Counting distinct elements in a data stream. In *International Workshop on Randomization and Approximation Techniques in Computer Science* (2002), pp. 1{10.

[15] Barbaro, M., and Zeller, T. A face is exposed for aol searcher no. 4417749. *New York Times* (01 2006).

[16] Bebensee, B. Local di erential privacy: a tutorial. *CoRR abs/1907.11908* (2019).

[17] Bhaskar, R., Bhowmick, A., Goyal, V., Laxman, S., and Thakurta, A. Noiseless database privacy. In *17th International Conference on the Theory and Application of Cryptology and Information Security, ASIACRYPT* (2011), vol. 7073 of *Lecture Notes in Computer Science*, pp. 215{232.

[18] Bioglio, V., Bianchi, T., and Magli, E. Secure compressed sensing over nite elds. In *International Workshop on Information Forensics and Security (WIFS)* (2014), pp. 191{196.

[19] Bittau, A., Erlingsson, Ú., Maniatis, P., Mironov, I., Raghunathan, A., Lie, D., Rudominer, M., Kode, U., Tinnés, J., and Seefeld, B. Prochlo: Strong privacy for analytics in the crowd. In *Symposium on Operating Systems Principles, SOSP* (2017), pp. 441{459.

[20] Blocki, J., Blum, A., Datta, A., and Sheffet, O. The Johnson-Lindenstrauss transform itself preserves di erential privacy. In *53rd Symposium on Foundations of Computer Science, FOCS* (2012), pp. 410{419.

[21] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for privacy-preserving machine learning. In *Conference on Computer and Communications Security, CCS* (2017), pp. 1175{1191.

[22] Boutsidis, C., Zouzias, A., Mahoney, M. W., and Drineas, P. Randomized dimensionality reduction for k-means clustering. *IEEE Trans. Inf. Theory 61*, 2 (2015), 1045{1062.

[23] Brenner, H., and Nissim, K. Impossibility of di erentially private universally optimal mechanisms. In *Symposium on Foundations of Computer Science, FOCS* (2010), pp. 71{80.

[24] Bringmann, K., Kuhn, F., Panagiotou, K., Peter, U., and Thomas, H. Internal DLA: e cient simulation of a physical growth model. In *Automata, Languages, and Programming, ICALP* (2014), vol. 8572 of *Lecture Notes in Computer Science*, pp. 247{258.

[25] Broder, A. Z., and Mitzenmacher, M. Survey: Network applications of bloom lters: A survey. *Internet Mathematics 1*, 4 (2003), 485{509.

[26] Calandrino, J. A., Kilzer, A., Narayanan, A., Felten, E. W., and Shmatikov, V. "you might also like: " privacy risks of collaborative ltering. In *32nd Symposium on Security and Privacy, S&P* (2011), pp. 231{246.

[27] Canonne, C. L., Kamath, G., and Steinke, T. The discrete gaussian for di erential privacy. *CoRR abs/2004.00010* (2020).

[28] Carole Cadwalladr & Emma Graham-Harrison, T. O. Revealed: 50 million facebook pro les harvested for cambridge analytica in major data breach. `https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election`. Published: 2018-03-17.

[29] Chan, T. H., Shi, E., and Song, D. Private and continual release of statistics. *ACM Trans. Inf. Syst. Secur. 14*, 3 (2011), 26:1{26:24.

[30] Chan, T. H., Shi, E., and Song, D. Privacy-preserving stream aggregation with fault tolerance. In *Financial Cryptography and Data Security - FC* (2012), vol. 7397 of *Lecture Notes in Computer Science*, pp. 200{214.

[31] Chen, J., and Rubin, H. Bounds for the di erence between median and mean of Gamma and Poisson distributions. *Statistics & Probability Letters 4*, 6 (October 1986), 281{283.

[32] Chen, R., Xiao, Q., Zhang, Y., and Xu, J. Di erentially private high-dimensional data publication via sampling-based inference. In *21st International Conference on Knowledge Discovery and Data Mining* (2015), pp. 129{138.

[33] Cheu, A., Smith, A. D., Ullman, J. R., Zeber, D., and Zhilyaev, M. Distributed di erential privacy via shu ing. In *Advances in Cryptology, EUROCRYPT* (2019), vol. 11476 of *Lecture Notes in Computer Science*, pp. 375{403.

[34] Choi, S. G., Dachman-Soled, D., Kulkarni, M., and Yerukhimovich, A. Di erentially-private multi-party sketching for large-scale statistics. *IACR Cryptol. ePrint Arch. 2020* (2020), 29.

[35] Clarkson, K. L. Tighter bounds for random projections of manifolds. In *24th Symposium on Computational Geometry* (2008), pp. 39{48.

[36] Clarkson, K. L., and Woodruff, D. P. Numerical linear algebra in the streaming model. In *41st Symposium on Theory of Computing, STOC* (2009), pp. 205{214.

[37] Clarkson, K. L., and Woodruff, D. P. Low rank approximation and regression in input sparsity time. In *Symposium on Theory of Computing Conference, STOC* (2013), pp. 81{90.

[38] Cohen, M. B., Elder, S., Musco, C., Musco, C., and Persu, M. Dimensionality reduction for k-means clustering and low rank approximation. In *47th Symposium on Theory of Computing, STOC* (2015), pp. 163{172.

[39] Cohen, R., Katzir, L., and Yehezkel, A. A uni ed scheme for generalizing cardinality estimators to sum aggregation. *Information Processing Letters 115*, 2 (2015), 336{342.

[40] Cormode, G., Garofalakis, M. N., Haas, P. J., and Jermaine, C. Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends in Databases 4*, 1-3 (2012), 1{294.

[41] Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, D., and Wang, T. Privacy at scale: Local di erential privacy in practice. In *International Conference on Management of Data, SIGMOD* (2018), pp. 1655{1658.

[42] Cyphers, B. Di erential privacy, part 3: Extraordinary claims require extraordinary scrutiny. https://www.accessnow.org/differential-privacy-part-3-extraordinary-claims-require-extraordinary-scrutiny/. Published: 2017-11-30.

[43] Dasgupta, A., Kumar, R., and Sarlós, T. A sparse Johnson-Lindenstrauss transform. In *42nd Symposium on Theory of Computing, STOC* (2010), pp. 341{350.

[44] De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. Unique in the crowd: The privacy bounds of human mobility. *Scienti c reports 3* (2013), 1376.

[45] Desfontaines, D., Lochbihler, A., and Basin, D. A. Cardinality estimators do not preserve privacy. *PoPETs 2019*, 2 (2019), 26{46.

[46] Differential Privacy Team, Apple. Learning with privacy at scale. https://machinelearning.apple.com/research/learning-with-privacy-at-scale. Published: December 2017.

[47] Differential Privacy Team, Google. Secure noise generation. Tech. rep., Google, 2020.

[48] Ding, B., Kulkarni, J., and Yekhanin, S. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems, NIPS* (2017), pp. 3571{3580.

[49] Dinur, I., and Nissim, K. Revealing information while preserving privacy. In *22nd Symposium on Principles of Database Systems, PODS* (2003), pp. 202{210.

[50] Domingo-Ferrer, J., Sánchez, D., and Blanco-Justicia, A. The limits of di erential privacy (and its misuse in data release and machine learning). *Commun. ACM 64*, 7 (2021), 33{35.

[51] Drew Harwell, T. W. P. Secret-sharing app whisper left users' locations, fetishes exposed on the web. https://www.washingtonpost.com/technology/2020/03/10/secret-sharing-app-whisper-left-users-locations-fetishes-exposed-web/. Published: 2020-03-10.

[52] Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlós, T. Faster least squares approximation. *Numerische Mathematik 117*, 2 (2011), 219{249.

[53] Dubhashi, D. P., and Panconesi, A. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[54] Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *54th Symposium on Foundations of Computer Science, FOCS* (2013), pp. 429{438.

[55] Dwork, C. Di erential privacy. In *Automata, Languages and Programming, ICALP* (2006), vol. 4052 of *Lecture Notes in Computer Science*, Springer, pp. 1{12.

[56] Dwork, C. Di erential privacy: A survey of results. In *Theory and Applications of Models of Computation, TAMC* (2008), vol. 4978 of *Lecture Notes in Computer Science*, pp. 1{19.

[57] Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology, EUROCRYPT* (2006), vol. 4004 of *Lecture Notes in Computer Science*, pp. 486{503.

[58] Dwork, C., Kohli, N., and Mulligan, D. K. Di erential privacy in practice: Expose your epsilons! *J. Priv. Con dentiality 9*, 2 (2019).

[59] Dwork, C., and Lei, J. Di erential privacy and robust statistics. In *41st Symposium on Theory of Computing, STOC* (2009), pp. 371{380.

[60] Dwork, C., McSherry, F., Nissim, K., and Smith, A. D. Calibrating noise to sensitivity in private data analysis. In *3rd Theory of Cryptography Conference, TCC* (2006), vol. 3876 of *Lecture Notes in Computer Science*, pp. 265{284.

[61] Dwork, C., Naor, M., Pitassi, T., and Rothblum, G. N. Di erential privacy under continual observation. In *42nd Symposium on Theory of Computing, STOC* (2010), pp. 715{724.

[62] Dwork, C., Naor, M., Pitassi, T., Rothblum, G. N., and Yekhanin, S. Pan-private streaming algorithms. In *ICS* (2010), pp. 66{80.

[63] Dwork, C., and Roth, A. The algorithmic foundations of di erential privacy. *Foundations and Trends in Theoretical Computer Science 9*, 3-4 (2014), 211{407.

[64] Erlingsson, Ú., Feldman, V., Mironov, I., Raghunathan, A., Talwar, K., and Thakurta, A. Ampli cation by shu ing: From local to central di erential privacy via anonymity. In *Symposium on Discrete Algorithms, SODA* (2019), SIAM, pp. 2468{2479.

[65] Erlingsson, Ú., Pihur, V., and Korolova, A. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In *Conference on Computer and Communications Security* (2014), pp. 1054{1067.

[66] EU, B. W. G. A guide to gdpr data privacy requirements. https://gdpr.eu/data-privacy/.

[67] Flajolet, P., Fusy, É., Gandouet, O., and Meunier, F. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. In *AofA: Analysis of Algorithms* (2007), pp. 137{156.

[68] Flajolet, P., and Martin, G. N. Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci. 31*, 2 (1985), 182{209.

[69] Geng, Q., Kairouz, P., Oh, S., and Viswanath, P. The staircase mechanism in di erential privacy. *IEEE J. Sel. Top. Signal Process. 9*, 7 (2015), 1176{1184.

[70] Geng, Q., and Viswanath, P. The optimal noise-adding mechanism in di erential privacy. *IEEE Trans. Inf. Theory 62*, 2 (2016), 925{951.

[71] Geng, Q., and Viswanath, P. Optimal noise adding mechanisms for approximate di erential privacy. *Transactions on Information Theory 62*, 2 (2016), 952{969.

[72] Ghazi, B., Golowich, N., Kumar, R., Manurangsi, P., Pagh, R., and Velingker, A. Pure di erentially private summation from anonymous messages. In *Information-Theoretic Cryptography, ITC* (2020), vol. 163 of *LIPIcs*, pp. 15:1{15:23.

[73] Ghazi, B., Kumar, R., Manurangsi, P., Pagh, R., and Sinha, A. Di erentially private aggregation in the shu e model: Almost central accuracy in almost a single message. In *International Conference on Machine Learning, ICML* (2021).

[74] Ghazi, B., Manurangsi, P., Pagh, R., and Velingker, A. Private aggregation from fewer anonymous messages. In *Advances in Cryptology - EUROCRYPT* (2020), vol. 12106 of *Lecture Notes in Computer Science*, pp. 798{827.

[75] Ghosh, A., Roughgarden, T., and Sundararajan, M. Universally utility-maximizing privacy mechanisms. *SIAM J. Comput. 41*, 6 (2012), 1673{1693.

[76] Goryczka, S., and Xiong, L. A comprehensive comparison of multiparty secure additions with di erential privacy. *Trans. Dependable Secur. Comput. 14*, 5 (2017), 463{477.

[77] Goryczka, S., Xiong, L., and Sunderam, V. S. Secure multiparty aggregation with di erential privacy: a comparative study. In *Joint 2013 EDBT/ICDT Conferences, EDBT/ICDT '13* (2013), pp. 155{163.

[78] Greenberg, A. Apple's `di erential privacy' is about collecting your data | but not your data. *Wired* (6 2016).

[79] Greenberg, A. How one of apple's key privacy safeguards falls short. *Wired* (9 2017).

[80] Gupte, M., and Sundararajan, M. Universally optimal privacy mechanisms for minimax agents. In *Symposium on Principles of Database Systems, PODS* (2010), J. Paredaens and D. V. Gucht, Eds., pp. 135{146.

[81] Haas, P. J., Naughton, J. F., Seshadri, S., and Stokes, L. Sampling-based estimation of the number of distinct values of an attribute. In *VLDB* (1995), vol. 95, pp. 311{322.

[82] Hardt, M., and Talwar, K. On the geometry of di erential privacy. In *42nd Symposium on Theory of Computing, STOC* (2010), pp. 705{714.

[83] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics 4*, 8 (2008), e1000167.

[84] Hsu, J., Khanna, S., and Roth, A. Distributed private heavy hitters. In *Automata, Languages, and Programming, ICALP* (2012), vol. 7391 of *Lecture Notes in Computer Science*, pp. 461{472.

[85] Indian Law and Justice Ministry. The personal data protection bill, 2019. https://prsindia.org/billtrack/the-personal-data-protection-bill-2019. Published: 2019-12-11.

[86] Indyk, P. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM 53*, 3 (2006), 307{323.

[87] Indyk, P., and Motwani, R. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *30th Annual Symposium on the Theory of Computing, STOC* (1998), pp. 604{613.

[88] Jain, P., Kothari, P., and Thakurta, A. Di erentially private online learning. In *25th Annual Conference on Learning Theory, COLT* (2012), vol. 23 of *JMLR Proceedings*, JMLR.org, pp. 24.1{24.34.

[89] Jayram, T. S., and Woodruff, D. P. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *Transactions on Algorithms 9*, 3 (2013), 26:1{26:17.

[90] Johnson, W. B., and Lindenstrauss, J. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics 26*, 189-206 (1984), 1.

[91] Joseph, M., Roth, A., Ullman, J. R., and Waggoner, B. Local di erential privacy for evolving data. *J. Priv. Con dentiality 10*, 1 (2020).

[92] Kane, D. M., Meka, R., and Nelson, J. Almost optimal explicit johnson-lindenstrauss families. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 14th International Workshop, APPROX and 15th International Workshop, RANDOM* (2011), vol. 6845 of *Lecture Notes in Computer Science*, pp. 628{639.

[93] Kane, D. M., and Nelson, J. Sparser Johnson-Lindenstrauss transforms. *J. ACM 61*, 1 (2014), 4:1{4:23.

[94] Kane, D. M., Nelson, J., and Woodruff, D. P. An optimal algorithm for the distinct elements problem. In *Proceedings of the 29th ACM symposium on Principles of database systems (PODS)* (2010), pp. 41{52.

[95] Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. D. What can we learn privately? *SIAM J. Comput. 40*, 3 (2011), 793{826.

[96] Kenthapadi, K., Korolova, A., Mironov, I., and Mishra, N. Privacy via the Johnson-Lindenstrauss transform. *J. Priv. Con dentiality 5*, 1 (2013).

[97] Kifer, D., Ben-David, S., and Gehrke, J. Detecting change in data streams. In *VLDB* (2004), vol. 4, Toronto, Canada, pp. 180{191.

[98] Kifer, D., and Machanavajjhala, A. No free lunch in data privacy. In *Proceedings of ACM International Conference on Management of data (SIGMOD)* (2011), pp. 193{204.

[99] Kushilevitz, E., Ostrovsky, R., and Rabani, Y. E cient search for approximate nearest neighbor in high dimensional spaces. In *Symposium on the Theory of Computing* (1998), pp. 614{623.

[100] Li, N., and Ye, Q. Mobile data collection and analysis with local di erential privacy. In *20th International Conference on Mobile Data Management, MDM* (2019), pp. 4{7.

[101] Marriott International. Marriott announces starwood guest reservation database security incident. https://news.marriott.com/2018/11/marriott-announces-starwood-guest-reservation-database-security-incident/. Published: 2018-11-30.

[102] MATHAI, A. On noncentral generalized laplacianness of quadratic forms in normal variables. *Journal of multivariate analysis 45*, 2 (1993), 239{246.

[103] MCGREGOR, A., MIRONOV, I., PITASSI, T., REINGOLD, O., TALWAR, K., AND VADHAN, S. The limits of two-party di erential privacy. In *51st Annual Symposium on Foundations of Computer Science* (2010), pp. 81{90.

[104] MCSHERRY, F., AND MIRONOV, I. Di erentially private recommender systems: Building privacy into the net ix prize contenders. In *15th International Conference on Knowledge Discovery and Data Mining* (2009), pp. 627{636.

[105] MCSHERRY, F., AND TALWAR, K. Mechanism design via di erential privacy. In *FOCS* (2007), vol. 7, pp. 94{103.

[106] MCSHERRY, F. D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of ACM International Conference on Management of data (SIGMOD)* (2009), pp. 19{30.

[107] MEISER, S. Approximate and probabilistic di erential privacy de nitions. *IACR Cryptol. ePrint Arch.* (2018), 277.

[108] MELIS, L., DANEZIS, G., AND CRISTOFARO, E. D. E cient private statistics with succinct sketches. In *23rd Annual Network and Distributed System Security Symposium, NDSS* (2016).

[109] MIR, D. J., MUTHUKRISHNAN, S., NIKOLOV, A., AND WRIGHT, R. N. Pan-private algorithms: When memory does not help. *CoRR abs/1009.1544* (2010).

[110] MIR, D. J., MUTHUKRISHNAN, S., NIKOLOV, A., AND WRIGHT, R. N. Pan-private algorithms via statistics on sketches. In *30th Symposium on Principles of Database Systems, PODS* (2011), pp. 37{48.

[111] MIRONOV, I. On signi cance of the least signi cant bits for di erential privacy. In *Conference on Computer and Communications Security, CCS* (2012), pp. 650{661.

[112] MIRONOV, I. Renyi di erential privacy. In *30th Computer Security Foundations Symposium, CSF* (2017), IEEE Computer Society, pp. 263{275.

[113] MIRONOV, I., PANDEY, O., REINGOLD, O., AND VADHAN, S. P. Computational di erential privacy. In *Advances in Cryptology - CRYPTO* (2009), S. Halevi, Ed., vol. 5677 of *Lecture Notes in Computer Science*, pp. 126{142.

[114] MITZENMACHER, M., PAGH, R., AND PHAM, N. E cient estimation for high similarities using odd sketches. In *Proceedings of 23rd international conference on World Wide Web (WWW)* (2014), pp. 109{118.

[115] NARAYANAN, A., AND SHMATIKOV, V. Robust de-anonymization of large sparse datasets. In *Symposium on Security and Privacy* (2008), pp. 111{125.

[116] NELSON, J., AND NGUYEN, H. L. OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings. In *54th Symposium on Foundations of Computer Science, FOCS* (2013), IEEE Computer Society, pp. 117{126.

[117] NIKOLOV, A. Personal communication. Clari cation, 2020.

[118] NIKOLOV, A., TALWAR, K., AND ZHANG, L. The geometry of di erential privacy: the sparse and approximate cases. In *Symposium on Theory of Computing Conference, STOC* (2013).

[119] PAGH, R., AND STAUSHOLM, N. M. E cient di erentially private F_0 linear sketching. In *24th International Conference on Database Theory, ICDT* (2021), vol. 186 of *LIPIcs*, pp. 18:1{18:19.

[120] Pagh, R., Stausholm, N. M., and Thorup, M. Hardness of bichromatic closest pair with jaccard similarity. In *27th Annual European Symposium on Algorithms, ESA* (2019), vol. 144 of *LIPIcs*, pp. 74:1{74:13.

[121] Public Information Office, US Census Bureau. Census bureau sets key parameters to protect privacy in 2020 census results. https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html. Published: 2021-06-09.

[122] Qardaji, W. H., Yang, W., and Li, N. Priview: practical di erentially private release of marginal contingency tables. In *International Conference on Management of Data, SIGMOD* (2014), pp. 1435{1446.

[123] Rob Bonta: State of California Department of Justice. California consumer privacy act (ccpa). https://oag.ca.gov/privacy/ccpa.

[124] Rubinstein, B. I. P., Bartlett, P. L., Huang, L., and Taft, N. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *J. Priv. Con dentiality 4*, 1 (2012).

[125] Shi, E., Chan, T. H., Rieffel, E. G., Chow, R., and Song, D. Privacy-preserving aggregation of time-series data. In *Distributed System Security Symposium, NDSS* (2011).

[126] Snowden, E. Reddit comment. https://www.reddit.com/r/IAmA/comments/36ru89/just_days_left_to_kill_mass_surveillance_under/crglgh2/. Published: 2015-05-21.

[127] Sparka, H., Tschorsch, F., and Scheuermann, B. P2KMV: A privacy-preserving counting sketch for e cient and accurate set intersection cardinality estimations. *IACR Cryptology ePrint Archive 2018* (2018), 234.

[128] Spielman, D. A., and Srivastava, N. Graph sparsi cation by e ective resistances. *SIAM J. Comput. 40*, 6 (2011), 1913{1926.

[129] Stanojevic, R., Nabeel, M., and Yu, T. Distributed cardinality estimation of set operations with di erential privacy. In *IEEE Symposium on Privacy-Aware Computing, PAC* (2017), pp. 37{48.

[130] Stausholm, N. M. Improved di erentially private euclidean distance approximation. In *40th Symposium on Principles of Database Systems, PODS* (2021), pp. 42{56.

[131] Stephen Castle, T. N. Y. T. Tv message by snowden says privacy still matters. https://www.nytimes.com/2013/12/26/world/europe/snowden-christmas-message-privacy.html. Published: 12-25-2013.

[132] Sweeney, L. Simple demographics often identify people uniquely. *Health (San Francisco) 671*, 2000 (2000), 1{34.

[133] Sweeney, L. Only you, your doctor, and many others may know. *Technology Science 2015092903*, 9 (2015), 29.

[134] Tan, P., Steinbach, M. S., and Kumar, V. *Introduction to Data Mining*. Addison-Wesley, 2005.

[135] Tschorsch, F., and Scheuermann, B. An algorithm for privacy-preserving distributed user statistics. *Computer Networks 57*, 14 (2013), 2775{2787.

[136] Upadhyay, J. Randomness e cient fast-Johnson-Lindenstrauss transform with applications in di erential privacy and compressed sensing. *arXiv preprint arXiv:1410.2470* (2014).

[137] Vadhan, S. P. The complexity of di erential privacy. In *Tutorials on the Foundations of Cryptography*. Springer, 2017, pp. 347{450.

[138] VON VOIGT, S. N., AND TSCHORSCH, F. Rrtxfm: Probabilistic counting for di erentially private statistics. In *Workshop on Trust and Privacy Aspects of Smart Information Environments (TPSIE)* (2019).

[139] WAGH, S., HE, X., MACHANAVAJJHALA, A., AND MITTAL, P. Dp-cryptography: marrying di erential privacy and cryptography in emerging applications. *Commun. ACM 64*, 2 (2021), 84{93.

[140] WANG, T., ZHANG, X., FENG, J., AND YANG, X. A comprehensive survey on local di erential privacy toward data statistics and analysis. *Sensors 20*, 24 (2020), 7030.

[141] WARNER, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association 60*, 309 (1965), 63{69.

[142] WIKIPEDIA. Arete. https://en.wikipedia.org/wiki/Arete.

[143] WILSON, R. J., ZHANG, C. Y., LAM, W., DESFONTAINES, D., SIMMONS-MARENGO, D., AND GIPSON, B. Di erentially private SQL with bounded user contribution. *Proceedings on Privacy Enhancing Technologies 2020*, 2 (2020), 230{250.

[144] WOODRUFF, D. P. Data streams and applications in computer science. *Bulletin of the EATCS 114* (2014).

[145] WOODRUFF, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science 10*, 1-2 (2014), 1{157.

[146] XIONG, X., LIU, S., LI, D., CAI, Z., AND NIU, X. A comprehensive survey on local di erential privacy. *Secur. Commun. Networks 2020* (2020), 8829523:1{8829523:29.

[147] XU, C., REN, J., ZHANG, Y., QIN, Z., AND REN, K. Dppro: Di erentially private high-dimensional data release via random projection. *IEEE Transactions on Information Forensics and Security 12*, 12 (2017), 3081{3093.

[148] YANG, M., LYU, L., ZHAO, J., ZHU, T., AND LAM, K. Local di erential privacy and its applications: A comprehensive survey. *CoRR abs/2008.03686* (2020).

[149] YAO, A. C. Protocols for secure computations (extended abstract). In *Symposium on Foundations of Computer Science, FOCS* (1982), pp. 160{164.

[150] ZHANG, J., CORMODE, G., PROCOPIUC, C. M., SRIVASTAVA, D., AND XIAO, X. Privbayes: private data release via bayesian networks. In *International Conference on Management of Data, SIGMOD* (2014), pp. 1423{1434.