Noiseless Database Privacy

Raghav Bhaskar
Microsoft Research India
rbhaskar@microsoft.com

Abhishek Bhowmick[*]
University of Texas at Austin
bhowmick@cs.utexas.edu

Vipul Goyal
Microsoft Research India
vipul@microsoft.com

Srivatsan Laxman
Microsoft Research India
slaxman@microsoft.com

Abhradeep Thakurta[†]
Pennsylvania State University
azg161@cse.psu.edu

June 14, 2012

Abstract

The notion of differential privacy has recently emerged as a gold standard in the field of database privacy. While this notion has the benefit of providing concrete theoretical privacy (compared to various previous ad-hoc approaches), the major drawback is that the mechanisms needs to inject some noise the output limiting its applicability in many settings.

In this work, we initiate the study of a new notion of privacy called *noiseless privacy*. The (very natural) idea we explore is to exploit the entropy already present in the database and substitute that in the place of external noise to the output. The privacy guarantee we provide is very similar to DP but where that guarantee "comes from" is very different in the two cases. While differential privacy focusses on generality, we make assumptions about the database distribution, the auxiliary information which the adversary may have and the type of queries. This allows us to obtain "privacy for free" whenever the underlying assumptions are satisfied.

In this work, we first formalize the notion of noiseless privacy, introduce two definitions and show that they are equivalent. We then study certain types of boolean and real queries and show natural (and well understood) conditions under which noiseless privacy can be obtained with good parameters. We also study the issue of composability and introduce models under which it can be achieved in the noiseless privacy framework.

1 Introduction

In the recent past, the field of data privacy has gone through a sea of change. Starting with techniques like removal of Personally identifiable information (PII), to sophisticated anonymization methods like k-anonymity [Swe02], ℓ-diversity [MGKV06], there has always been a tension between different privacy models and breaches against them. Almost all of these ad-hoc notions of privacy are presently known to be vulnerable to various easy to implement practical attacks [GKS08]. Dwork *et al.* [DMNS06] came up with a rigorous notion of privacy called *differential privacy*. This notion was one of the first few to provably guarantee privacy under worst case adversarial setting. Any algorithm (mechanism) that preserves differential privacy (DP), does it by injecting controlled randomness (noise) in the output. The definition is as below:

Definition 1 (ϵ-differential privacy or ϵ-DP [DMNS06])**.** *A randomized algorithm $\mathcal{A}$ is said to be ϵ-differentially private (or is said to satisfy ϵ-DP) if for all databases $T, T' \in \mathcal{D}^n$ differing in at most one record and all events $\mathcal{O} \subseteq Range(\mathcal{A})$,* $\Pr[\mathcal{A}(T) \in \mathcal{O}] \leq e^{\epsilon} \Pr[\mathcal{A}(T') \in \mathcal{O}]$.

Although this is an extremely strong notion of privacy, a draw back being that in many practical scenarios differentially private algorithms inject too much of noise in the output and as a result degrades the utility of the output sharply [MKA$^+$08]. Several applications, in fact, completely breakdown upon addition of any amount of noise to the output (see the *financial audit* example below). Besides, if the database curator has to give a noisy output, both

[*] Work done in part while visiting Microsoft Research, India.

[†] Work done in part while visiting Microsoft Research, India.

the mechanism to generate the output as well as the one which *consumes* the output has to be reanalyzed and reprogrammed[1] Hence the fact that addition of some noise to the output is inherent in DP is argued to be a major barriers towards the adoption of DP in practice.

Financial Audits: Consider a company going under a financial audit. The auditor, in order to verify the quarterly report, would like to know the total expenditure under the head "employee salary". In other words, the auditor request for the sum-total salary of all the employees of the company. In order to preserve privacy of each employee, say the company adds noise to the sum-total salary and then outputs the total salary (call it $\tilde{X}$). Now, let Y be the actual total amount spent by the company for employee salaries. Since $\tilde{X}$ is a noisy output, with high probability $|Y - \tilde{X}| \neq 0$. From the perspective of an auditor, it would not be clear whether the non-zero gap between Y and $\tilde{X}$ is because of the injected noise or there was some fraud w.r.t. the money spent for salaries. In such cases adding any amount of noise to the output $\tilde{X}$ is completely unacceptable.

In this work, we ask the following natural question: *"Is it always necessary to add noise the output to achieve provable privacy guarantees? Are there any scenarios where one can do without adding any noise?"*

Our Notion – Noiseless Privacy. Towards answering the above question, we initiate the study of a new notion of privacy called *noiseless privacy*. The (very natural) idea we explore is to exploit the entropy already present in the database and substitute that in the place of external noise to the output. At a high level, our definition is as follows. Consider an entry t_i in the database and two possible values a and b which it can take. We simply require that the probability of the output (or the vector of outputs in case of multiple queries) lying in a certain measurable set remains similar in both of these cases. Here, the probability is taken over the choice of the database (coming from a certain distribution) and is conditioned on the auxiliary information (present with the adversary) about the database. See Definition 2 for formal details.

The privacy guarantee we provide is very similar to DP but where that guarantee "comes from" is very different in the two cases. Differential Privacy focusses on attaining generality and works in a large number of scenarios (the data could be coming from any distribution, the adversary may have any auxiliary information about the database, etc). However to achieve the required guarantees, DP requires addition of external noise to the output. The philosophy of this work is quite different from DP:

- The key "ground rule" we follow in our work is that the output is always given correctly; there is no external noise added to the query output.
- We are willing to make strong (but still realistic) assumptions about the databases: in particular we make assumptions about the distribution from which the entries are coming from and that a large number of entries are independent of each other.
- We are willing to restrict the class of queries we can handle: for example a particular class of boolean functions, linear queries over reals, etc.
- We are willing to make assumptions about the adversary: in particular that the adversary has limited or no auxiliary information about the database in question.

At this point, we do not know how widely our privacy framework will be applicable in real systems. However whenever privacy can be obtained in our framework, it comes for "free". Our work shows that there are significant non-trivial classes of systems where our framework is applicable to obtained privacy. We believe that the gains made by making the assumptions required by our framework are quite significant:

- The output can be provided without any loss of utility, there are no tradeoffs to be made between the privacy parameter and the utility. Whatever privacy comes, it comes for free.
- Only an analysis to "okay the system" is required and whenever such an analysis goes through, no changes need to be made to the mechanism to answer the queries. Hence, existing applications will work as they are without any additional investment to change the mechanism (and how the output is handled) etc.

[1] For example, consider a mechanism which takes various inputs and produces various outputs. If all inputs are noisy, the outputs may not now satisfy a property which they earlier did.

Given the growing importance of data privacy, we believe that a fruitful approach towards obtaining data privacy will be the following two step approach:

1. First check if the conditions for obtaining noiseless privacy are satisfied by the system. If so, privacy is automatic and nothing needs to be done.
2. If not, proceed to obtain privacy via alternate "explicit" means such as injection of noise to the output. Thus, we *pay for privacy only if strictly necessary*.

Thus, we believe it is important to understand and develop techniques to analyze when privacy comes automatically. The current work represents an initial step in that direction. We stress that our results should not be seen as "end results" by any means and much work remains to be done in this direction.

Our Results. In this work, we study certain types of boolean and real queries and show natural (and well understood) conditions under which noiseless privacy can be obtained with good parameters. We first focus on boolean setting; i.e., the entries of the database as well as the query output have one bit of information each. For simplicity, our starting assumption is that each bit of the database is independently drawn from the uniform distribution (this assumption can be partially relaxed; see Section 3). We show that functions which are sufficiently "far" away from both 0-junta and 1-junta functions[2] satisfy noiseless privacy with "good" parameters. Note that functions which are close to either 0-junta or 1-junta do not represent an "aggregate statistics" of the database (which should depend on a large number of database entries). Hence, in real systems releasing some aggregate information about the database, we do expect such a condition to be naturally satisfied. Our proof of this theorem is rather intuitive and interestingly shows that these two (well understood) characteristics of the boolean functions are the only ones on which the privacy parameter depends.

For functions over the reals with real outputs, we study two types of functions: (a) linear functions (i.e., where the output is a linear combination of the rows of the database), and, (b) sum of arbitrary functions of the database rows. These functions together cover a large class of aggregate information to support various data mining algorithms used in real systems. We show natural conditions on the database distribution for which noiseless privacy can be obtained with good parameters. We refer the reader to section 4.1 for more details.

Multiple Queries. The above results are only for the case where the adversary is allowed to ask a single query. In the noiseless privacy framework, composition is tricky and privacy can completely breakdown even given a response to two different (carefully crafted) queries. This is in contrast to the differential privacy framework which provides a composition theorem showing that privacy does not completely break down with multiple queries (although it still degrades *exponentially*). The reason why such a composition is difficult to obtain in our setting is the lack of independence between the responses to the queries; the queries operate on the same database and might have complex interdependence on each other to enable an entry of the database to be deduced fully give the responses. In differential privacy, such an interdependence is broken by means of adding (independent) external noise to the output.

To break such interdependence in our setting, we introduce what we call the changing database model; we assume that between any two queries, a nontrivial fraction of the database has been "refreshed". The newly added entries (which may either replace some existing entries or be in addition to the existing entries) are independent of the old entries already present in the database. The effect of these new entries will reflect in the response to a given query. This helps us maintain some weak independence between different responses as in differential privacy thus providing us something to rely on while trying to obtain composability. We note that the setting of the changing database model is quite realistic. We present one real world example below.

HR Salary Surveys: A standard practice to determine the budget for pay raises in most large organizations is the conduct what is called an yearly HR salary survey. As part of this survey, the organization would submit relevant statistics about the salaries of its employees to some market research firms. These firms in turn aggregate data received from all organizations and generate salary statistics for the entire industry (which in turn forms the primary basis for the organizations to decide the salary budget for the upcoming year). A key requirement when an organization submits such statistics about the employee salaries to the market research firms is to maintain anonymity and hide all personally identifiable information about its employees (and only give salary statistics based on the department, years of experience, etc). A reasonable assumption in this setting is that a constant fraction of the employees will change

[2] Roughly, an i-junta function is one which depends only upon i of the total input variables.

every year (i.e., if the attrition rate of a firm is five percent, then roughly five percent of the entries can be expected to be refreshed every year).

Apart from the above example, there are various other scenarios where the changing database model is realistic (i.e., when one is dealing with streaming data, data with a time window, etc). Under such changing database model, we provide generalizations of our boolean as well as real query theorems to the case of multiple queries.

Other than the above, we also have other interesting results like obtaining noiseless privacy for symmetric boolean functions, "decomposable" function, etc. In various cases, we in fact show positive results for noiseless privacy under multiple queries even in the *static database* model.

Future Work. Our works opens up an interesting direction for research in the area of database privacy. Our attempt to understand what type of queries and database distributions inherently provide privacy guarantees is far from comprehensive. One can ask about what type of boolean function classes support multiple queries while maintaining good privacy parameters. What other real value functions support good privacy for the case of single as well as multiple queries? In the present work, much of our results are with independence assumption on the database generating distribution. A possible direction of research can be on how to show privacy for databases with correlated entries.

Related Works. The line of works most related to our work is that of *query auditing* (see Kenthapadi *et al.* [KMN05] and Nabar *et al.* [NMK$^+$06]). Roughly, in this line of work there is a *query auditor* which decides for a database $T = \langle t_1, \cdots, t_n \rangle$ with real entries, whether to answer a particular query or not. If the auditor decides to answer the query, then the answer to the query is output without adding any noise. Since the decision of whether to answer a query or not in itself can leak information about the database, the decision is itself randomized. So, in essence it is possible that a query is not answered when ideally it should have been answered and vice versa. The cases where the query is rejected and the randomized decision of when to query a query can be viewed as some form of noise injection into the system. However, the good aspect of query auditing is that if an answer is output, it is without any noise which is in harmony with the motivation of our present work. See appendix for a more detailed comparison of our work to this and other related works.

2 Our privacy notion

In our present work, we investigate the possibility of guaranteeing privacy without adding any external noise. The main idea is to look for (and systematically categorize) query functions which under certain assumptions on the data generating distribution are inherently private (under our formal notion of privacy that we define shortly). Since, the output of the function itself is inherently private, there is no need to inject external noise. As a result the output of the function has no utility degradation. Formally, we define our new notion of privacy (called *Noiseless Privacy*) as follows:

Definition 2 (ϵ-Noiseless Privacy or ϵ-NP)**.** *Let $\mathcal{D}$ be the domain from which the entries of the database are drawn. A deterministic query function $f : \mathcal{D}^n \rightarrow \mathcal{Y}$ is said to be ϵ-noiseless private (or is said to satisfy ϵ-NP) under a distribution D on $\mathcal{D}^n$ and some auxiliary information $\mathcal{A}ux$ (which the adversary might have), if for all measurable sets $\mathcal{O} \subseteq \mathcal{Y}$, for all $\ell \in [n]$ and for all $a, a' \in \mathcal{D}$,*

$$\Pr_{T \sim D}[f(T) \in \mathcal{O} | t_\ell = a, \mathcal{A}ux] \leq e^\epsilon \Pr_{T \sim D}[f(T) \in \mathcal{O} | t_\ell = a', \mathcal{A}ux]$$

where t_ℓ is the ℓ-th entry of the database T.

In comparison to Differential Privacy (see Definition 1), our notion of Noiseless Privacy (Definition 2) differs at least in the following three aspects:

- unlike in Definition 1, it is possible for a non-trivial deterministic function f to satisfy Definition 2 with reasonable ϵ. For *e.g.,* XOR of all the bits of a boolean database (where each entry of the database is an unbiased random bit) satisfies Definition 2 with $\epsilon = 0$ where as Definition 1 is not satisfied for any finite ϵ.

- the privacy guarantee of Definition 2 is under a specific distribution D, where as Definition 1 is agnostic to any distributional assumption on the database.

- the privacy guarantee of Definition 2 is w.r.t. an auxiliary information $\mathcal{A}ux$ whereas differential privacy is oblivious to auxiliary information.

Intuitively, the above definition captures the change in adversary's belief about a particular output in the range of f in the presence or absence of a particular entry in the database. A comparable (and seemingly more direct) notion is to capture the change in adversary's belief about a particular entry before and after seeing the output. This is stated formally in the next definition.

Definition 3 (ϵ-Aposteriori Noiseless Privacy or ϵ-ANP)**.** *A deterministic query function $f : \mathcal{D}^n \to \mathcal{Y}$ is said to be ϵ-Aposteriori Noiseless Private (or is said to satisfy ϵ-ANP) under a distribution D on $\mathcal{D}^n$ and some auxiliary information $\mathcal{A}ux$, if for all measurable sets $\mathcal{O} \subseteq \mathcal{Y}$, for all $\ell \in [n]$ and for all $a \in \mathcal{D}$,*

$$e^{-\epsilon} \leq \frac{\Pr_{T\sim D}[t_\ell = a | f(T) \in \mathcal{O}, \mathcal{A}ux]}{\Pr_{T\sim D}[t_\ell = a | \mathcal{A}ux]} \leq e^{\epsilon}$$

where t_ℓ is the ℓ-th entry of the database T.

The following lemma shows that Definition 3 implies Definition 2 and vice versa with a two-fold degradation in privacy parameter ϵ in one direction.

Lemma 1. *Let $f : \mathcal{D}^n \to \mathcal{Y}$ be a deterministic query function. Given a distribution, D, over $\mathcal{D}^n$, and some auxiliary information, $\mathcal{A}ux$, the following hold:*

1. *If f satisfies ϵ-NP under D and $\mathcal{A}ux$, then it also satisfies 2ϵ-ANP under the same D and $\mathcal{A}ux$.*
2. *If f satisfies ϵ-ANP under D and $\mathcal{A}ux$, then it is also ϵ-NP under the same D and $\mathcal{A}ux$.*

Proof. Let us first prove Definition 3$\Rightarrow$ Definition 2. For all measurable $\mathcal{O} \subseteq Range(f)$, for all $\ell \in [n]$ (where n is the number of rows in the database T) and for all $a \in \mathcal{D}$, the following is true.

$$\Pr[f(T) \in \mathcal{O} | t_\ell = a, \mathcal{A}ux] = \frac{\Pr[t_\ell = a | f(T) \in \mathcal{O}, \mathcal{A}ux] \Pr[f(T) \in \mathcal{O} | \mathcal{A}ux]}{\Pr[t_\ell = a | \mathcal{A}ux]}$$

Now, $\frac{\Pr[f(T)\in\mathcal{O}|t_\ell = a, \mathcal{A}ux]}{\Pr[f(T)\in\mathcal{O}|t_\ell = a', \mathcal{A}ux]}$ (where $a, a' \in \mathcal{D}$ and $a \neq a'$) can be written as

$$\begin{aligned}
&\frac{\Pr[t_\ell = a | f(T) \in \mathcal{O}, \mathcal{A}ux] \Pr[f(T) \in \mathcal{O} | \mathcal{A}ux]}{\Pr[t_\ell = a | \mathcal{A}ux]} \cdot \frac{\Pr[t_\ell = a' | \mathcal{A}ux]}{\Pr[t_\ell = a' | f(T) \in \mathcal{O}, \mathcal{A}ux] \Pr[f(T) \in \mathcal{O} | \mathcal{A}ux]} \\
&= \frac{\Pr[t_\ell = a | f(T) \in \mathcal{O}, \mathcal{A}ux]}{\Pr[t_\ell = a | \mathcal{A}ux]} \cdot \frac{\Pr[t_\ell = a' | \mathcal{A}ux]}{\Pr[t_\ell = a' | f(T) \in \mathcal{O}, \mathcal{A}ux]}
\end{aligned}$$

Now, $\frac{\Pr[t_\ell = a|f(T)\in\mathcal{O},\mathcal{A}ux]}{\Pr[t_\ell = a|\mathcal{A}ux]} \leq e^{\epsilon}$ and $\frac{\Pr[t_\ell = a'|\mathcal{A}ux]}{\Pr[t_\ell = a'|f(T)\in\mathcal{O},\mathcal{A}ux]} \leq e^{\epsilon}$ by Definition 3. Therefore,

$$\begin{aligned}
\frac{\Pr[f(T) \in \mathcal{O} | t_\ell = a, \mathcal{A}ux]}{\Pr[f(T) \in \mathcal{O} | t_\ell = a', \mathcal{A}ux]} &\leq e^{2\epsilon} \\
\Rightarrow \Pr[f(T) \in \mathcal{O} | t_\ell = a, \mathcal{A}ux] &\leq e^{2\epsilon} \Pr[f(T) \in \mathcal{O} | t_\ell = a', \mathcal{A}ux]
\end{aligned}$$

The proof in the other direction goes as follows. We want to upper and lower bound the ratio $\frac{\Pr[t_\ell = a|f(T)\in\mathcal{O},\mathcal{A}ux]}{\Pr[t_\ell = a|\mathcal{A}ux]}$. Equivalently, we have,

$$\begin{aligned}
\frac{\Pr[t_\ell = a | f(T) \in \mathcal{O}, \mathcal{A}ux]}{\Pr[t_\ell = a | \mathcal{A}ux]} &= \frac{\Pr[f(T) \in \mathcal{O} | t_\ell = a, \mathcal{A}ux] \Pr[t_\ell = a | \mathcal{A}ux]}{\Pr[f(T) \in \mathcal{O} | \mathcal{A}ux] \Pr[t_\ell = a | \mathcal{A}ux]} \\
&= \frac{\Pr[f(T) \in \mathcal{O} | t_\ell = a, \mathcal{A}ux]}{\Pr[f(T) \in \mathcal{O} | \mathcal{A}ux]} \\
&= \frac{\Pr[f(T) \in \mathcal{O} | t_\ell = a, \mathcal{A}ux]}{\sum_{a' \in \mathcal{D}} \Pr[f(T) \in \mathcal{O} | t_\ell = a', \mathcal{A}ux] \Pr[t_\ell = a' | \mathcal{A}ux]} \\
&= \frac{1}{\sum_{a' \in \mathcal{D}} \frac{\Pr[f(T)\in\mathcal{O}|t_\ell = a', \mathcal{A}ux]}{\Pr[f(T)\in\mathcal{O}|t_\ell = a, \mathcal{A}ux]} \Pr[t_\ell = a' | \mathcal{A}ux]}
\end{aligned}$$

This ratio is upper and lower bounded by e^{ϵ} and $e^{-\epsilon}$ respectively because each of $\frac{\Pr[f(T)\in\mathcal{O}|t_\ell = a', \mathcal{A}ux]}{\Pr[f(T)\in\mathcal{O}|t_\ell = a, \mathcal{A}ux]}$ is upper and lower bounded due to Definition 2 and the fact $\sum_{a' \in \mathcal{D}} \Pr[t_\ell = a' | \mathcal{A}ux] = 1$. □

Finally, we introduce a relaxed notion of Noiseless Privacy called (ϵ, δ)-Noiseless Privacy, where with a small probability δ the ϵ-Noiseless Privacy does not hold. Here, the probability is taken over the choice of the database and the two possible values for the database entry in question. While for a strong privacy guarantee a negligible δ is desirable, a non-negligible δ may be tolerable in certain applications. The following definition captures this notion formally.

Definition 4 ((ϵ, δ)-Noiseless Privacy)**.** *Let $f : \mathcal{D}^n \to \mathcal{Y}$ be a deterministic query function on a database of length n drawn from domain $\mathcal{D}$. Let D be a distribution on $\mathcal{D}^n$. Let $S_1 \subseteq \mathcal{Y}$ and $S_2 \subseteq \mathcal{D}$ be two sets such that for all $j \in [n]$, $\Pr_{T\sim D}[f(T) \in S_1] + \Pr_{T\sim D}[t_j \in S_2] \leq \delta$, where t_j is the j-th entry of T.*

The function f is said to be (ϵ, δ)-Noiseless Private under distribution D and some auxiliary information $\mathcal{A}ux$, if there exists S_1, S_2 as defined above such that, for all measurable sets $\mathcal{O} \subseteq \mathcal{Y} - S_1$, for all $a, a' \in \mathcal{D} - S_2$, and for all $\ell \in [n]$ the following holds:

$$\Pr_{T\sim D}[f(T) \in \mathcal{O} | t_\ell = a, \mathcal{A}ux] \leq e^\epsilon \Pr_{T\sim D}[f(T) \in \mathcal{O} | t_\ell = a', \mathcal{A}ux]$$

Composability. In many applications, privacy has to be achieved under multiple (partial) disclosures of the database. For instance, in database applications, several thousand user queries about the database entries are answered in a day. Thus, a general result which tells how the privacy guarantee changes (typically degrades) as more and more queries are answered is very useful and is referred to as *composability* of privacy under multiple queries. While in some scenarios (eg. streaming applications) the database can change in between queries (dynamic database), in other scenarios it remains the same (static database). Also, the queries can be of different types or multiple instances of the same type. As mentioned earlier, while the DP definition admits a general composition result, the privacy guarantee degrades exponentially with the number of queries on a static database.

In case of NP, obtaining general composition results is a challenge because the data distribution can change somewhat intractably when conditioned on the query responses returned so far. Moreover, if the data distribution allowed *dependence* among data entries, then revealing exact answers to deterministic queries involving some entries, can easily reveal information about other (dependent) entries. Hence we explore the possibility of composition under independence assumptions. It turns out that, even with independence, not all query-sequences compose under NP. We provide one interesting example of the kinds of query-sequences for which composition can be achieved under independence assumption on the data.

Consider data entries, $t_j, j = 1, 2, \ldots$, drawn *iid* according to some distribution (over some fixed domain $\mathcal{D}$), and let $\{T_1, \ldots, T_m\}$ denote a sequence of subsets of $\{t_j\}$. We now study the composition of a sequence of queries $f_i(T_i)$, $i = 1, \ldots, m$ (Notice that each $f_i(\cdot)$ is a function of only the corresponding subset T_i of $\{t_j\}$). If T_i's do not share any elements at all, then composition is trivial (given the independence of $\{t_i\}$) assuming each individual f_i satisfies noiseless privacy (with the suitable parameters). Even when T_i's contain common or shared elements, composition is easy if we only required guarantees for the unique elements (i.e., for those that participate in only one query). This is because, entries participating in previous queries can be regarded as auxiliary information and if we can obtain NP under leakage of suitable auxiliary information, composition of NP for the unique entries follows. However, we obviously require composition of NP for all entries in the database (not just the unique entries). Our composition result below shows that this is possible if each T_i has "sufficiently many" unique entries.

Theorem 1 (Composition)**.** *Let $t_j, j = 1, \ldots,$ denote the data entries drawn* iid *according to distribution D (over the data domain $\mathcal{D}$). Let $T_i, i = 1, \ldots, m,$ denote a sequence of subsets of $\{t_j\}$. Consider a sequence of m queries $f_i(T_i), i = 1, \ldots, m,$ where the i^{th} query operates on the i^{th} data set T_i. Let T^* denote the set of all data entries that appear in two or more data sets, i.e. $T^* = \{t : t \in T_i \cap T_j,\ i \neq j\}$. If each $f_i(T_i)$ is (ϵ_i, δ_i) noiseless private under auxiliary disclosure of $|T_i \cap T^*|$ entries of T_i, then the sequence of queries $f_i(T_i), i = 1, \ldots, m$ is $(\sum_{i=1}^m \epsilon_i, \sum_{i=1}^m \delta_i)$ noiseless private.*

Proof. To assess the privacy of the ℓ^{th} element t_ℓ, we consider the joint probability of observing the m query responses given $[t_\ell = a]$ and compare it with the corresponding probability for $[t_\ell = a']$. Let $\mathcal{A}_f$ denote the collection of random variables in $(T^* \setminus \{t_\ell\})$. By marginalizing over all possible value assignments to the entries in $\mathcal{A}_f$, we can write the joint probability of the m query responses given $[t_\ell = a]$ as follows:

$$\begin{aligned}
&\Pr_{T\sim D}[f_1(T_1) \in \mathcal{O}_1, \ldots, f_m(T_m) \in \mathcal{O}_m \mid t_\ell = a] \\
&\qquad = \int_{\mathcal{A}_f} dF(f_1(T_1) \in \mathcal{O}_1, \ldots, f_m(T_m) \in \mathcal{O}_m, \mathcal{A}_f \mid t_\ell = a) \\
&\qquad = \int_{\mathcal{A}_f} \Pr_{T\sim D}[f_1(T_1) \in \mathcal{O}_1, \ldots, f_m(T_m) \in \mathcal{O}_m \mid t_\ell = a, \mathcal{A}_f] \times dF(\mathcal{A}_f) \qquad (1)
\end{aligned}$$

The last expression above, namely (1), follows from the independence of $\mathcal{A}_f$ and t_ℓ. By factorizing the conditional probability in (1) we obtain

$$\begin{aligned}
&\Pr_{T\sim D}[f_1(T_1) \in \mathcal{O}_1, \ldots, f_m(T_m) \in \mathcal{O}_m \mid t_\ell = a] \\
&\qquad = \int_{\mathcal{A}_f} \prod_{i=1}^{m} \Pr_{T\sim D}[f_i(T_i) \in \mathcal{O}_i \mid t_\ell = a, \mathcal{A}_f] \times dF(\mathcal{A}_f) \qquad (2) \\
&\qquad = \int_{\mathcal{A}_f} \prod_{i=1}^{m} \Pr_{T\sim D}[f_i(T_i) \in \mathcal{O}_i \mid t_\ell = a, (T^* \cap T_i)] \times dF(\mathcal{A}_f) \qquad (3)
\end{aligned}$$

where (2) is obtained by observing that the event $[f_1(T_1) \in \mathcal{O}_1, \ldots, f_{i-1}(T_{i-1}) \in \mathcal{O}_{i-1}]$ can reveal no more about the entries in T_i than $[t_\ell = a, \mathcal{A}_f]$ put-together (Note that this again requires the independence between data entries as per the data generating distribution D). The next expression (3) then follows since, again, elements in $\mathcal{A}_f$ outside of $(T^* \cap T_i)$ are independent of T_i. Under auxiliary leakage of $|T^* \cap T_i|$ elements of T_i we are given that $f_i(T_i)$ is (ϵ_i, δ_i)-Noiseless Private, i.e., there exist appropriate sets S_1^i and S_2^i (see *Definition 4*) with $\Pr_{T\sim D}[f_i(T_i) \in S_1^i] + \Pr_{T\sim D}[t_j \in S_2^i] \leq \delta_i$ such that, for all measurable sets $\mathcal{O}_i \subseteq \mathcal{Y}_i - S_1^i$, for all $a, a' \in \mathcal{D} - S_2^i$, we have

$$\begin{aligned}
&\Pr_{T\sim D}[f_i(T_i) \in \mathcal{O}_i \mid t_\ell = a, (T^* \cap T_i)] \\
&\leq e^{\epsilon_i} \Pr_{T\sim D}[f_i(T_i) \in \mathcal{O}_i \mid t_\ell = a', (T^* \cap T_i)] \qquad (4)
\end{aligned}$$

Let S_1 be the set of all possible m responses where the i^{th} response lies in S_1^i for some i, $1 \leq i \leq m$, and let $S_2 = \cup_{i=1}^{m} S_2^i$. Using these constructions of S_1 and S_2, and denote the vector of responses, $[f_1(T_1), \ldots, f_m(T_m)]$ as $f(T)$ we get

$$\Pr_{T\sim D}[f(T) \in S_1] + \Pr_{T\sim D}[t_j \in S_2] \leq \sum_{i=1}^{m} \delta_i \qquad (5)$$

and applying (4) for each of the m terms inside the product in (3) we get, for all measurable sets $\mathcal{O}_1 \times \cdots \times \mathcal{O}_m$, where each $\mathcal{O}_i \subseteq \mathcal{Y}_i - S_1^i$, for all $a, a' \in \mathcal{D} - S_2$,

$$\begin{aligned}
&\Pr_{T\sim D}[f_1(T_1) \in \mathcal{O}_1, \ldots, f_m(T_m) \in \mathcal{O}_m \mid t_\ell = a] \\
&\leq e^{\sum_{i=1}^{m} \epsilon_i} \Pr_{T\sim D}[f_1(T_1) \in \mathcal{O}_1, \ldots, f_m(T_m) \in \mathcal{O}_m \mid t_\ell = a'] \qquad (6)
\end{aligned}$$

This completes the proof. □

One way of gaining intuition behind this theorem is the following. In this theorem, we are interested in the privacy of entries appearing in multiple datasets T_i's. However if each dataset has some unique entries, the "effect" these unique entries have on the output can be compared to the noise added by a differentially private mechanism to the output. Besides, because of the independence property, the effect of unique entries in each query is independent. Hence, this can be viewed as similar to independent noise being added to each query by a differentially private mechanism. Then, the privacy of entries appearing in multiple datasets degrades similar to how the privacy degrades in differential privacy setting under multiple queries.

There are many practical settings where the T_i's satisfy the condition of our composition theorem. For example, dynamic databases arise in many scenarios: (a) Growing database model: Here the database keeps growing with time, *e.g.* database of all registered cars. Thus, in-between subsequent releases of information, the database grows by some number k, (b) Streaming model: This is the more commonly encountered scenario, where the availability of limited memory/storage causes the replacement of some old data with new one. Thus, at the time of each query the database has some k new entries out of the total (fixed) n , and (c) Random replacement model: A good generalization of the above two models, it replaces randomly chosen k entries from the database of size n with the new incoming entries. In all such cases, our composition results will apply.

In all the above models of dynamic databases, we assume that the number of new elements form a constant fraction of the database. In particular, if n is the current database size, then some $\rho n, (0 \leq \rho \leq 1)$ number of entries are old and the remaining $k = (1-\rho)n$ entries are new. Our main result about composability of Noiseless Privacy holds for any query which has (ϵ, δ)-Noiseless Privacy under any auxiliary information about at most $\rho n, (0 \leq \rho \leq 1)$ elements of the database. Note that in the growing database model, the size of the largest database on which the query is made is assumed to be n and the maximum fraction of old entries is ρ.

3 Boolean queries

In this section we study queries of the form $f : T \rightarrow \{0,1\}$, *i.e.*, the query function f acts on a database T and outputs a bit. In all the cases except section A.1, we consider that database $T \in \mathcal{D}^n$, where $\mathcal{D}$ is the domain from which the data entries are drawn. In section A.1 where we discuss the growing database model, we do not fix the length of the database, *i.e.*, the query f can be over a variable length database. We discuss later why such a relaxation is necessary.

3.1 The No Auxiliary Information Setting

We first study a simple and clean setting: the database entries are all drawn independently and the adversary has no auxiliary information about them. We discuss generalizations later on. Before we get into the details of privacy friendly functions under our setting, we need some of the terminologies from analysis of boolean functions literature.

Definition 5 (k-junta [KLM$^+$09]). *A function $f : \{0,1\}^n \rightarrow \{0,1\}$ is said to be k-junta if it depends only on some subset of the n coordinates of size k .*

Definition 6 ($(1-\tau)$-far from k-junta). *Let $\mathcal{F}$ be the class of all k-junta functions $f' : \{0,1\}^n \rightarrow \{0,1\}$ and let D be a distribution on $\{0,1\}^n$. A function $f : \{0,1\}^n \rightarrow \{0,1\}$ is $(1-\tau)$-far from k-junta under D if*

$$\max_{f' \in \mathcal{F}} \left| \Pr_{T \sim D}[f(T) = f'(T)] - \Pr_{T \sim D}[f(T) \neq f'(T)] \right| = \tau$$

An alternative way of stating the above condition is that for any such f', $|Pr[f(T) = f'(T)] - 1/2| \leq \frac{\tau}{2}$. That is, f' can predict f only with "advantage" at most $\frac{\tau}{2}$ (over a random guess). It is easy to see that when D is a uniform distribution over n-bits, a k-junta is 0-far from the class of k-juntas and the parity function is 1-far from the class of all 1-juntas.

Now, we are in a position to state our result in the boolean to boolean setting. The theorem below is for the setting where the adversary has no auxiliary information about the database. Later on in this section, we show how to handle the case when the adversary may have a subset of the database entries. We for clarity of exposition, we first state the theorem assuming that each bit of the database is individually random and generalize it later on.

Theorem 2. *Let D be an arbitrary distribution over $\{0,1\}^n$ such that the marginal probability of the i-th bit equaling 1 is $1/2$. Let $f : \{0,1\}^n \rightarrow \{0,1\}$ be a boolean function which is $(1-\tau_1)$-far from 0-junta and $(1-\tau_2)$-far from 1-junta under D. If $\tau_1 + \tau_2 < 1$, then f is $\ln \frac{1+(\tau_1+\tau_2)}{1-(\tau_1+\tau_2)}$-noiseless private.*

We now state and prove a more general version of the above theorem.

Theorem 3. *Let D be an arbitrary distribution over $\{0,1\}^n$ such that the marginal probability of the i-th bit equaling 1 is p_i. Let $f : \{0,1\}^n \rightarrow \{0,1\}$ be a boolean function which is $(1-\tau_1)$-far from 0-junta and $(1-\tau_2)$-far from 1-junta under D. If $\frac{\tau_1+\tau_2}{2} < \min_{i \in [n]} p_i$ and $\max_{i \in [n]} p_i < 1 - \frac{\tau_1+\tau_2}{2}$, then f is $\left(\max_{i \in [n]} \max \left\{ \ln \frac{1+(\tau_1+\tau_2)/(2(1-p_i))}{1-(\tau_1+\tau_2)/(2p_i)}, \ln \frac{1+(\tau_1+\tau_2)/(2p_i)}{1-(\tau_1+\tau_2)/(2(1-p_i))} \right\}\right)$-noiseless private. (Observe that the two terms for the max function are obtained by simply switching p_i by $1-p_i$.)*

Proof. In the following, we denote by T, a random instance of the database under distribution D and by t_i, the i-th bit of T. For a boolean function f to be $(1-\tau_1)$-far from 0-junta, it means that either $\Pr_T[f(T)=0]=1/2+\tau_1/2$ or $\Pr_T[f(T)=0]=1/2-\tau_1/2$ (as 0 and 1 are the only 0-junta functions). Similarly, for a function to be $1-\tau_2$ far from a 1-junta, it means that there exists $i\in[n]$ such that $\Pr_T[f(T)=t_i]=1/2-\tau_2/2$ or $\Pr_T[f(T)=\bar{t_i}]=1/2+\tau_2/2$.

Let us fix the following notation. Let $A=\Pr_T[f(T)=0]$. For any i, let $\frac{1}{2}+B_i=\Pr_T[f(T)=t_i]$. We note that $A\in\{1/2-\tau_1/2, 1/2+\tau_1/2\}$ and for all i, $B_i\in[-\tau_2/2,\tau_2/2]$.

One can write

$$\begin{aligned}\frac{1}{2}+B_i=Pr_T[f(T)=t_i] &= (1-p_i)\Pr_T[f(T)=0|t_i=0]+p_i\Pr_T[f(T)=1|t_i=1]\\ &= (1-p_i)\Pr_T[f(T)=0|t_i=0]+p_i(1-\Pr_T[f(T)=0|t_i=1])\end{aligned}$$

Again,

$$A=\Pr_T[f(T)=0]= (1-p_i)\Pr_T[f(T)=0|t_i=0]+p_i\Pr_T[f(T)=0|t_i=1]$$

Solving for $\Pr_T[f(T)=0|t_i=0]$ and $\Pr_T[f(T)=0|t_i=1]$ we get,

$$\begin{aligned}\Pr_T[f(T)=0|t_i=0] &= \frac{A+B_i+(1/2-p_i)}{2(1-p_i)}\\ \Pr_T[f(T)=0|t_i=1] &= \frac{A-B_i-(1/2-p_i)}{2p_i}\end{aligned}$$

Therefore, the ratio

$$\frac{\Pr_T[f(T)=0|t_i=0]}{\Pr_T[f(T)=0|t_i=1]}=\frac{A+B_i+(1/2-p_i)}{2(1-p_i)}\frac{2p_i}{A-B_i-(1/2-p_i)}$$

To obtain noiseless privacy guarantee we need to upper (lower) bound the above ratio. It can be verified that the above ratio is upper bounded by substituting the maximum value for B_i and by substituting the minimum value for A in the denominator and the maximum value in the numerator. This leads to (after some simple calculations) an upper bound of $\frac{1+(\tau_1+\tau_2)/(2(1-p_i))}{1-(\tau_1+\tau_2)/(2p_i)}$ which would make sense as long as the denominator is positive which is enforced by $\frac{\tau_1+\tau_2}{2}<p_i$.

To lower bound the above ratio, it suffices to upper bound $\frac{\Pr_T[f(T)=0|t_i=1]}{\Pr_T[f(T)=0|t_i=0]}$. Using similar argument as above it follows that this ratio is upper bounded by

$$\frac{1+(\tau_1+\tau_2)/(2p_i)}{1-(\tau_1+\tau_2)/(2(1-p_i))}$$

With this the proof is complete. □

Note that in the above theorem we do not assume independence among the entries in T. As a result we can handle databases with correlated entries. It is also worth mentioning here that all the other results in this section assume the entries in the database to be uncorrelated.

To get some more insight into the result let, us consider $f(T)$ to be the XOR of all the bits of T. Let T be drawn from the uniform distribution. Then f is 1-far from both a 0-junta and a 1-junta. Hence, f is 0-noiseless private. Instead of the XOR, if we let f be the AND function, then we see that it is just $1-\frac{1}{2^{n-1}}$-far from a 0-junta. The ratio in this case becomes ∞, which shows AND is not a very good function for providing ϵ-noiseless privacy for small ϵ. This is indeed the case because $\Pr_T[f(T)=1|t_i=0]=0$ for all i. However, we can capture functions like AND if we try to guarantee (ϵ,δ)-noiseless privacy. If we fix $\delta=\frac{1}{2^n}$ (which is basically the probability of the AND function yielding 1), we get $(0,\frac{1}{2^n})$-noiseless privacy for AND. This property is in fact not specific to AND. In fact one can easily guarantee (ϵ,δ)-noiseless privacy for any *symmetric boolean functions* (*i.e.,* the functions whose output does not change on any permutation of the input bits). We will discuss this result in a more general setting later in Section 3.4.

3.2 Handling Auxiliary Information

We now study the setting where the adversary may have auxiliary information about a subset of the entries in the database. We study the privacy of the entries about whom the adversary has no auxiliary information.

Theorem 4. *Let D be the distribution over $\{0,1\}^n$ where the i-th bit is chosen to be 1 independently with probability p_i. Let T be a database drawn from D. Let $f : \{0,1\}^n \rightarrow \{0,1\}$ be a boolean function which is $(1-\tau)$-far away from $d+1$ junta (that is, for any function g that depends only on a maximum of $d+1$ input variables, $|Pr[f(T) = g(T)] - 1/2| \leq \tau/2$). Let Γ be any adversarially chosen subset of variables that has been leaked with $|\Gamma| = d$. If $\frac{\tau}{\delta} < \min_{i\in[n]} p_i$ and $\max_{i\in[n]} p_i < 1 - \frac{\tau}{\delta}$, then f is $(\max_{i\in[n]-\Gamma}\left(\max\left\{\ln\left(\frac{1+\frac{\tau}{\delta(1-p_i)}}{1-\frac{\tau}{\delta p_i}}\right), \ln\left(\frac{1+\frac{\tau}{\delta p_i}}{1-\frac{\tau}{\delta(1-p_i)}}\right)\right\}\right), 2\delta)$-noiseless private with respect to the bit $t_i \in T$, where $i \in [n] - \Gamma$.*

Proof. We analyze the probability that the auxiliary information $\Gamma = t$ is such that $|Pr_R[f(R||t) = 0] - 1/2| \leq \tau/2\delta$ and $|Pr_R[f(R||t) = t_i] - 1/2| \leq \tau/2\delta$. We prove that this happens with probability at least $1 - \delta - \delta = 1 - 2\delta$. The proof is as follows. Here the notation $R||t$ refers to a database formed by combining R and t.

Lemma 2. *Let the underlying distribution be any D where each bit is 1 independently with probability p_i. Under D, let f be $(1-\tau)$-far away from d junta, that is for any function g that depends only on d variables, $|Pr_D[f(T) = g(T)] - 1/2| \leq \tau/2$. Here T is the database drawn from D and let Γ (with $|\Gamma| = d$) be any adversarial subset of entries of T that has been leaked. Then, with probability at least $1-\delta$ over the choice of assignments t to Γ, $|Pr_R[f(R||t) = 0] - 1/2| \leq \tau/2\delta$.*

Proof. Let $|\Gamma| = d$, be the set of indices leaked. Note that we use Γ to represent both the indices and the set of leaked variables. Let $R = [n] - \Gamma$. We prove the lemma by contradiction. Suppose the claim is wrong. That is, with probability more than δ over Γ, $|Pr_R[f(R||t) = 0] - 1/2| \geq \tau/2\delta$. Construct $g : \{0,1\}^d \rightarrow \{0,1\}$ as follows.

$$g(t) = \begin{cases} 0 & \text{if } Pr_R[f(R||t) = 0] \geq 1/2 \\ 1 & \text{otherwise} \end{cases}$$

Observe that g just depends on d variables. We shall now show predictability of f using g which contradicts farness from d junta. Let us evaluate $Pr[f(T) = g(\Gamma)]$. To that end, we partition the assignments t to T into three sets, S_1, S_2 and S_3. S_1 is the set of t such that $Pr_R[f(R||t) = 0] \geq 1/2 + \tau/2\delta$, S_2 is the set of t such that $Pr_R[f(R||t) = 0] \leq 1/2 - \tau/2\delta$ and S_3 is the set of remaining assignments. Now, from our assumption, we are given that $\Pr[T \in S_1 \cup S_2] > \delta$. Also, it is easy to observe that for any t, $Pr_R[f(R||t) = g(t)] \geq 1/2$ by the choice of g (since in all cases, g does at least as well as a random guess). Now, we lower bound $Pr[f(T) = g(\Gamma)]$.

$$\begin{aligned} Pr[f(T) = g(\Gamma)] &= \mathbb{E}_\Gamma Pr_R[f(R||\Gamma) = g(\Gamma)] \\ &\geq Pr[\Gamma \in S_1](1/2 + \tau/2\delta) + Pr[\Gamma \in S_2](1/2 + \tau/2\delta) + Pr[\Gamma \in S_3](1/2) \\ &\geq 1/2 + (\tau/2\delta)\Pr[\Gamma \in S_1 \cup S_2] \\ &> 1/2 + \tau/2 \end{aligned}$$

This leads to a contradiction.

□

Lemma 3. *Let D be a distribution over $\{0,1\}^n$ where each bit is 1 independently with probability p_i. Under D, let f be far away from $d+1$ junta. Let T be a database drawn from D and let Γ (with $|\Gamma| = d$) be any adversarial subset of entries of T that has been leaked. Then, with probability at least $1-\delta$ over the choice of assignments t to Γ, $|Pr_R[f(R||t) = t_i] - 1/2| \leq \tau/2\delta$, where t_i is the i-th entry of the database T.*

Proof. Let $|\Gamma| = d$, be the set of indices leaked. As before that we use Γ to represent both the indices and the set of leaked variables. Let $R = [n] - \Gamma$. We prove the lemma by contradiction. Suppose the claim is wrong. That is, with probability more than δ over Γ, $|Pr_R[f(R||t) = t_i] - 1/2| \geq \tau/2\delta$. Now, let $f' = f \oplus t_i$. Equivalently, with probability at least δ over Γ, $|Pr_R[f'(R||t) = 0] - 1/2| \geq \tau/2\delta$. The argument now is similar to the previous lemma owing to the transformation from f to f'. We nevertheless repeat it for completeness. Construct $g : \{0,1\}^d \rightarrow \{0,1\}$ as follows.

$$g(t) = \begin{cases} 0 & \text{if } Pr_R[f'(R||t) = 0] \geq 1/2 \\ 1 & \text{otherwise} \end{cases}$$

Observe that $g' = g \oplus t_i$ just depends on $d+1$ variables. We shall now show the predictability of f using g' which contradicts farness from $d+1$ junta. Let us evaluate $Pr[f(T) = g'(\Gamma)] = Pr[f'(T) = g(\Gamma)]$. To that end, we partition the assignments t to Γ into three sets, S_1, S_2 and S_3. S_1 is the set of t such that $Pr_R[f'(R||t) = 0] \geq 1/2 + \tau/2\delta$, S_2 is the set of t such that $Pr_R[f'(R||t) = 0] \leq 1/2 - \tau/2\delta$ and S_3 is the set of remaining assignments. Now, from our assumption, we are given that $\Pr[\Gamma \in S_1 \cup S_2] > \delta$. Also, it is easy to observe that for any t, $Pr_R[f'(R||t) = g(t)] \geq 1/2$ by the choice of g. Now, we lower bound $Pr[f(T) = g'(\Gamma)] = Pr[f'(T) = g(\Gamma)]$.

$$\begin{aligned} Pr[f'(T) = g(\Gamma)] &= \mathbb{E}_\Gamma Pr_R[f'(R||\Gamma) = g(\Gamma)] \\ &\geq \Pr[\Gamma \in S_1](1/2 + \tau/2\delta) + \Pr[\Gamma \in S_2](1/2 + \tau/2\delta) + \Pr[\Gamma \in S_3](1/2) \\ &\geq 1/2 + (\tau/2\delta)\Pr[\Gamma \in S_1 \cup S_2] \\ &> 1/2 + \tau/2 \end{aligned}$$

This leads to a contradiction. □

Following the proof structure in Theorem 3, let $N = Pr[f = 0|\Gamma = t, t_i = 0]$ and $D = Pr[f = 0|\Gamma = t, t_i = 1]$. Now,

$$(1 - p_i)N + p_i(1 - D) = 1/2 + B_i, \quad \text{where } |B_i| \leq \tau/2\delta$$
$$(1 - p_i)N + p_i D = A, \quad \text{where } |A - 1/2| \leq \tau/2\delta$$

From here we use the argument from Theorem 3 to upper (lower) bound the ratio N/D. Since the bound holds with probability $1-2\delta$, we get that if $\max_{i\in[n]} p_i < 1-\frac{\tau}{\delta}$, then f is $\left(\max_{i\in[n]-\Gamma}\left(\max\left\{\ln\left(\frac{1+\frac{\tau}{\delta(1-p_i)}}{1-\frac{\tau}{\delta p_i}}\right), \ln\left(\frac{1+\frac{\tau}{\delta p_i}}{1-\frac{\tau}{\delta(1-p_i)}}\right)\right\}\right), 2\delta\right)$-noiseless private which again makes sense as long as $\frac{\tau}{\delta} < \min_{i\in[n]} p_i$ and $\max_{i\in[n]} p_i < 1 - \frac{\tau}{\delta}$. □

3.3 Handling multiple queries in Adversarial Refreshment Model

Unlike the static model, in this model we assume that every query is run on a database where some significant part of it is new. We focus on the *adversarial replacement model*.

Definition 7 (d-Adversarial Refreshment Model)**.** *In this model, except for an adversarially chosen d bits of the database T, the remaining bits are refreshed under the data generating distribution D before every query f_i.*

We demonstrate the composability of boolean to boolean queries (*i.e.*, $f : \{0,1\}^n \rightarrow \{0,1\}$) under this model.

By the reduction shown in Theorem 1, privacy under multiple queries follows from the privacy in single query under auxiliary information. We use Theorems 1 and 4 to obtain the following *composition* theorem for boolean functions.

Corollary 1. *Let f be $(1-\tau)$-far away from $d+1$ junta (with $d = O(n)$), that is for any function g that depends only on a subset of variables of size $d+1$, $|Pr[f(T) = g(S)] - 1/2| < \tau/2$. Let T be changed as per the d-Adversarial Refreshment Model and let $\hat{T}$ be the database formed by concatenating the new entries (in the d-Adversarial Refreshment Model) with the existing entries (instead of refreshing them). Let the number of times that f has been queried is m. Under the conditions of Theorem 4, f is $\left(m\max_{i\in[n]}\left(\max\left\{\ln\left(\frac{1+\frac{\tau}{\delta(1-p_i)}}{1-\frac{\tau}{\delta p_i}}\right), \ln\left(\frac{1+\frac{\tau}{\delta p_i}}{1-\frac{\tau}{\delta(1-p_i)}}\right)\right\}\right), 2m\delta\right)$-noiseless private, where n is the size of the database $\hat{T}$ and p_i is the probability of the i-th bit of $\hat{T}$ being one.*

3.4 Symmetric Functions on $\{0,1\}^n$

In this section we look at functions where the domain of the function is a vector of boolean entries but the range is over the reals, *i.e.*, $f : \{0,1\}^n \to \mathbb{R}$. Note that even if f maps to the set of reals, the number of elements of $\mathbb{R}$ which f maps to is at most 2^n.

Consider a boolean database $T \in \{0,1\}^n$ and the query function $f : \{0,1\}^n \to \mathbb{R}$, where each entry of T is drawn from a biased bernoulli trial with bias p. We want to provide ϵ-noiseless privacy for the query function f. Note that even if f maps to the set of reals, the number of elements of $\mathbb{R}$ which f maps to is at most 2^n. The specific class of query functions f we consider are called *symmetric functions*. Formally,

Definition 8 (Symmetric function)**.** *A function $f : \{0,1\}^n \to \mathbb{R}$ is said to be symmetric, if for any input string $T = \langle t_1 t_2 \cdots t_n \rangle \in \{0,1\}^n$ and for all permutations σ from the set of permutations over the set $\{1,\cdots,n\}$ the following holds*

$$f(t_1,\cdots,t_n) = f(t_{\sigma(1)},\cdots,t_{\sigma(n)})$$

The lemma below follows directly from the definition above.

Lemma 4. *Any symmetric function $f : \{0,1\}^n \to \mathbb{R}$ depends only on the number of bits in the input which are* 1.

So, to guarantee noiseless privacy for a symmetric function f, we essentially need to bound the following ratio for all $i \in [n]$ and for all $a \in \{0,\cdots,n\}$

$$\frac{\Pr_T[c_T = a | t_i = 0]}{\Pr_T[c_T = a | t_i = 1]}$$

where c_T is the count of the number of 1 in T. The following theorem provides a bound for it. We will make use of the fact that the entire view of the adversary (which has responses to multiple different queries) can be constructed based only on the count of the number of bits in $T \in \{0,1\}^n$ which are 1.

Theorem 5. *Let T be from the distribution D where each bit is* 1 *with probability p (a constant). Let c_T be the count of number of bits in $T \in \{0,1\}^n$ which are* 1 *and let a be any element from the set $\{0,\cdots,n\}$. For any constant $\epsilon > 0$, c_T is $(\epsilon, negl(n))$-noiseless private.*

Proof. First note that since the vector T is generated from the set $\{0,1\}^n$ under D, the random variable c_T follows a $Binomial(n,p)$ distribution. For any given $v \in \{0,1\}$, the ratio $\frac{\Pr_T[c_T=a|t_i=v]}{\Pr_T[c_T=a|t_i=\bar{v}]}$ is exactly the ratio between the probability masses of the binomial distribution $B = Binomial(n' = n-1, p)$ at two adjacent points a and $a-1$.

For such a distribution and for $0 < x < n' - pn'$, we have

$$\frac{\Pr[B = pn' + x]}{\Pr[B = pn' + x + 1]} = \frac{n - 1 + (x+1)/p}{n - 1 - x/(1-p)}$$

Now, we bound the above ratio by $1 + \epsilon < e^\epsilon$. It follows that as long as $x < \underbrace{\frac{\epsilon p(n-1)(1-p)}{(1+p\epsilon)} - \frac{1-p}{1+p\epsilon}}_{A}$, the above bound holds. By Chernoff bound we have $|x| < A$ w.p. greater than $1 - negl(n)$ (assuming ϵ and p to be constant).

This completes the proof.

□

Handling Auxiliary Information. The above result implies that the privacy guarantee for a sequence of symmetric boolean functions does not degrade across multiple queries, which is an interesting property of symmetric boolean functions.

Corollary 2. *Let T be a static database drawn from the distribution D where each bit initially* 1 *with probability p (a constant). Let $f_1, f_2, \cdots f_m$ be m symmetric queries on T. Then for any constant $\epsilon > 0$, the query sequence $f_1, f_2, \cdots f_m$ is $(\epsilon, negl(n))$-noiseless private.*

4 Real queries

In this section, we study the privacy of functions which operate on databases with real entries and compute a real value as output. We view the database T as a collection of n random variables $\langle t_1, t_2, \ldots, t_n\rangle$ with the i^{th} random variable representing the i^{th} database entry. First, we study the privacy of a query that outputs the sum of functions of database entries, that is, $f_n(T) = \frac{1}{s_n}\sum_{i\in[n]} g_i(t_i)$, $s_n = \sum_{i\in[n]} \mathbb{E}[g_i^2(t_i)]$ in Section 4.1. We provide the set of assumptions about the functions g_i's, under which the response of a single such query can be provided with Noiseless Privacy guarantees in Theorem 7. While Theorem 7 is for an adversary that has no auxiliary information about the database, Theorem 8 is for an adversary that may have auxiliary information about some constant fraction of the database. We note that this query function is important as many learning algorithms, including principal component analysis, k-means clustering and any algorithm in the *statistical query* framework can be captured by this type of query (see [BDMN05]). Next, in section 4.2, we study the case of simple linear queries of the form $f_n(T) = \sum_{i\in[n]} a_i t_i$, $a_i \in \mathbb{R}$ when t_i are drawn i.i.d. from a normal distribution. Clearly, linear queries on database entries are captured by the earlier query type, but here we study how multiple such queries (for a static database) can be answered, albeit with stronger assumptions about the database. We show that we can allow upto $\sqrt[5]{n}$ query-responses (on a static database) while still providing (ϵ, δ)-Noiseless Privacy for any arbitrary ϵ and for δ negligible in n. Again, we give a theorem each for an adversary with no auxiliary information as well as for an adversary who may have auxiliary information about some constant fraction of the database. We present several results about the privacy of these two queries under the various changing databases models in section 4.2.

4.1 Sums of functions of database entries

Let $T = \langle t_1, \cdots, t_n\rangle$ be a database where each $t_i \in \mathbb{R}$ is independently chosen and let $g_i : \mathbb{R} \to \mathbb{R}, \forall i \in [n]$ be a set of n real valued functions. We study the privacy of the following function on the database T: $Y_n = \frac{1}{s_n}\sum_{i=1}^n g_i(t_i)$ where $s_n^2 = \sum_{i=1}^n \mathbb{E}[g_i^2(t_i)]$. We state Hertz Theorem [Her69] below and use it to derive the uniform convergence of the cdf of Y_n to the cdf of the standard normal.

Theorem 6 (Hertz theorem)**.** *Let $X_1, X_2, \cdots, X_n$ be independent random variables with distribution functions V_1, V_2, $\cdots$, V_n, zero means and finite non-zero variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$. Let $s_n^2 = \sum_1^n \sigma_i^2$ and $\Phi(x) = (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{x} e^{-t^2/2}\,\mathrm{dt}$ be the distribution function of $\mathcal{N}(0,1)$. Then,*

$$\left|\Pr[(X_1 + X_2 + \cdots + X_n)s_n^{-1} \le x] - \Phi(x)\right| \le \frac{K}{s_n^3}\int_{u=0}^{s_n} \psi_n(u)du \tag{7}$$

where $\psi_n(u) = \sum_{i=1}^n \int_{|x|>u} x^2 dV_i(x)$.

Corollary 3 (Uniform Convergence of F_n to Φ)**.** *Let F_n be the cdf of $Y_n = \frac{1}{s_n}\sum_{i=1}^n g_i(t_i)$ where $s_n^2 = \sum_{i=1}^n \mathbb{E}[g_i^2(t_i)]$ and let Φ denote the standard normal cdf. If $\forall i \in [n]$, $\mathbb{E}[g_i(t_i)] = 0$ and $\mathbb{E}[g_i^2(t_i)]$, $\mathbb{E}[|g_i(t_i)|^3] = O(1)$ (constants independent of i), then Y_n converges in distribution uniformly to the standard normal random variable as follows: $|F_n(x) - \Phi(x)| = O\left(\frac{1}{\sqrt{n}}\right)$*

Proof. Setting $X_i = g_i(t_i)$ $\forall i$ and solving the integral on the RHS of (7) we obtain the corollary about uniform convergence of the cdf (F_n) of Y_n to the cdf (Φ) of the standard normal random variable.

Let $I_i = \int_{u=0}^{s_n}(\int_{|x|>u} x^2 dV_i(x))du$ where $V_i(\cdot)$ is the distribution function of $g_i(t_i)$. We have

$$I_i = \int_{u=0}^{s_n}\left(\int_{|x|>s_n} x^2 v_i(x)dx\right)du + \int_{u=0}^{s_n}\left(\int_{u<|x|\le s_n} x^2 v_i(x)dx\right)du$$

where $v_i(x)$ is the density function of $g_i(t_i)$. Changing the order of integration, we get

$$\begin{aligned}
I_i &= \int_{|x|>s_n} x^2 v_i \left(\int_{u=0}^{s_n} du \right) dx + \int_{|x|\leq s_n} x^2 v_i \left(\int_{u=0}^{|x|} du \right) dx \\
&= \int_{|x|>s_n} s_n x^2 v_i dx + \int_{|x|\leq s_n} |x|^3 v_i dx \\
&\leq \int_{|x|>s_n} |x|^3 v_i dx + \int_{|x|\leq s_n} |x|^3 v_i dx \\
&= \mathbb{E}[|g_i^3(t_i)|] = O(1)
\end{aligned}$$

This gives us $\sum_{i=1}^n I_i = O(n)$. Now, since $s_n = O(\sqrt{n})$ and since $|\Pr[Y_n \leq x] - \Phi(x)| \leq \frac{K}{s_n^3} \sum_{i=1}^n I_i$, we get $|F_n(x) - \Phi(x)| = O(\frac{1}{\sqrt{n}})$. □

If the pdf f_n of Y_n exists and has a bounded derivative, we can further derive the convergence rate of the pdf f_n to the pdf ϕ of the standard normal random variable. This result about pdf convergence is required because we will need to calculate the conditional probabilities in our privacy definitions over all measurable sets $\mathcal{O}$ in the range of the query output (see Definitions 2 & 4). The result is presented in the following Lemma.

Lemma 5 (Uniform Convergence of f_n to ϕ)**.** *Let $f_n(\cdot)$ be the pdf of $Y_n = \frac{1}{s_n} \sum_{i=1}^n g_i(t_i)$ where $s_n^2 = \sum_{i=1}^n \mathbb{E}[g_i^2(t_i)]$ and let $\phi(\cdot)$ denote the standard normal pdf. If $\mathbb{E}[g_i(t_i)] = 0$, $\mathbb{E}[g_i^2(t_i)]$, $\mathbb{E}[|g_i(t_i)|^3] = O(1)$ $\forall i \in [n]$, and if $\forall i$, the densities of $g_i(t_i)$ exist and have bounded derivative then f_n converges uniformly to the standard normal pdf as follows: $|f_n(x) - \phi(x)| = O\left(\frac{1}{\sqrt[4]{n}}\right)$*

Proof. We are given that the density of $g_i(t_i)$ exists and has bounded derivative $\forall i \in [n]$. Let $f_{g_1}(\cdot)$ and $f_{g_2}(\cdot)$ denote the densities of $g_1(t_1)$ and $g_2(t_2)$ respectively. The density $f_{g_1+g_2}(u)$ of their sum $(g_1(t_1) + g_2(t_2))$ is given by the convolution integral

$$f_{g_1+g_2}(u) = \int_{-\inf}^{\inf} f_{g_1}(\tau) f_{g_2}(u-\tau) d\tau.$$

The derivative of the $f_{g_1+g_2}(\cdot)$ can be written as

$$\lim_{h\to 0} \frac{1}{h} \int_{-\inf}^{\inf} f_{g_1}(\tau) \left(f_{g_2}(u+h-\tau) - f_{g_2}(u-\tau) \right) d\tau.$$

Applying mean value theorem for the $f_{g_2}(\cdot)$ terms and letting C denote the upper-bound on $|f'_{g_2}(\cdot)|$, we obtain that C is also an upper-bound for $|f'_{g_1+g_2}(\cdot)|$. By induction, we conclude that $\forall x, n, |f'_n(x)| \leq C$. We will now show that $\forall \gamma > 0, \exists \bar{N}$ s.t. $\forall x$ and $\forall n \geq \bar{N}$ the following is true:

$$|f_n(x) - \phi(x)| \leq \gamma$$

We have $f_n(x) = \lim_{h\to 0} \frac{F_n(x+h) - F_n(x)}{h}$ where $F_n(\cdot)$ denotes the cdf of Y_n. Define $q_x(n,h) = \frac{F_n(x+h)-F_n(x)}{h}$. Similarly, we can write $\phi(x) = \lim_{h\to 0} \frac{\Phi(x+h)-\Phi(x)}{h}$ and define $r_x(h) = \frac{\Phi(x+h)-\Phi(x)}{h}$. For some constant h^* which is independent of n and x (and which we will fix later)

$$|f_n(x) - \phi(x)| \leq \underbrace{|f_n(x) - q_x(n,h^*)|}_{(I)} + \underbrace{|q_x(n,h^*) - r_x(h^*)|}_{(II)} + \underbrace{|r_x(h^*) - \phi(x)|}_{(III)}$$

The approach for proving the lemma is as follows:

- Find h^* independent of x and n, such that (I) and (III) are $\leq \frac{\gamma}{3}$.
- Find $\bar{N}$ s.t. $\forall n \geq \bar{N}$, we have $(II) \leq \frac{\gamma}{3}$.

Again, using Mean Value Theorem and the fact that $|f'_n(\cdot) \leq C|$, we find that $\exists m$ s.t. , $f_n(x) - f_n(y) = f'_n(m)(x-y)$. This implies $|f_n(x) - f_n(y)| \leq C|x-y|$. If we have $|x-y| \leq \frac{\gamma}{3C}$, we have $|f_n(x) - f_n(y)| \leq \frac{\gamma}{3}$ Now, $\frac{q_n(x,h)}{h} = \frac{F_n(x+h)-F_n(x)}{h}$. From Mean Value Theorem it follows that $\exists w \in [x, x+h]$ such that

$$f_n(w) = \frac{F_n(x+h) - F_n(x)}{h}$$

if $|h| \leq \frac{\gamma}{3C}$, then we can get $|f_n(x) - q_x(n,h)| \leq \frac{\gamma}{3}\ \forall x, \forall n$. Similarly, since $\phi(\cdot)$ is differentiable with $|\phi'(\cdot)| \leq \frac{1}{e\sqrt{2\pi}}$, if we picked $|h| \leq \frac{\gamma e\sqrt{2\pi}}{3}$, $\forall x$ we get, $|r_x(h) - \phi(x)| \leq \frac{\gamma}{3}$. Therefore, if we choose $h^* = M\gamma$ where M is a constant given by $M = \min\{\frac{1}{3C}, \frac{e\sqrt{2\pi}}{3}\}$ then we can ensure that (I) and (III) are both $\leq \frac{\gamma}{3}$.

To obtain an upper-bound for (II) we express $|q_x(n,h^*) - r_x(h^*)|$ as follows:

$$\begin{aligned}
&|q_x(n,h^*) - r_x(h^*)| \\
&= \left|\frac{F_n(x+h^*) - F_n(x)}{h^*} - \frac{\Phi(x+h^*) - \Phi(x)}{h^*}\right| \\
&\leq \left|\frac{F_n(x+h^*) - \Phi(x+h^*)}{h^*}\right| + \left|\frac{F_n(x) - \Phi(x)}{h^*}\right|
\end{aligned}$$

By uniform convergence of F_n to Φ (*Corollary 3*) we can pick $\bar{N}$ such that $\forall n \geq \bar{N}, |F_n(x+h^*) - \Phi(x+h^*)|$ and $|F_n(x) - \Phi(x)|$ are both less than $\frac{\gamma h^*}{6}$ and thus (II) is also upper-bounded by $\frac{\gamma}{3}$. Finally, since $\frac{\gamma h^*}{6} = \frac{M\gamma^2}{6}$, we note that for any $\gamma > 0$, if $|F_n(x) - \Phi(x)| \leq \frac{M\gamma^2}{6}, \forall n \geq \bar{N}$ and $\forall x$, then we get $|f_n(x) - \phi(x)| \leq \gamma, \forall n \geq \bar{N}$ and $\forall x$. This gives us $|f_n(x) - \phi(x)| = O(\frac{1}{\sqrt[4]{n}})$.

From corollary 3 we have

$$\forall x, (\lambda > 0), \left|\int_{x-\lambda}^{x} f_n(x)dx - \int_{x-\lambda}^{x} \phi(x)dx\right| \leq O\left(\frac{1}{\sqrt{n}}\right)$$

We set $\frac{\gamma h^*}{6} = \theta\left(\frac{1}{\sqrt{n}}\right)$. Thus, $\forall n \geq 0, |F_n(x+h^*) - \Phi(x+h^*)|$ and $|F_n(x) - \Phi(x)|$ are both $\leq \frac{\gamma h^*}{6}$. Since $h^* = \theta(\gamma)$. Replacing h^* in $\frac{\gamma h^*}{6}$ and equating with $\left(\frac{1}{\sqrt{n}}\right)$, we have $\gamma = O\left(\frac{1}{n^{\frac{1}{4}}}\right)$.

Hence, $\forall n \geq 0, \forall x, |f_n(x) - \phi(x)| \leq O\left(\frac{1}{n^{\frac{1}{4}}}\right)$.

□

Our results concerning the privacy of $Y_n = \frac{1}{s_n}\sum_{i=1}^{n} g_i(t_i)$ are presented in Theorem 7 (for an adversary with no auxiliary information) and Theorem 8 (for an adversary with auxiliary information about at most a constant fraction of the database).

Theorem 7 (Sum of functions of database entries)**.** *Let $T = \langle t_1, \cdots, t_n\rangle$ be a database where each t_i is an independent, real-valued random varaible. Let $g_i : \mathbb{R} \to \mathbb{R}, \forall i \in [n]$ be a set of one-to-one real valued functions and let $Y_n = \frac{1}{s_n}\sum_{i=1}^{n} g_i(t_i)$, where $s_n^2 = \sum_{i=1}^{n} \mathbb{E}[g_i^2(t_i)]$ and $\forall i \in [n]$, $\mathbb{E}[g_i(t_i)] = 0$, $\mathbb{E}[g_i^2(t_i)]$, $\mathbb{E}[|g_i(t_i)|^3] = O(1)$ (constants independent of i) and $\forall i \in [n]$ the density function for $g_i(t_i)$ exists and has bounded derivative ($\frac{dg(t_i)}{dt_i}$ is a constant, independent of i). Then, Y_n is $\left(O\left(\frac{\ln n}{\sqrt[6]{n}}\right), O\left(\frac{1}{\sqrt{n}}\right)\right)$-Noiseless Private.*

Proof. We will show that $\forall a, \alpha, \beta \in \mathbb{R}$, the following ratio is bounded from above and below.

$$R = \frac{\mathrm{pdf}(Y_n = a|t_\ell = \alpha)}{\mathrm{pdf}(Y_n = a|t_\ell = \beta)}$$

Define $Z = \frac{1}{s_z}\sum_{i=1, i\neq \ell}^{n} g_i(t_i)$, where $s_z^2 = \sum_{i=1, i\neq\ell}^{n} \mathbb{E}[g_i^2(t_i)]$. If g_ℓ is a one-to-one function, we can rewrite the above ratio as

$$R = \frac{\mathrm{pdf}\left(Z = \frac{as_n - g_\ell(\alpha)}{s_z}\right)}{\mathrm{pdf}\left(Z = \frac{as_n - g_\ell(\beta)}{s_z}\right)}$$

.

The uniform convergence result of Lemma 5 gives us

$$R \leq \frac{\phi\left(\frac{as_n - g_\ell(\alpha)}{s_z}\right) + \frac{\xi}{(n-1)^{\frac{1}{4}}}}{\phi\left(\frac{as_n - g_\ell(\beta)}{s_z}\right) - \frac{\xi}{(n-1)^{\frac{1}{4}}}}$$

where $\xi \geq 0$ is some constant. Let us assume that $\frac{\xi}{(n-1)^{\frac{1}{4}}} \leq \frac{1}{10} \cdot \phi\left(\frac{as_n - g_\ell(\alpha)}{s_z}\right)$ and also $\frac{\xi}{(n-1)^{\frac{1}{4}}} \leq \frac{1}{10} \cdot \phi\left(\frac{as_n - g_\ell(\beta)}{s_z}\right)$. These two assumptions imply that we should have,

$$\left(\frac{as_n - g_\ell(\alpha)}{s_z}\right) \leq \sqrt{\ln\left(\frac{\sqrt{n-1}}{\xi^2 \cdot 200\pi}\right)} \tag{8}$$

and

$$\left(\frac{as_n - g_\ell(\beta)}{s_z}\right) \leq \sqrt{\ln\left(\frac{\sqrt{n-1}}{\xi^2 \cdot 200\pi}\right)} \tag{9}$$

We will later choose a, $g_\ell(\alpha)$ and $g_\ell(\beta)$ so that the conditions in (8)-(9) hold. Under these assumptions, the ratio becomes:

$$R = \frac{1.1}{0.9} \frac{\phi\left(\frac{as_n - g_\ell(\alpha)}{s_z}\right)}{\phi\left(\frac{as_n - g_\ell(\beta)}{s_z}\right)}$$

Now, since $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

$$\begin{aligned}
\frac{\phi\left(\frac{as_n - g_\ell(\alpha)}{s_z}\right)}{\phi\left(\frac{as_n - g_\ell(\beta)}{s_z}\right)} &\leq \exp\left(\frac{1}{2}\left|\left(\frac{as_n - g_\ell(\alpha)}{s_z}\right)^2 - \left(\frac{as_n - g_\ell(\beta)}{s_z}\right)^2\right|\right) \\
&= \exp\left(\frac{1}{2s_z^2}\left|(as_n - g_\ell(\alpha))^2 - (as_n - g_\ell(\beta))^2\right|\right) \\
&\leq \exp\left(\frac{1}{2s_z^2}\left|2as_n - (g_\ell(\alpha) + g_\ell(\beta))\right|\left|g_\ell(\alpha) - g_\ell(\beta)\right|\right)
\end{aligned}$$

Now, if we choose $|a| = O(\sqrt{\ln n})$, and $|g_\ell(\alpha)|, |g_\ell(\beta)| = O(\sqrt[3]{n}\sqrt{\ln n})$, since $s_n = \theta(n)$ and $s_z = \theta(n-1)$, we get

$$\ln R \;\; = \;\; O\left(\frac{\ln n}{\sqrt[6]{n}}\right)$$

Also, for the chosen values of $a, g_\ell(\alpha)$ and $g_\ell(\beta)$, equations 8 and 9 are satisfied for appropriate choices of constants. Now we estimate the probabilities of $|g_\ell(\alpha)| \geq c_1\sqrt[3]{n}\sqrt{\ln n}$, $|g_\ell(\beta)| \geq c_2\sqrt[3]{n}\sqrt{\ln n}$ and $|a| \geq c_3\sqrt{\ln n}$ (This is the probability with which our privacy guarantee does not hold). Both $g_\ell(\alpha)$ and $g_\ell(\beta)$ represent realizations from zero mean distributions with constant variances. Thus, using Chebyshev's inequality,

$$\Pr[|g_\ell(\alpha)| \geq c_1\sqrt[3]{n}\sqrt{\ln n}] \leq \frac{1}{n^{\frac{2}{3}}\ln n}$$

$$\Pr[|g_\ell(\beta|) \geq c_2\sqrt[3]{n}\sqrt{\ln n}] \leq \frac{1}{n^{\frac{2}{3}}\ln n}$$

Since a represents a realization of Z, from Corollary 3 we have $|\Pr[Z \leq x] - \Phi(x)| = O\left(\frac{1}{\sqrt{n-1}}\right)$,

$$\begin{aligned}
&\Pr[|a| > c_3\sqrt{\ln n}] \\
&= 1 - \Pr[|a| \leq c_3\sqrt{\ln n}] \\
&= 1 - \left(\Pr[Z \leq c_3\sqrt{\ln n}] - \Pr[Z \leq -c_3\sqrt{\ln n}]\right) \\
&\leq 1 - \left[\Phi(c_3\sqrt{\ln n}) - \Phi(-c_3\sqrt{\ln n}) - O\left(\frac{1}{\sqrt{n-1}}\right)\right] \\
&= 1 - \Pr[|X| \leq c_3\sqrt{\ln n}] + O\left(\frac{1}{\sqrt{n-1}}\right) \text{(where } X \sim \mathcal{N}(0,1)) \\
&= \Pr[|X| > c_3\sqrt{\ln n}] + O\left(\frac{1}{\sqrt{n-1}}\right)
\end{aligned}$$

If $X \sim \mathcal{N}(0, \sigma^2)$, then Mill's inequality bounds the tail of the distribution as follows:

$$Pr[|X| \geq x] \leq \sqrt{\frac{2}{\pi}}\left(\frac{\sigma}{x}\right)\exp(-\frac{x^2}{2\sigma^2})$$

Thus, finally we get $\Pr[|a| > c_3\sqrt{\ln n}] \leq \sqrt{\frac{2}{\pi}}\left(\frac{1}{c_3\sqrt{\ln n}}\right)\exp(-\frac{c_3^2 \ln n}{2}) + O\left(\frac{1}{\sqrt{n-1}}\right) = O\left(\frac{1}{\sqrt{n}}\right)$

Hence, it follows that Y_n is $\left(O\left(\frac{\ln n}{\sqrt[6]{n}}\right), O\left(\frac{1}{\sqrt{n}}\right)\right)$-noiseless private for the ℓ^{th} entry of the database T. □

Theorem 8 (Sum of functions of database entries with auxiliary information). *Let $T = \langle t_1, \cdots, t_n\rangle$ be a database where each t_i is an independent, real-valued random variable. Let $g_i : \mathbb{R} \to \mathbb{R}, \forall i \in [n]$ be a set of one-to-one real valued functions and let $Y_n = \frac{1}{s_n}\sum_{i=1}^n g_i(t_i)$, where $s_n^2 = \sum_{i=1}^n \mathbb{E}[g_i^2(t_i)]$ and $\forall i \in [n]$, $\mathbb{E}[g_i(t_i)] = 0$, $\mathbb{E}[g_i^2(t_i)]$, $\mathbb{E}[|g_i(t_i)|^3] = O(1)$ and $\forall i \in [n]$ the density functions for $g_i(t_i)$ exist and have bounded derivative. Let the auxiliary information $\mathcal{A}ux$ be any subset of T of size ρn. Then, Y_n is $\left(O\left(\frac{\ln(n(1-\rho))}{\sqrt[6]{n(1-\rho)}}\right), O\left(\frac{1}{\sqrt{n(1-\rho)}}\right)\right)$-Noiseless Private (for entries in $T \setminus \mathcal{A}ux$).*

Sketch of the proof: To analyze the privacy of the ℓ^{th} entry in the database T, we consider the ratio $R = \text{pdf}(Y_n = a|t_\ell = \alpha, \mathcal{A}ux)/\text{pdf}(Y_n = a|t_\ell = \beta, \mathcal{A}ux)$. Setting $Z = \frac{1}{s_z}\sum_{i\in[n]\setminus I(\mathcal{A}ux), i\neq\ell} g_i(t_i)$, where $s_z^2 = \sum_{i\in[n]\setminus I(\mathcal{A}ux), i\neq\ell}^n \mathbb{E}[g_i^2(t_i)]$, we can rewrite this ratio as $R = \text{pdf}(Z = \frac{z_0 - g_\ell(\alpha)}{s_z})/\text{pdf}(Z = \frac{z_0 - g_\ell(\beta)}{s_z})$, where $I(\mathcal{A}ux)$ is the index set of $\mathcal{A}ux$ and $z_0 = as_n - \sum_{j\in I(\mathcal{A}ux)} g_j(t_j)$. Thereafter, the proof is similar to the proof of Theorem 7 except that Z is now a sum of $n(1-\rho)$ random variables instead of $n-1$.

4.2 Privacy analysis of linear queries

We consider a sequence of linear queries $f_n^i(T) = \sum_{j\in[n]} a_{ij}t_j$, $i = 1, 2, \ldots$ with constant and bounded coefficients a_{ij} for a static database T. For each $m = 1, 2, \ldots$, we ask if the set $\{f_n^i(T) : i = 1, \ldots, m\}$ of queries can have Noiseless Privacy guarantees. We first present Gershgorin circle theorem which we need for our privacy theorem for linear queries.

Theorem 9 (Gershgorin circle theorem). *Let A be a complex $m \times m$ matrix, with entries a_{ij}. For $i \in \{1, \cdots, m\}$, let $R_i = \sum_{j\neq i}|a_{ij}|$ be the sum of the absolute values of the entries in the i-th row. Let $D(a_{ii}, R_i)$ be the closed disc centered at a_{ii} with radius R_i. Such a disc is called a* Gershgorin disc. *Then, every eigenvalue of A lies within at least one of the Gershgorin discs $D(a_{ii}, R_i)$.*

Theorem 10 (Linear Queries). *Consider a database $T = \langle t_1, \ldots, t_n\rangle$ where each t_j is drawn* i.i.d *from $\mathcal{N}(0, 1)$. Let $f_n^i(T) = \sum_{j\in[n]} a_{ij}t_j$, $i = 1, 2, \ldots$, be a sequence of linear queries (over T) with constant coefficients a_{ij}, $|a_{ij}| \leq 1$ and at least two non-zero coefficients in each query. For every m, $1 \leq m \leq \sqrt[5]{n}$, the set of queries $\{f_n^1(T), \ldots, f_n^m(T)\}$ is $(\epsilon, negl(n))$-Noiseless Private for any constant ϵ, provided the following conditions hold: (Diagonal Dominance) For all $i \in [m]$, $\ell \in [n]$, $R(\ell, i) \leq 0.99\sum_{j=1,j\neq\ell}^n a_{ij}^2$, where $R(\ell, i) = \sum_{k=1,k\neq i}^m |\sum_{j=1,j\neq\ell}^n a_{ij}a_{kj}|$.*

Proof. We will first prove the privacy of the ℓ^{th} data item, t_ℓ. Let $Y_i = \sum_{j=1}^{n} a_{ij} t_j$, where t_j are sampled i.i.d. from $\mathcal{N}(0, 1)$. For any $\alpha, \beta \in \mathbb{R}$ and any $\vec{v} = (y_1, \cdots, y_m) \in \mathbb{R}^m$ the following ratio needs to be bounded from above and below by $exp(\epsilon)$ to guarantee noiseless privacy:

$$\frac{\text{pdf}(Y_1 = y_1, \cdots, Y_m = y_m | t_\ell = \alpha)}{\text{pdf}(Y_1 = y_1, \cdots, Y_m = y_m | t_\ell = \beta)}$$

If we define $Z_i = \sum_{j=1, j\neq\ell}^{n} a_{ij} t_j$ for $i \in [m]$, the above ratio is equivalent to:

$$\frac{\text{pdf}(Z_1 = y_1 - a_{1\ell}\alpha, \cdots, Z_m = y_m - a_{m\ell}\alpha)}{\text{pdf}(Z_1 = y_1 - a_{1\ell}\beta, \cdots, Z_m = y_m - a_{m\ell}\beta)}$$

Let $\widetilde{A}$ denote the $m \times (n-1)$ matrix obtained by dropping the ℓ^{th} column of A (the $m \times n$ coefficient matrix). We have $Z_i \sim \mathcal{N}(0, \sum_{j=1, j\neq\ell}^{n} a_{ij}^2)$ and the vector $\vec{Z} = (Z_1, \cdots, Z_m)$ follows the distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma = \widetilde{A}\widetilde{A}^T$. The entries of Σ look like $\Sigma_{ik} = \sum_{j=1, j\neq\ell}^{n} a_{ij} a_{kj}$ and $dim(\Sigma) = m \times m$. The sum of absolute values of non-diagonal entries in the i^{th} row of Σ is given by $R(\ell, i)$ and the i^{th} diagonal entry is $\sum_{j=1, j\neq\ell}^{n} a_{ij}^2$ (denoted Σ_{ii}). By Gershgorin Circle Theorem, the eigenvalues of Σ are lower-bounded by $\Sigma_{ii} - R(\ell, i)$ for some $i \in [m]$. The condition $R(\ell, i) \leq 0.99\Sigma_{ii}$ implies that every eigenvalue is at least $0.01 \times \sum_{j=1, j\neq\ell}^{n} a_{ij}^2$. Since each row of $\widetilde{A}$ has atleast one strictly positive element (each row of A has two non-zero coefficients), this ensures that Σ has strictly positive eigenvalues, and since Σ is also real and symmetric, we know Σ is invertible. Hence, for a given vector $\vec{z} \in \mathbb{R}^m$, we can write

$$\text{pdf}(\vec{Z} = \vec{z}) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} exp(-\frac{1}{2} \vec{z}^T \Sigma^{-1} \vec{z})$$

Let us define $\vec{z_\alpha} = \vec{y} - \alpha\vec{A_\ell}$ and $\vec{z_\beta} = \vec{y} - \beta\vec{A_\ell}$ where $\vec{A_\ell}$ denotes the ℓ^{th} column of A. Therefore, the ratio we try to bound becomes:

$$exp\left(-\frac{1}{2}\left(\vec{z_\alpha}^T \Sigma^{-1} \vec{z_\alpha} - \vec{z_\beta}^T \Sigma^{-1} \vec{z_\beta}\right)\right)$$

Let $\Sigma^{-1} = Q\Lambda Q^T$ be the eigen decomposition and let $\vec{z'_\alpha} = Q^T \vec{z_\alpha}$ and $\vec{z'_\beta} = Q^T \vec{z_\beta}$ under the eigen basis.

The ratio becomes

$$\begin{aligned} & exp\left(-\frac{1}{2}\left(\vec{z'_\alpha}^T \Lambda \vec{z'_\alpha} - \vec{z'_\beta}^T \Lambda \vec{z'_\beta}\right)\right) \\ = \; & exp\left(-\frac{1}{2}\sum_{i=1}^{m} \lambda_i \left((z'_{\alpha,i})^2 - (z'_{\beta,i})^2\right)\right) \end{aligned}$$

where $z'_{\alpha,i}$ is the i-th entry of $\vec{z'_\alpha}$, $z'_{\beta,i}$ is the i-th entry of $\vec{z'_\beta}$ and λ_i is the i-th eigen value of Σ^{-1}. Examining each term in the summation individually, we have $(z'_{\alpha,i})^2 - (z'_{\beta,i})^2 = (z'_{\alpha,i} + z'_{\beta,i})(z'_{\alpha,i} - z'_{\beta,i})$. Further we have,

$$\begin{aligned} & exp\left(-\frac{1}{2}\sum_{i=1}^{m} \lambda_i \left((z'_{\alpha,i})^2 - (z'_{\beta,i})^2\right)\right) \\ = \; & exp\left(-\frac{1}{2}\sum_{i=1}^{m} \lambda_i \left((z'_{\alpha,i} + z'_{\beta,i})(z'_{\alpha,i} - z'_{\beta,i})\right)\right) \\ \leq \; & exp\left(\frac{\lambda_{\max}}{2}\sum_{i=1}^{m} \left|(z'_{\alpha,i} + z'_{\beta,i})(z'_{\alpha,i} - z'_{\beta,i})\right|\right) \\ \leq \; & exp\left(\frac{\lambda_{\max}}{2}\sum_{i=1}^{m} |(z'_{\alpha,i} + z'_{\beta,i})| \sum_{i=1}^{m} |(z'_{\alpha,i} - z'_{\beta,i})|\right) \end{aligned}$$

where $\lambda_{\max} = arg\max_i \lambda_i$. Note that $\sum_{i=1}^{m} |(z'_{\alpha,i} + z'_{\beta,i})|$ and $\left|\sum_{i=1}^{m} (z'_{\alpha,i} - z'_{\beta,i})\right|$ are L_1 norm of the vectors

$(\vec{z'_\alpha} + \vec{z'_\beta})$ and $(\vec{z'_\alpha} - \vec{z'_\beta})$ respectively. We know that L_1 norm $\leq \sqrt{m} L_2$ norm. Using this inequality we have

$$\begin{aligned} & exp\left(\frac{\lambda_{\max}}{2}\sum_{i=1}^{m}|(z'_{\alpha,i} + z'_{\beta,i})|\sum_{i=1}^{m}|(z'_{\alpha,i} - z'_{\beta,i})|\right) \\ \leq \quad & exp\left(\frac{m\lambda_{\max}}{2}\sqrt{\sum_{i=1}^{m}(z'_{\alpha,i} + z'_{\beta,i})^2}\sqrt{\sum_{i=1}^{m}(z'_{\alpha,i} - z'_{\beta,i})^2}\right) \end{aligned}$$

Since, the L_2 norm remains preserved under orthonormal transformation, hence $||\vec{z'_\alpha} + \vec{z'_\beta}||_2 = ||\vec{z_\alpha} + \vec{z_\beta}||_2$ and $||\vec{z'_\alpha} - \vec{z'_\beta}||_2 = ||\vec{z_\alpha} - \vec{z_\beta}||_2$. Hence, we have

$$\begin{aligned} & exp\left(\frac{m\lambda_{\max}}{2}\sqrt{\sum_{i=1}^{m}(z'_{\alpha,i} + z'_{\beta,i})^2}\sqrt{\sum_{i=1}^{m}(z'_{\alpha,i} - z'_{\beta,i})^2}\right) \\ \leq \quad & exp\left(\frac{m\lambda_{\max}|\alpha-\beta|}{2}\sqrt{\sum_{i=1}^{m}(2y_i - a_{i\ell}(\alpha+\beta))^2}\sqrt{\sum_{i=1}^{m}a_{i\ell}^2}\right) \end{aligned}$$

Thus, this ratio will be less than $exp(\epsilon)$ if the following inequality is satisfied:

$$\sqrt{\textstyle\sum_{i=1}^{m}(2y_i - a_{i\ell}(\alpha+\beta))^2} \leq \tfrac{2\epsilon}{m|(\alpha-\beta)|\lambda_{\max}\|A_\ell\|} \tag{10}$$

For $i \in [m]$ let G_i denote the event $\left[|2y_i - a_{i\ell}(\alpha+\beta)| \leq \frac{2\epsilon}{m^{3/2}|(\alpha-\beta)|\lambda_{\max}\|A_\ell\|}\right]$. The conjunction of events represented by $G = \wedge_i G_i$ implies the inequality in (10). We will now estimate the probability of the event G^c (compliment of G) and show that it is negligible in n. By union bound, we have

$$Pr[G^c] = \Pr[G_1^c \vee G_2^c \cdots \vee G_m^c] \leq \textstyle\sum_{i=1}^{m}\Pr[G_i^c]$$

and for each i we get

$$\begin{aligned} \Pr[G_i^c] &= \Pr\left[|2y_i - a_{i\ell}(\alpha+\beta)| > \frac{2\epsilon}{m^{3/2}|(\alpha-\beta)|\lambda_{\max}\|A_\ell\|}\right] \\ &\leq \Pr\left[|2y_i| + |a_{i\ell}(\alpha+\beta)| > \frac{2\epsilon}{m^{3/2}|(\alpha-\beta)|\lambda_{\max}\|A_\ell\|}\right] \\ &= \Pr\left[|2y_i| > \frac{2\epsilon}{m^{3/2}|(\alpha-\beta)|\lambda_{\max}\|A_\ell\|} - |a_{i\ell}(\alpha+\beta)|\right] \\ &= \Pr\left[|y_i| > e_i\right] \end{aligned}$$

where $e_i = \frac{\epsilon}{m^{3/2}|(\alpha-\beta)|\lambda_{\max}\|A_\ell\|} - \left|\frac{a_{i\ell}(\alpha+\beta)}{2}\right|$. We observe that $Pr[|y_i| > e_i]$ is a non-trivial upper-bound for $Pr[G_i^c]$ only if $e_i > 0$. If $|\alpha|, |\beta| < B$ for some $B > 0$, then $e_i \geq \frac{\epsilon}{2m^2 B\lambda_{\max}} - B$ as $0 < a_{ij} < 1$. We denote this lower-bound for e_i by e_*,

$$e_* = \frac{\epsilon}{2m^2 B\lambda_{\max}} - B \tag{11}$$

Thus, the event $\{|y_i| \leq e_i\} \supseteq \{|\alpha| < B, |\beta| < B, |y_i| \leq e_*\}$. Again, we will need $e_* > 0$ for $\{|\alpha| < B, |\beta| < B, |y_i| \leq e_*\}$ to be non-trivial. Therefore,

$$\begin{aligned} \Pr[|y_i| \leq e_i] &\geq \Pr[(|\alpha| < B) \wedge (|\beta| < B) \wedge (|y_i| \leq e_*)] \\ \Pr[|y_i| > e_i] &\leq \Pr[(|\alpha| \geq B) \vee (|\beta| \geq B) \vee (|y_i| > e_*)] \end{aligned}$$

Since $\Pr[G_i^c] \leq \Pr[|y_i| > e_i]$, $\Pr[G_i^c] \leq \Pr[|\alpha| \geq B \vee |\beta| \geq B \vee |y_i| > e_*]$. Again using Union Bound, we get $\Pr[G_i^c] \leq \Pr[|\alpha| \geq B] + \Pr[|\beta| \geq B] + \Pr[|y_i| > e_*]$. Thus,

$$\Pr[G_1^c \vee G_2^c \cdots \vee G_m^c] \leq m(\Pr[|\alpha| \geq B] + \Pr[|\beta| \geq B]) + \sum_{i=1}^{m}\Pr[|y_i| > e_*] \tag{12}$$

If $X \sim \mathcal{N}(0, \sigma^2)$, then Mill's inequality bounds the tail of the distribution as follows:

$$Pr[|X| \geq x] \leq \sqrt{\frac{2}{\pi}} \left(\frac{\sigma}{x}\right) \exp(-\frac{x^2}{2\sigma^2})$$

Now, since $\alpha, \beta \sim \mathcal{N}(0, 1)$ and since $y_i \sim \mathcal{N}(0, \sigma_i^2)$ where $\sigma_i^2 = \sum_{j=1}^{n} a_{ij}^2$ the upper-bound of (12) can be written as

$$\begin{aligned} &\leq \sqrt{\frac{2}{\pi}} \left[\frac{m}{B} \exp(-\frac{B^2}{2}) + \sum_{i=1}^{m} \frac{\sigma_i}{e_*} \exp\left(-\frac{e_*^2}{2\sigma_i^2}\right)\right] \\ &\leq \sqrt{\frac{2}{\pi}} \left[\frac{m}{B} \exp(-\frac{B^2}{2}) + \frac{m\sqrt{n}}{e_*} \exp\left(-\frac{e_*^2}{2n}\right)\right] \qquad (13) \end{aligned}$$

For the second inequality we use the fact that since a_{ik}'s are constants, we have σ_i^2 is $O(n)$. Now, we fix $B = n^{r_1}$ and $m = n^{r_2}$, $r_1, r_2 \in (0, \frac{1}{2})$. This makes the first of two summands in (13) negligible in n. The second summand (and consequently the upper-bound) will be negligible in n if $e_* = \Omega(n^{r_3 + \frac{1}{2}})$ for some $r_3 > 0$. We now show that this is true provided that $(2r_2 + r1 < \frac{1}{2})$. First, since $\lambda_{\max}$ is the maximum eigenvalue of Σ, we know that $\frac{1}{\lambda_{\max}}(= 0.01 \times \sum_{j=1, j\neq \ell}^{n} a_{ij}^2)$ is the *minimum* eigenvalue of Σ^{-1}. Since a_{ij}'s are all constants between 0 and 1, we have $\frac{1}{\lambda_{\max}} = \Omega(n)$. Putting all this together in (11) we observe that there exist constants C_1 and C_2 such that:

$$\begin{aligned} e_* &\geq \frac{C_1 n \epsilon}{2n^{2r_2} n^{r_1}} - n^{r_1} \\ &= \frac{C_1 \epsilon}{2} n^{1 - 2r_2 - r_1} - n^{r_1} \\ &= C_2 n^{r_3 + \frac{1}{2}} - n^{r_1} \end{aligned}$$

where we have used $r_3 = \frac{1}{2} - 2r_2 - r_1$. Since $r_1 > 0, r_2 > 0$ such that $(2r_2 + r1 < \frac{1}{2})$, we have $r_3 + \frac{1}{2} > r_1 > 0$. This implies $e_* = \Omega(n^{r_3 + \frac{1}{2}})$ for an $r_3 > 0$. Hence the second summand in (13) is also negligible in n, and this gives us our theorem. □

The above theorem is also true if the expected value of the database entries is a non-zero constant. This is our next claim.

Claim 1. *If $Y = \sum_{i=1}^{n} a_i t_i$ is (ϵ, δ)-Noiseless Private for a database*
$T = \langle t_1, \cdots, t_n\rangle$ such that $\forall i, \mathbb{E}[t_i] = 0$, then $Y^* = \sum_{i=1}^{n} a_i t_i^*$, where $t_i^* = t_i + \mu_i$, is also (ϵ, δ)-Noiseless Private.

Proof. Since Y is (ϵ, δ)-noiseless private, there exist sets $S_1, S_2 \in \mathbb{R}$ such that $\Pr[Y \in S_1] + \Pr[t_i \in S_2] \leq \delta$ and $\forall a \in \mathbb{R} \setminus S_1, \alpha, \beta \in \mathbb{R} \setminus S_2, \ell \in [1, n]$, the ratio $r = \frac{\text{pdf}(Y=a|t_\ell=\alpha)}{\text{pdf}(Y=a|t_\ell=\beta)}$ is bounded between $exp(-\epsilon)$ and $exp(\epsilon)$. Let $S_1^* = \{x \in \mathbb{R} | x - \sum_{i=1}^{n} a_i \mu_i \in S_1\}$ be a subset of $\mathbb{R}$ and Let $S_2^* = \{x \in \mathbb{R} | x - \mu_i \in S_2\}$ be a subset of $\mathbb{R}$. Clearly, cardinality of S_1 and S_1^* is same and so is that of S_2 and S_2^*. Now consider the ratio $r = \frac{\text{pdf}(Y^*=a|t_\ell^*=\alpha)}{\text{pdf}(Y^*=a|t_\ell^*=\beta)}$. r is equivalent to $\frac{\text{pdf}(Y=a-\sum_{i=1}^{n} a_i\mu_i|t_\ell=\alpha-\mu_\ell)}{\text{pdf}(Y=a-\sum_{i=1}^{n} a_i\mu_i|t_\ell=\beta-\mu_\ell)}$, which is bounded by $exp(\epsilon)$ as long as $a - \sum_{i=1}^{n} a_i \mu_i \notin S_1$ and $\alpha - \mu_\ell, \beta - \mu_\ell \notin S_2$. Thus Y^* is also (ϵ, δ)-noiseless private. □

The results of *Theorem 10* can be extended to the case when adversary has access to some auxiliary information, $\mathcal{A}ux$, provided that $\mathcal{A}ux$ only contains information about a constant fraction of entries, albeit with a stricter requirement on the coefficients of the queries ($0 < a_{ij} \leq 1$ instead of $|a_{ij}| \leq 1$).

Theorem 11 (Linear queries with auxiliary information)**.** *Consider a database $T = \langle t_1, \ldots, t_n\rangle$ where each t_j is drawn* i.i.d *from $\mathcal{N}(0, 1)$. Let $f_n^i(T) = \sum_{i\in[n]} a_{ij} t_j$, $i = 1, 2, \ldots$, be a sequence of linear queries (over T) with constant coeficients a_{ij}, $0 < a_{ij} \leq 1$ and at least two non-zero coefficients in each query. Let $\mathcal{A}ux$ denote the auxiliary information that the adversary can access. If $\mathcal{A}ux$ only contains information about a constant fraction, ρ, of data entries in T, then, for every m, $1 \leq m \leq \sqrt[5]{n}$, the set of queries $\{f_n^1(T), \ldots, f_n^m(T)\}$ is $(\epsilon, negl(n))$-Noiseless Private for any constant ϵ, provided the following conditions hold: For all $i \in [m], \ell \in [n]$ and $(n - \rho n) \leq r \leq n$*

$$\min_{S_r} \sum_{j\in S_r} \left(0.99a_{ij}^2 - \sum_{k=1, k\neq l}^{m} a_{ij} a_{kj}\right) \geq 0 \qquad (14)$$

where S_r is the collection of all possible $(r-1)$-size subsets of $[n] \setminus \{\ell\}$. The test in (14) can be performed efficiently in $O(n \log n)$ time.

Sketch of the proof: We first give a proof for the case when the auxiliary information $\mathcal{A}ux$ is full disclosure of any r entries of the database which implies privacy for the case when $\mathcal{A}ux$ is any partial information about at most r entries of the database. Fix a set $\widehat{I}$ of indices (out of $[n]$) that correspond to the elements in $\mathcal{A}ux$ (This set is known to the adversary, but not to the mechanism). Let $|\widehat{I}| = r$. The response Y_i to the i^{th} query can be written as $Y_i = \widehat{Y_i} + \sum_{j \in \widehat{I}} a_{ij} t_j$, where $\widehat{Y_i} = \sum_{j \in [n] \setminus \widehat{I}} a_{ij} t_j$. Since the second term in the above summation is known to the adversary, the ratio R that we need to bound for Noiseless Privacy is given by

$$R = \frac{\text{pdf}(Y_1 = y_1, \ldots, Y_m = y_m \mid t_\ell = \alpha, \mathcal{A}ux)}{\text{pdf}(Y_1 = y_1, \ldots, Y_m = y_m \mid t_\ell = \beta, \mathcal{A}ux)} \tag{15}$$

$$= \frac{\text{pdf}(\widehat{Y_i} = y_i - \sum_{j \in \widehat{I}} a_{ij} t_j,\ i = 1, \ldots m \mid t_\ell = \alpha)}{\text{pdf}(\widehat{Y_i} = y_i - \sum_{j \in \widehat{I}} a_{ij} t_j,\ i = 1, \ldots, m \mid t_\ell = \beta)} \tag{16}$$

Applying *Theorem 10* to $\widehat{Y_i}$'s we get $(\epsilon, negl(n))$-Noiseless Privacy for any $m \le \sqrt[5]{n}$, if $\forall i \in [m], \ell \in [n]$:

$$\sum_{j \in [n] \setminus \widehat{I}, j \neq \ell} 0.99 a_{ij}^2 - \sum_{k=1, k \neq i}^{m} \left| \sum_{j \in [n] \setminus \widehat{I}, j \neq \ell}^{n} a_{ij} a_{kj} \right| \ge 0 \tag{17}$$

Theorem 11 uses the stronger condition of $0 < a_{ij} \le 1$ (compared to $|a_{ij}| \le 1$ in *Theorem 10*). Hence, we can remove the mod signs and change order of summation to get the following equivalent test: For all $i \in [m], \ell \in [n]$,

$$\sum_{j \in [n] \setminus \widehat{I}, j \neq \ell} \left(0.99 a_{ij}^2 - \sum_{k=1, k \neq i}^{m} a_{ij} a_{kj} \right) \ge 0 \tag{18}$$

Since $\widehat{I}$ is not known to the mechanism, we need to perform this check for all $\widehat{I}$ and ensure that even the $\widehat{I}$ that minimizes the LHS above must be non-negative. This gives us the test of (14). We can first compute all entries inside the round braces of (18), and then sort and picking the first $(n-r)$ entries. This takes $O(n \log n)$ time. This completes the proof.

Finally, we point out that although *Theorem 11* requires $0 < a_{ij} \le 1$, we can obtain a very similar result for the $|a_{ij}| \le 1$ case as well. This is because (17) is true even for $|a_{ij}| \le 1$. However, unlike for $0 < a_{ij} \le 1$ (when (18) could be derived), testing (17) for all $\widehat{I}$ becomes combinatorial and inefficient.

Privacy under multiple queries on changing databases *Theorems 8 & 11* provide (ϵ, δ)-privacy guarantees under leakage of constant fraction of data as auxiliary information. From *Theorem 1*, this implies composition results under dynamically changing databases (e.g., if each query is (ϵ, δ)-Noiseless Private, composition of m such queries will be $(m\epsilon, m\delta)$-Noiseless Private). As discussed in Sec. 2, we get composition under various models where the database keeps changing.

5 Related Works.

Comparison to differential privacy.

In the present section we compare our notion of privacy (*i.e.*, *noiseless privacy*) with other existing formal notions of privacy. The notion of privacy closest to our notion is *differential privacy* [DMNS06]. Roughly, it says that the output of an algorithm $\mathcal{A}$ satisfying differential privacy is not too much dependent on any individual entry in the database. Differential privacy guarantees that on any two databases T and T' differing in one single entry and for any measurable set $\mathcal{O}$ in the output space of $\mathcal{A}$, we have $\frac{\Pr[A(T)=\mathcal{O}]}{\Pr[A(T')=\mathcal{O}]} \le e^{\epsilon}$. Note that the randomness in the above expression is totally drawn from the coin tosses of the algorithm itself and no assumption is made on the data generating distribution. Having no dependence on the data generating distribution makes differential privacy a very strong privacy notion. The flip side is that since the privacy guarantee totally relies on the randomness of the algorithm $\mathcal{A}$, often differentially private algorithms tend to add too much of noise in the output to have reasonable utility. Noiseless privacy essentially tries to address the utility perspective of an output of a private algorithm. Firstly, data collection is a costly affair, so by adding noise to degrade the utility of output will incur non-trivial additional cost. Secondly, providing worst case privacy guarantee (*i.e.*, without making any assumptions the adversary's knowledge about the database) might be too restrictive in certain scenarios. For example, for a reasonably large scale database it is highly unlikely that the adversary knows all but one row of the database. So, in our setting we solely try to leverage the uncertainty of an adversary about a database in order to guarantee privacy. Roughly speaking Definition 2 says that if there is sufficient entropy in the database (from an adversary's point of view), the probability distribution on the output will remain same even if one entry of the database is changed. The philosophy of differential privacy and our definition is the same, namely, from an adversary's view, the output of a private algorithm should not depend too much on an individual entry. Since, in our setting we draw the total randomness from the uncertainty in the data, we can afford not to add any noise in the output and at the same time guarantee privacy.

One of the difficulties in drawing randomness from the uncertainty in the data is that, it is hard to quantify privacy across multiple queries. Primarily the reason being, with each query the uncertainty in the data reduces in a manner that is hard to quantify. For *e.g.*, it can be shown that on a fixed real valued database T of size n, if n linearly independent queries are posed, database T is completely determined (even if each query is ϵ-noiseless private in itself). Where as since differential privacy draws its randomness from the coin tosses of the algorithm (and not from the uncertainty about the database), it is easy to show that if algorithms $\mathcal{A}_1, \cdots, \mathcal{A}_k$ are each ϵ-differentially private, the combined output of $\mathcal{A}_1, \cdots, \mathcal{A}_k$ is $k\epsilon$-differentially private. This property is also called *composability* [DL09]. We handle the issue of composability with the dynamic model discussed earlier. The idea is that for every query a fraction of the database is totally new (*i.e.*, no queries have been answered on that database yet). As a result to guarantee privacy we will rely on the uncertainty of the new fraction of the database. We show via various theorems that it is possible to achieve composability across multiple queries under the *dynamic* model. We would like to emphasize the fact that a lot of large scale databases (for *e.g.*, web stores of different search engines) are continuously evolving, so the dynamic model is realistic.

Comparison to query auditing and different relaxations of differential privacy.

Apart from our present work, previously there have been attempts to relax the notion of differential privacy in order to improve utility. Works of Kenthapadi *et al.* [KMN05] and Nabar *et al.* [NMK$^+$06] in field of *query auditing* is one such direction of progress. Roughly, in this line of work there is a *query auditor* which decides for a database $T = \langle t_1, \cdots, t_n \rangle$ with real entries, whether to answer a particular query or not. If the auditor decides to answer the query, then the answer to the query is output without adding any noise. The auditor has a data generating distribution X in mind (similar to our model). Now, while deciding whether to answer a query or not, the auditor analyzes the following: it first fixes a length of an interval (call it α) and then partitions the real line into intervals of length α. For any such interval $\mathcal{I}$ and a data entry t_i in the database, it computes $\frac{Pr_X[t_i \in \mathcal{I} | \mathcal{O}]}{Pr_X[t_i \in \mathcal{I}]}$, where $\mathcal{O}$ is the answer to the current query. If the ratio goes beyond e^ϵ, the auditor does not answer the query else it gives the correct answer. Clearly for the auditor to check whether for any interval I if the ratio goes beyond e^ϵ, the number of such intervals $\mathcal{I}$ have to be finite. What that in turn means is that the data entries t_i have to come from a bounded range. And also the runtime of the algorithm will be directly proportional to the number of such intervals which can be potentially large depending on the interval length α. Unlike the query auditing, our proposed approach allows the data coming from unbounded domain, and also there is no runtime overhead as we do not run any sanitization algorithm. Although the query auditing approach has a flavour similar to that of Definition 3, query auditing cannot handle arbitrary small interval range (*i.e.*,$\alpha \rightarrow 0$). Where as Definition 3 can be viewed as giving privacy guarantees even under arbitrary small interval range. Another caveat in the query auditing approach is that since the decision of whether to answer a query or not in itself can leak information about the database, the decision is itself randomized. So, in essence it is possible that a query is not answered when ideally it should have been answered and vice versa. One can view it as some form of noise injection into the system. However, the good aspect of query auditing is that if an answer is output, it is without any noise which is in harmony with the motivation of our present work.

The work of Yitao Duan [Dua09] presents some results about differential privacy of summation queries without addition of any external noise. Although the goal of our work is similar, our results have a broader scope and are based on tighter, more accurate, analysis. First, unlike our work, [Dua09] does not provide any analysis under leakage of auxiliary information and/or composition under multiple queries. Even in the single query setting, [Dua09] only provides asymptotic results based on CLT, without taking into consideration rates of convergence. More importantly, since DP typically requires point-wise convergence of pdfs, not merely that of cdfs (as provided by CLT) it is not even clear from the proof if differential privacy is eventually achieved. In our work, we derive explicit convergence rates of the pdfs to the Gaussian pdf, and bring out the additional regularity conditions on the data distribution that are needed for this to happen. Using this, our analysis is able to make explicit, the relationship between the privacy parameter and other data parameters (such as data size).

Some of the other approaches towards relaxing the notion of differential privacy is by Rastogi *et al.* [RHMS09] and by Machanavajjhala *et al.* [MGG09]. Both these approaches assume a distribution over the database, and guarantee a notion of privacy similar to that of differential privacy by extracting randomness partly from the data generating distribution and partly from the coin tosses of the privacy preserving algorithm. As a result the output has some noise in it. In Rastogi *et al.* [RHMS09] approach it is not clear how one can deal with multiple queries when the number of queries are not known in advance. Approach of Machanavajjhala *et al.* [MGG09] is incomparable to our approach because we are concerned with interactive setting (*i.e.,* where a user poses a query and receives an answer in return), whereas [MGG09] is concerned with the non-interactive setting (*i.e.,* where a sanitized version of the database is published and any user query is on that sanitized database).

Acknowledgements. We would like to thank Cynthia Dwork for suggesting to us the changing database model and other useful comments. We would also like to thank Adam Smith and Piyush Srivastava for various intriguing conversations during the course of the project.

References

[BDMN05] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In *PODS*, pages 128–138, 2005.

[DL09] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, STOC '09, pages 371–380, New York, NY, USA, 2009. ACM.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

[Dua09] Yitao Duan. Differential privacy for sum queries without external noise. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*, 2009.

[GKS08] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, pages 265–273, 2008.

[Her69] Ellen S. Hertz. On convergence rates in the central limit theorem. In *Ann. Math. Statist.*, volume 40, pages 475–479, 1969.

[KLM$^+$09] Mihail N. Kolountzakis, Richard J. Lipton, Evangelos Markakis, Aranyak Mehta, and Nisheeth K. Vishnoi. On the fourier spectrum of symmetric boolean functions. *Combinatorica*, 29(3):363–387, 2009.

[KMN05] Krishnaram Kenthapadi, Nina Mishra, and Kobbi Nissim. Simulatable auditing. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '05, pages 118–127, New York, NY, USA, 2005. ACM.

[MGG09] Ashwin Machanavajjhala, Johannes Gehrke, and Michaela Götz. Data publishing against realistic adversaries. volume 2, pages 790–801. VLDB Endowment, August 2009.

[MGKV06] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-ano-nymity. In *ICDE*, page 24, 2006.

[MKA$^+$08] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *ICDE '08: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 277–286, Washington, DC, USA, 2008. IEEE Computer Society.

[NMK$^+$06] Shubha U. Nabar, Bhaskara Marthi, Krishnaram Kenthapadi, Nina Mishra, and Rajeev Motwani. Towards robustness in query auditing. In *In VLDB*, pages 151–162, 2006.

[RHMS09] Vibhor Rastogi, Michael Hay, Gerome Miklau, and Dan Suciu. Relationship privacy: output perturbation for queries with joins. In *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '09, pages 107–116, New York, NY, USA, 2009. ACM.

[Swe02] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

A Boolean Queries

A.1 Growing database Model

In this model we assume that the database is constantly growing in size. This in turn means that every time a query is posed, sizable number of entries in the database are new (*i.e.,* has not been used for answering any of the previous queries). Intuition suggests, its easier to guarantee privacy across multiple queries under this kind of a setting simply because every query is answered on a different database. We would like to emphasize on the fact that this kind of setting is not unnatural. In fact, the kinds of databases we are interested in providing privacy (such as internet scale databases) actually are constantly changing. Lets define formally, what we mean by the *growing database model.*

Definition 9 (k-growing database)**.** *A database T is k-growing w.r.t. a sequence of queries $f_1, \cdots$ if between every two queries f_i and f_{i-1} ($i > 1$), at least k new entries get added to the database. The addition of the new entries preserve ordering, i.e., the new entries are always added at the end of the database.*

Under the k-growing database model we can answer a class of queries which we call *extendible queries.* One of the salient point of such queries is that the privacy guarantee does not degrade across multiple queries. So, potentially there is no limit on the number of extendible queries that can be asked without violating privacy guarantee.

Definition 10 (Extendible sequence w.r.t. k-growing database)**.** *A sequence of queries $f_1, \cdots, f_m$ are said to be extendible w.r.t. k-growing database T if response to every query f_i can be generated from the responses to queries $f_1, \cdots, f_{i-1}$ and a query f'_i (*called the auxiliary query w.r.t. *f_i) which only depends on $\geq k$ new entries that have been added on to the database after query f_{i-1} and before query f_i.*

The following theorem guarantees noiseless privacy for extendible sequence of queries.

Theorem 12. *Let $f_1, \cdots, f_m$ be an extendible query sequence w.r.t. k-growing database. For a given query f_i, let f'_i be the auxiliary query corresponding to it. If for all $i \in [m]$, f'_i satisfies ϵ-noiseless privacy for any ϵ, the extendible query sequence itself is ϵ-noiseless private.*

The idea behind this the proof of this theorem is simple. Consider the queries $f'_1, \cdots, f'_m$. Every entry is the database influences the answer to only one of these queries (since they operate on disjoint set of entries). Thus, the responses to the queries still preserve ϵ-noiseless privacy of the database. Finally we observe that the view of the adversary can be fully reconstructed based on the responses to the queries $f'_1, \cdots, f'_m$. A full proof of this theorem follows below.

Proof. Let the query functions $f_1, \cdots, f_m$ map to some range $\mathcal{R}$ and let the domain of the database be $\mathcal{D}$. Let T_i be the database on which the i-th query is executed. Also, let a_i be the response to the i-th query. In order to prove that the extendible query sequence is ϵ-noiseless private, we need to show the following for any j and $\alpha, \beta \in \mathcal{D}$

$$\frac{\Pr_{T_1,\cdots,T_m}[f_1(T_1) = a_1, \cdots, f_m(T_m) = a_m | t_j = \alpha]}{\Pr_{T_1,\cdots,T_m}[f_1(T_1) = a_1, \cdots, f_m(T_m) = a_m | t_j = \beta]} \leq e^{\epsilon} \tag{19}$$

where t_j is the j-th entry of the database T_n and $a_1, \cdots, a_m$ be any query sequence. Note that $T_1 \subset \cdots \subset T_m$.

Let us denote the new entries that has come in between the $i-1$-th query and the i-th query by $T_i' = T_i - T_{i-1}$. If we break the numerator in the above inequality via conditioning, we have the following

$$\Pr[f_m(T_m) = a_m | a_1, a_2, \cdots, a_m - 1, t_j = \alpha] \cdots \Pr[f_1(T_1) = a_1 | t_j = \alpha]$$

where the probability is over the data generating distribution.

Now, if the j-th entry t_j is not in T_i this expression $\Pr[f_i(T_i) = a_i | a_1, a_2, \cdots, a_i - 1, t_j = \alpha]$ is independent of the conditioning. Hence, they get canceled in the numerator and denominator of the expression in equation 19.

The other case is that t_j is actually in the database T_i. We can write $\Pr[f_i(T_i) = a_i | a_1, a_2, \cdots, a_i - 1, t_j = \alpha]$ as $\Pr[f_i'(T_i') = a_i' | t_j = \alpha]$, where a_i' is the response to f_i' (the auxillary query) is required to get $f_i(T_i) = a_i$ conditioned on $a_1, \cdots, a_{i-1}$.

If $t_j \notin T_i'$, then $\Pr[f_i'(T_i') = a_i' | t_j = \alpha]$ is essentially equal to $\Pr[f_i'(T_i') = a_i']$ (because the conditioning does not matter). Hence, this term gets canceled in the numerator and the denominator of the ratio we are trying to bound. If $t_j \in T_i'$, by our initial assumption, $\frac{\Pr[f_i'(T_i') = a_i' | t_j = \alpha]}{\Pr[f_i'(T_i') = a_i' | t_j = \beta]} \le e^{\epsilon}$. Again, by assumption it is possible to have only one i such that $t_j \in T_i'$. Hence, over all

$$\frac{\Pr_{T_1, \cdots, T_m}[f_1(T_1) = a_1, \cdots, f_m(T_m) = a_m | t_j = \alpha]}{\Pr_{T_1, \cdots, T_m}[f_1(T_1) = a_1, \cdots, f_n(T_m) = a_m | t_j = \beta]} \le e^{\epsilon}$$

□

With a similar argument one can show the same result for (ϵ, δ)-noiseless privacy. To see the utility of the result consider the following example. Let $f_1, \cdots, f_m$ be a query sequence, where each f_i is a symmetric boolean function over database entries $\langle t_1, \cdots, t_{ik} \rangle$, where $t_j \in \{0, 1\}$ are the entries of a k-growing database. Let $\forall i \in [m], c_i$ be the count of number of *ones* in the database entries $\langle t_{i-1}k + 1, \cdots, t_{ik} \rangle$. For any $i \in [m]$, Theorem 5 guarantees the privacy for the count c_i. Since, $c_1, \cdots, c_m$ are enough to answer queries $f_1, \cdots, f_m$, Theorem 12 suggests that if $\forall i \in [m], c_i$ is (ϵ, δ)-noiseless private, the query sequence $f_1, \cdots, f_m$ is (ϵ, δ)-noiseless private.

Instead of modeling the database to be growing in size (*i.e.*, the *growing database model*), we can model the changing database under conventional streaming model. In streaming model, the data is continuously arriving and the system which handles the data has a bounded memory. So, as an when new data arrives, the system purges out the old data. This can also be viewed as a *FIFO* (First In First Out) data structure, for *e.g.*, a queue. From a privacy perspective we assume that every time a query is posed, a significant fraction of the data stored in the storage is *new*, *i.e.,* no queries have been answered yet on these newly arrived data entries yet. We define it more formally below.

Let $f_1, \cdots, f_m$ be the queries on the data stream T_{stream} under the streaming model. Let λn entries of the database of size n that gets purged before every query. A sequence of queries $f_1', \cdots, f_m'$ is called a λ-*disjoint sequence* on T_{stream} if f_i' depends only on a part of the stream T_{stream} whose length is $\ge \lambda n$ and is disjoint from the parts of the stream on which the other queries from the sequence depend on.

Theorem 13. *If $f_1, \cdots, f_m$ can be computed from a λ-disjoint sequence only and if each query in the λ-disjoint sequence is (ϵ, δ)-noiseless private, then the query sequence $f_1, \cdots, f_m$ is also (ϵ, δ)-noiseless private.*

The proof is very similar to that of Theorem 12 and hence omitted. This theorem demonstrates that queries which are noiseless private in the growing database model can be moulded into queries which are noiseless private in the streaming model. The difference being that the queries in the streaming model depend on a database of size exactly n unlike the growing database model. For *e.g.,* symmetric boolean queries discussed above are also private in the streaming model.

A.2 Random Boolean Functions

The following theorem suggests that most Boolean functions satisfy very good privacy parameters under the definition of noiseless privacy.

Theorem 14. *With probability at least $1 - e^{-\Theta(\epsilon^2 2^n)}$ over the choice of f, a random boolean function $f : \{0,1\}^n \to \{-1, 1\}$ is 4ϵ-noiseless private.*

Proof. In the following, $N = 2^n$ and we associate a number $i \in [N]$ with the characteristic set S over n indices. Also, as f is random, for a fixed $T \in \{0,1\}^n$, $E_f[f(T)] = 0$. For a fixed set $S \subseteq [N]$ and a random boolean function $f : \{0,1\}^n \to -1, 1$, define N real random variables Y_i^S's, $i \in [N]$, where $Y_i^S = f(i)\chi_S(i)$. Let $Y^S = \sum_{i \in [N]} Y_i^S$. The following observations are now in order.

- $\hat{f}(S) = \frac{Y^S}{N}$.
- $E_f[Y_i^S] = 0$ for all $S \subseteq [N]$ as $E_f[f(T)]] = 0$ for a random f.
- $E_f[Y^S] = 0$ for all $S \subseteq [N]$ by linearity of expectation.

As each $Y_i \in [-1, 1]$, we apply the Chernoff Hoeffding bound to get

$$\Pr_f[|\frac{Y^S}{N} - 0| > t] \leq 2e^{-\frac{t^2 N}{2}}$$

This reduces to the following. For a fixed set $S \subseteq [N]$ and $t > 0$,

$$\Pr_f[|\hat{f}(S)| > t] \leq 2e^{-\frac{t^2 N}{2}}$$

By a union bound over all sets of cardinality 0 and 1, we thus see that with probability at least $1 - \Theta\left(ne^{-t^2 N}\right)$, a random function has $|\hat{f}(S)| \leq t$. We now choose t to be an arbitrarily small constant, say ϵ. Then using Theorem 3, we obtain a privacy parameter of at most $\ln\left(\frac{1+2\epsilon}{1-2\epsilon}\right)$ which is approximately equal to 4ϵ for small $\epsilon > 0$. □