Noiseless Database Privacy

Raghav Bhaskar[1], Abhishek Bhowmick[2], Vipul Goyal[1], Srivatsan Laxman[1], and Abhradeep Thakurta[3]

[1] Microsoft Research India {rbhaskar,vipul,slaxman}@microsoft.com
[2] University of Texas, Austin bhowmick@cs.utexas.edu
[3] Pennsylvania State University azg161@cse.psu.edu

Abstract. Differential Privacy (DP) has emerged as a formal, flexible framework for privacy protection, with a guarantee that is agnostic to auxiliary information and that admits simple rules for composition. Benefits notwithstanding, a major drawback of DP is that it provides noisy[4] responses to queries, making it unsuitable for many applications. We propose a new notion called Noiseless Privacy that provides exact answers to queries, without adding any noise whatsoever. While the form of our guarantee is similar to DP, where the privacy comes from is very different, based on statistical assumptions on the data and on restrictions to the auxiliary information available to the adversary. We present a first set of results for Noiseless Privacy of arbitrary Boolean-function queries and of linear Real-function queries, when data are drawn independently, from nearly-uniform and Gaussian distributions respectively. We also derive simple rules for composition under models of dynamically changing data.

1 Introduction

Developing a mathematically sound notion of privacy is a difficult problem. Several definitions for database privacy have been proposed over the years, many of which were subsequently broken. For example, methods like k-anonymity [Swe02] and ℓ-diversity [MGKV06] are vulnerable to simple, practical attacks that can breach privacy of individual records [GKS08]. In 2006, Dwork *et al.* [DMNS06] made significant strides toward formal specification of privacy guarantees by introducing an information-theoretic notion called *Differential Privacy* (DP). For a detailed survey on DP see [Dwo08].

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

DP provides a flexible framework for privacy protection based on mechanisms that provide noisy responses to the database queries. The amount of noise

[4] By noise we broadly refer to any external randomization introduced in the output by the privacy mechanism.

introduced in the query-response is: 1) Independent of the actual data entries, 2) Based on the sensitivity of the query to "arbitrary" change of a small number of entries in the data, and 3) Agnostic to the auxiliary information available to the adversary. Their benefits notwithstanding, these properties of DP also result in high levels of noise in the DP output, oftentimes leading to unusable query responses [MKA+08]. Several applications, in fact, completely breakdown when even the slightest amount of noise is added to the output (For example, during a *financial audit*, noisy query-responses may reveal inconsistencies that may be wrongly interpreted as fraud). Besides, when transitioning from a noise-free regime, to incorporate privacy guarantees, the query-response mechanism must be re-programmed (to inject a calibrated amount of noise) and the mechanism *consuming* the DP output must be re-analyzed for its utility/effectiveness (since it must now operate on noisy, rather than exact, query-responses). Hence, the addition of noise to query-responses in the DP framework can be a major barrier to the adoption of DP in practice. Moreover, it is unclear if the DP guarantee (or for that matter, if *any* privacy guarantee) can provide meaningful privacy protection when the adversary has access to arbitrary auxiliary information. On the positive side, however, the structure of the DP guarantee makes it easy to derive simple rules of composition under multiple queries.

Noiseless Privacy: In this paper, we propose a new, also information-theoretic, notion of privacy called *Noiseless Privacy* that provides *exact* answers to database queries, without adding any noise whatsoever. While the form of our guarantee is similar to DP, where the privacy comes from is very different, and is based on: 1) A statistical (generative) model assumption for the database, 2) Restrictions on the kinds of auxiliary information available to the adversary. Both these assumptions are reasonable in many real-world settings; the former is, e.g., commonly used in machine learning, while the latter is natural when data is collected from a diverse network/collection of sources (e.g., from users of the world-wide web).

Consider an entry t_i in the database and two possible values a and b which it can take. Noiseless Privacy simply requires that the probability of the output (or the vector of outputs in-case of multiple queries) lying in a certain measurable set remains similar whether t_i takes value a or b. 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.

While the DP framework makes no assumptions about the data distribution or the auxiliary information available to the adversary, it requires the addition of external noise to query-responses. By contrast, in Noiseless Privacy, we study the privacy implications of providing noise-free responses to queries, but under assumptions governing the data distribution and limited auxiliary information.

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 (and our work shows there are significant non-trivial cases where Noiseless Privacy can be achieved) it comes for "free." Another practical benefit

is that no changes are needed in the query-response or response-consumption mechanisms, only an analysis to "okay the system" to establish the necessary privacy guarantees is required. Moving forward, we believe that checking the feasibility of Noiseless Privacy is a useful first-step when designing privacy-preserving systems. Only when sufficient intrinsic entropy in the data cannot be established, do we need external noise-injection in the query-responses. This way, we would *pay for privacy only when strictly necessary.*

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 the (single) boolean query setting; i.e., the entries of the database as well as the query output have one bit of information each, with no auxiliary information available to the adversary. 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[5] satisfy Noiseless Privacy with "good" parameters. Note that functions which are close to either 0-junta or 1-junta do not represent an "aggregate statistic" 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. We extend our result to the case when the adversary has auxiliary information about some records in the database.

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 aggregation functions that can support various data mining and machine learning tasks in the real-world. We show natural conditions on the database distribution for which Noiseless Privacy can be obtained with good parameters, even when the adversary has auxiliary information about some constant fraction of the dataset. We refer the reader to section 4.1 for more details.

Multiple Queries: The above results are for the case where the adversary is allowed to ask a single query, except for the case of linear real queries, where we have a result for multiple queries. In general, achieving composition in the Noiseless Privacy framework is tricky and privacy can completely breakdown even given a response to two different (carefully crafted) queries. 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 given the responses.

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

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. This helps us maintain some weak independence between different queries. We note that the setting of the changing database model is not unrealistic. Consider an organization that participates in a yearly industry-wide salary survey, where each organization submits relevant statistics about the salaries of its employees to some market research firms. A key requirement in such surveys is to maintain anonymity of 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 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.

We also present other interesting results like obtaining Noiseless Privacy for symmetric boolean functions, "decomposable" functions, etc. In some 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. An obvious line to pursue is to expand the classes of functions and data distributions for which Noiseless Privacy can be achieved. Relaxing the independence assumption that our current results make on database records is another important topic. There is also scope to explore alternative ways of specifying the auxiliary information available to the adversary. In general, we believe that developing new techniques for analyzing statistical queries for Noiseless Privacy is an important direction of privacy research, that must go hand-in-hand with efforts toward new, more clever ways of adding smaller amounts of noise to achieve Differential Privacy.

Related Works: The line of works most related to ours is that of *query auditing* (see [KMN05] and [NMK+06]) where, given a database $T = \langle t_1, \cdots, t_n \rangle$ with real entries, a *query auditor* makes a decision as to whether or not a particular query can be answered. If the auditor decides to answer the query, then the answer is output without adding any noise. Since the decision of whether to answer a query can itself leak information about the database, the decision is randomized. This randomization can be viewed as injection of some form of noise into the query response. However, on the positive side, if a decision is made to answer the query, the answer never contains any noise, which is in harmony with the motivation of our present work. See our full version [BBG+11] 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) *Let $\mathcal{D}$ be the domain from which the entries of the database are drawn. A deterministic query function $f : \mathcal{D}^n \to \mathcal{Y}$ is ϵ-noiseless private 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 Definition 1, the present definition differs at least in the following aspects, namely:

- 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. Formally,

Definition 3 (ϵ-Aposteriori Noiseless Privacy) *A deterministic query function $f : \mathcal{D}^n \to \mathcal{Y}$ is ϵ-Aposteriori Noiseless Private 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 fact shows that Definition 3 implies Definition 2 and vice versa with at most two times degradation in the privacy parameter ϵ. See the full version [BBG+11] for the proof.

Fact 1 *A query function f satisfies Definition 3 under a database generating distribution D and auxiliary information $\mathcal{A}ux$, if and only if it satisfies Definition 2 under the same distribution D and same auxiliary information $\mathcal{A}ux$. There is a possible deterioration of the privacy parameter ϵ by at most a factor of two in either direction.*

Hereafter, we will use Definition 2 as our defintion of Noiseless Privacy. We also 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:

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

One kind of auxiliary information ($\mathcal{A}ux$) that we will consider is partial information about some subset of entries of the database (*i.e.* partial disclosure). But often, it is easier to analyze the privacy when $\mathcal{A}ux$ corresponds to a full disclosure (complete revelation) of a subset of entries rather than partial disclosure because it may be difficult to characterize the corresponding conditional probabilities. The following result shows that the privacy degradation when $\mathcal{A}ux$ corresponds to a partial disclosure of information about a subset of entries can never be worse than the privacy degradation under full disclosure of the same set of entries.

Theorem 1 (Auxiliary Information) *Consider a database T and a query function $f(\cdot)$ over T. Let $\mathcal{A}_p$ denote some partial information regarding some fixed (but typically unknown to the mechanism) subset $T' \subset T$. Let $\mathcal{A}_f$ denote the corresponding full information about the entries of T'. If $f(T)$ is (ϵ, δ)-Noiseless Private under (every possible value of) the auxiliary information $\mathcal{A}_f$ (full disclosure) provided to the adversary, then it is also (ϵ, δ)-Noiseless Private under auxiliary information $\mathcal{A}_p$ (partial disclosure).*

Sketch of the proof:

The partial information $\mathcal{A}_p$ induces a distribution over the space of possible full disclosures $\mathcal{A}_f$. Using the law of total probability, we can write

$$\Pr_{T\sim D}[f(T) \in \mathcal{O} | t_\ell = a, \mathcal{A}_p] = \int_{\mathcal{A}_f} \Pr_{T\sim D}[f(T) \in \mathcal{O} | t_\ell = a, \mathcal{A}_f] \; dF(\mathcal{A}_f | \mathcal{A}_p, t_\ell = a) \tag{1}$$

where $F(\mathcal{A}_f|\mathcal{A}_p, t_\ell = a)$ denotes the conditional distribution for $\mathcal{A}_f$ given $\mathcal{A}_p$ and $[t_\ell = a]$. Since $f(T)$ is (ϵ, δ)-Noiseless Private given $\mathcal{A}_f$, there exist appropriate sets S_1 and S_2 (see *Definition 4*) with $\Pr_{T\sim D}[f(T) \in S_1] + \Pr_{T\sim D}[t_j \in S_2] \leq \delta$ 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 $\ell \in [n]$ we have

$$\Pr_{T\sim D}[f(T) \in \mathcal{O} | t_\ell = a, \mathcal{A}_f] \leq e^\epsilon \Pr_{T\sim D}[f(T) \in \mathcal{O} | t_\ell = a', \mathcal{A}_f] \tag{2}$$

The conditional distribution on F given $\mathcal{A}_p$ and t_ℓ in (1) is in fact independent of t_ℓ (since we can only argue about the privacy of the ℓ^{th} entry of T if it has not been already disclosed *fully* in $\mathcal{A}_f$). Now, since $F(\mathcal{A}_f|\mathcal{A}_p, t_\ell = a) = F(\mathcal{A}_f|\mathcal{A}_p, t_\ell = a')$, we can integrate both sides of (2) with respect to the same distribution and obtain, for the same sets S_1 and S_2 as in (2):

$$\Pr_{T\sim D}[f(T) \in \mathcal{O} | t_\ell = a, \mathcal{A}_p] \leq e^\epsilon \Pr_{T\sim D}[f(T) \in \mathcal{O} | t_\ell = a', \mathcal{A}_p] \tag{3}$$

This completes the proof.

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, in Differential Privacy, the privacy guarantees degrade exponentially with the number of queries on a static database. The notion of Noiseless Privacy often fails to compose in the presence of multiple queries on a static database (an exception to this is given in Section 4.2). But we do present several composability results for multiple queries under dynamic databases.

Dynamic databases may arise in practical scenarios in several ways: (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 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.

Theorem 2 (Composition) *Consider a sequence of m queries, $f_i(\cdot),\ i \in [m]$, over dynamically changing data, such that, the i^{th} query operates on the subset T_i of data elements. For each $i \geq 2$, let T_i share no more than a constant fraction $\rho, (0 \leq \rho \leq 1)$ of elements with $\cup_{i'<i} T_{i'}$ (i.e., all except ρ fraction of the elements in the database are new). If every query $f_i(T_i)$, individually, is (ϵ_i, δ_i)-Noiseless Private under the release of auxiliary information about a constant fraction ρ of elements in T_i, then the sequence of queries is $(\sum_{i=1}^m \epsilon_i, \sum_{i=1}^m \delta_i)$-Noiseless Private over the entire data.*

Sketch of the proof:

To assess the privacy of the ℓ^{th} element t_ℓ, we write down the following probability:

$$\Pr_{T\sim D}[f_1(T_1) \in \mathcal{O}_1, \ldots, f_m(T_m) \in \mathcal{O}_m \mid t_\ell = a] = \Pr_{T\sim D}[f_1(T_1) \in \mathcal{O}_1 \mid t_\ell = a] \times \prod_{i=2}^{m} \Pr_{T\sim D}[f_i(T_i) \in \mathcal{O}_i \mid f_1(T_1) \in \mathcal{O}_1, \ldots, f_{i-1}(T_{i-1}) \in \mathcal{O}_{i-1}, t_\ell = a] \quad (4)$$

Since T_i shares at most a constant fraction ρ of elements with $\cup_{i'<i} T_{i'}$, the sequence of query responses $\langle f_1(T_1), \ldots, f_{i-1}(T_{i-1})\rangle$, can be thought of as revealing auxiliary (possibly partial) information about at most ρ fraction of elements in T_i. Under such auxiliary leakage, 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 $\Pr_{T\sim D}[f(T) \in S_1^i] + \Pr_{T\sim D}[t_j \in S_2^i] \leq \delta_i$ such that, for all measurable sets $\mathcal{O} \subseteq \mathcal{Y} - S_1^i$, for all $a, a' \in \mathcal{D} - S_2^i$, we have

$$\Pr_{T\sim D}[f_i(T_i) \in \mathcal{O}_i \mid f_1(T_1) \in \mathcal{O}_1, \ldots, f_{i-1}(T_{i-1}) \in \mathcal{O}_{i-1}, t_\ell = a] \leq e^{\epsilon_i} \Pr_{T\sim D}[f_i(T_i) \in \mathcal{O}_i \mid f_1(T_1) \in \mathcal{O}_1, \ldots, f_{i-1}(T_{i-1}) \in \mathcal{O}_{i-1}, t_\ell = a'] \quad (5)$$

Setting $S_1 = \cup_i S_1^i$ and $S_2 = \cup_i S_2^i$, we have $\Pr_{T\sim D}[f(T) \in S_1] + \Pr_{T\sim D}[t_j \in S_2] \leq \sum_{i=1}^m \delta_i$ and using (5) for each of the m terms in the RHS of (4) we get, for all

measurable sets $\mathcal{O}_i \subseteq \mathcal{Y} - S_1$, for all $a, a' \in \mathcal{D} - S_2$,

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

This completes the proof. See the full version [BBG+11] for other results under multiple queries.

3 Boolean queries

In this section we study queries of the form $f : T \to \{0,1\}$, *i.e.*, the query function f acts on a database $T \in \mathcal{D}^n$, where $\mathcal{D}$ is the domain from which the data entries are drawn.

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 $f : \{0,1\}^n \to \{0,1\}$ 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 $f' : \{0,1\}^n \to \{0,1\}$ and let D be a distribution on $\{0,1\}^n$. A function $f : \{0,1\}^n \to \{0,1\}$ is $(1-\tau)$-far from k-junta under D if*

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

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.

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.

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 $f : \{0,1\}^n \to \{0,1\}$ be a boolean function which is $(1-\tau_1)$-far from 0-junta and $(1-\tau_2)$-far from 1-junta under D. If $\frac{\tau_1+\tau_2}{2} \leq \min_{i\in[n]} p_i$ and $\max_{i\in[n]} p_i \leq 1 - \frac{\tau_1+\tau_2}{2}$, 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.*

Proof. Please refer to [BBG+11] for the proof.

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.

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 $f : \{0,1\}^n \to \{0,1\}$ be a boolean function which is $(1-2B)$-far away from $d+1$ junta, that is, for any function g that depends only on a subset S of $U = [n]$ of size $d+1$, $|Pr[f(U) = g(S)] - 1/2| < B$. Let T be a database drawn from D and let Γ be any adversarially chosen subset of variables that has been leaked with $|\Gamma| = d$. If $\frac{B}{\delta} < \min_{i \in [n]} p_i$ and if $\max_{i \in [n]} p_i \leq 1 - \frac{B}{\delta}$, then function f is $\left(\max_{i \in [n]-\Gamma} \left(\max \left\{ \ln \left(\frac{1+\frac{B}{\delta(1-p_i)}}{1-\frac{B}{\delta p_i}} \right), \ln \left(\frac{1+\frac{B}{\delta p_i}}{1-\frac{B}{\delta(1-p_i)}} \right) \right\} \right), 2\delta\right)$-Noiseless Private with respect to the bit $t_i \in T$, where $i \in [n] - \Gamma$.*

Proof. We analyze the ratio given that $\Gamma = t$ is such that $|Pr_R[f(R||t) = 0] - 1/2| < B/\delta$ and $|Pr_R[f(R||t) = t_i] - 1/2| < B/\delta$. 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 1 *Let the underlying distribution be an arbitrary D where each bit is 1 independently with probability p_i. Under D, let f be far away from d junta, that is for any function g that depends only on a subset S (with $|S| = d$) of $U = [n]$, $|Pr_D[f(U) = g(S)] - 1/2| < A$. 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, $|Pr_R[f(R||t)=0]-1/2| < A/\delta$.

Proof. Let $\Gamma \subset U = [n]$, $|\Gamma| = d$, be the set of indices leaked. Note that we use Γ to represent both the indices and the variables itself. Let $R = [n] - \Gamma$. We prove the lemma by contradiction. Suppose the claim is wrong. That is, with probability at least δ over Γ, $|Pr_R[f(R||t)=0]-1/2| > A/\delta$. Construct $g : \{0,1\}^d \to \{0,1\}$ 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 $Pr[f(U) = g(\Gamma)]$. 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 $Pr_R[f(R||t)=0] \geq 1/2 + A/\delta$, S_2 is the set of t such that $Pr_R[f(R||t)=0] \leq 1/2 - A/\delta$ 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. Now, we lower bound $Pr[f(U) = g(\Gamma)]$.

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

This leads to a contradiction.

Lemma 2 *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 *junta, that is for any function* g *that depends only on a subset* S *(with* $|S| = d$*) of* $U = [n]$, $|Pr_D[f(U) = g(S)] - 1/2| < B$. *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, $|Pr_R[f(R||t) = t_i] - 1/2| < B/\delta$, *where* t_i *is the* i*-th entry of the database* T.

Proof. The proof of this lemma is identical to the previous proof. Please see [BBG+11] for the complete proof.

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

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

We now use the argument from the proof of Theorem 3 to upper (lower) bound N/D. Since the bound holds with probability $1-2\delta$, we get $\max_{i\in[n]} p_i \leq$

$1 - \frac{B}{\delta}$; hence f is $(\max_{i\in[n]-\Gamma}\left(\max\left\{\ln\left(\frac{1+\frac{B}{\delta(1-p_i)}}{1-\frac{B}{\delta p_i}}\right), \ln\left(\frac{1+\frac{B}{\delta p_i}}{1-\frac{B}{\delta(1-p_i)}}\right)\right\}\right), 2\delta)$-Noiseless Private which again makes sense as long as $\frac{B}{\delta} < \min_{i\in[n]} p_i$ and $\max_{i\in[n]} p_i \leq 1 - \frac{B}{\delta}$.

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 following *adversarial replacement model*.

Definition 7 (d-Adversarial Refreshment Model) *Except for d adversarially chosen 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.*, $f : \{0,1\}^n \to \{0,1\}$) under this model.

By the reduction shown in Theorem 2, privacy under multiple queries follows from the privacy in single query under auxiliary information. We use Theorems 2 and 4 to obtain the following *composition* theorem for boolean functions.

Corollary 1. *Let f be far away from $d+1$ junta (with $d = O(n)$), that is for any function g that depends only on a subset S of $U = [n]$ of size $d+1$, $|Pr[f(U) = g(S)] - 1/2| < B$. Let the database 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. Let the number of times that f has been queried is m. Under the conditions of Theorem 4, f is $(m\max_{i\in[n]}\left(\max\left\{\ln\left(\frac{1+\frac{B}{\delta(1-p_i)}}{1-\frac{B}{\delta p_i}}\right), \ln\left(\frac{1+\frac{B}{\delta p_i}}{1-\frac{B}{\delta(1-p_i)}}\right)\right\}\right), 2m\delta)$-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.*

Please refer to the full version of the paper [BBG+11] for results on the privacy of symmetric functions.

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 item. First we analyze the privacy of a query that outputs the sum of functions of database rows, 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 a set of assumptions about g_i, under which the response of a single such query can be provided with $(\frac{\ln n}{\sqrt[6]{n}}, \frac{1}{\sqrt{n}})$-Noiseless Privacy guarantees in Theorem 5. While Theorem 5 is for an adversary that has no auxiliary information about the database,

Theorem 6 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. 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 4.3.

4.1 Sums of functions of database rows

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 one-to-one real valued functions with the following properties: (i) $\forall i \in [n], \mathbb{E}[g_i(t_i)] = 0$, (ii) $\forall i \in [n], \mathbb{E}[g_i^2(t_i)] = O(1)$, (iii) $\forall i \in [n], \mathbb{E}[|g_i(t_i)|^3] = O(1)$, and (iv) The density function for $g_i(t_i), \forall i \in [n]$ exists and has a bounded derivative. 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 $s_n^2 = \sum_{i=1}^n \mathbb{E}[g_i^2(t_i)]$. Using Hertz Theorem [Her69] (see [BBG$^+$11]) we can derive the following uniform convergence result for the cdf of Y_n to the cdf of the standard normal.

Corollary 2 (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 $\mathbb{E}[g_i(t_i)] = 0$ and if $\mathbb{E}[g_i^2(t_i)]$, $\mathbb{E}[|g_i(t_i)|^3] \sim O(1)\ \forall i \in [n]$, then Y_n converges in distribution uniformly to the standard normal random variable as follows: $|F_n(x) - \Phi(x)| \sim O\left(\frac{1}{\sqrt{n}}\right)$*

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 (Please refer to [BBG$^+$11] for the proof).

Lemma 3 (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 $\mathbb{E}[g_i(t_i)] = 0$, $\mathbb{E}[g_i^2(t_i)]$, $\mathbb{E}[|g_i(t_i)|^3] \sim O(1)\ \forall i \in [n]$, 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: $|f_n(x) - \phi(x)| \sim O\left(\frac{1}{\sqrt[4]{n}}\right)$*

Theorem 5 (Privacy) *Let $T = \langle t_1, \cdots, t_n\rangle$ be a database where each $t_i \in \mathcal{D}$ is independently chosen. 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* $s_n^2 = \cdot \sum_{i=1}^{n} \mathbb{E}[g_i^2(t_i)]$ *and* $\forall i \in [n]$, $\mathbb{E}[g_i(t_i)] = 0$, $\mathbb{E}[g_i^2(t_i)]$, $\mathbb{E}[|g_i(t_i)|^3] \sim O(1)$ *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 empty. Then,* Y_n *is* $\left(O\left(\frac{\ln n}{\sqrt[6]{n}}\right),\ O\left(\frac{1}{\sqrt{n}}\right)\right)$*-Noiseless Private.*

Sketch of the proof: Please see [BBG+11] for the full proof. To analyze the privacy of the ℓ^{th} entry in the database T, we consider the ratio $R = \text{pdf}(Y_n = a|t_\ell = \alpha)/\text{pdf}(Y_n = a|t_\ell = \beta)$. Setting $Z = \frac{1}{s_z}\sum_{i=1, i\neq\ell}^{n} g_i(t_i)$, where $s_z^2 = \sum_{i=1,i\neq\ell}^{n} \mathbb{E}[g_i^2(t_i)]$, we can rewrite this ratio as $R = \text{pdf}(Z = \frac{as_n - g_\ell(\alpha)}{s_z})/\text{pdf}(Z = \frac{as_n - g_\ell(\beta)}{s_z})$. Applying Lemma 3 to the convergence of the pdf of Z to ϕ, we can upper-bound R using a ratio of appropriate standard normal pdf evaluations. For suitable choice of parameters, this leads to $\ln R \sim O(\frac{\ln n}{\sqrt[6]{n}})$. Using Corollary 2, we can show that the probability of data corresponding to the unsuitable parameters is $O(\frac{1}{\sqrt{n}})$.

Theorem 6 (Privacy with auxiliary information) *Let* $T = \langle t_1, \cdots, t_n\rangle$ *be a database where each* $t_i \in \mathbb{R}$ *is independently chosen. 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* $s_n^2 = \cdot \sum_{i=1}^{n} \mathbb{E}[g_i^2(t_i)]$ *and* $\forall i \in [n]$, $\mathbb{E}[g_i(t_i)] = 0$, $\mathbb{E}[g_i^2(t_i)]$, $\mathbb{E}[|g_i(t_i)|^3] \sim O(1)$ *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.*

Sketch of the proof: Please see [BBG+11] for the full 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 5 except that Z is now a sum of $n(1-\rho)$ random variables instead of $n-1$.

The above theorem and Theorem 1 together imply privacy of $Y_n = \frac{1}{s_n}\sum_{i=1}^{n} g_i(t_i)$ under any auxiliary information about a constant fraction of the database.

4.2 Privacy analysis of $f_n^i(T) = \sum_{j\in[n]} a_{ij}t_j$

We consider a sequence of linear queries $f_n^i(T)$, $i = 1, 2, \ldots$ with constant and bounded coefficients 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.

Theorem 7 (Privacy) *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 coefficients* a_{ij}, $|a_{ij}| \leq 1$ *and at least two non-zero coefficients in each query. Assume the adversary does not have access*

to any auxiliary information. 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 $i \in [m], \ell \in [n]$, $R(\ell, i) \leq 0.99 \sum_{j=1, j\neq\ell}^{n} a_{ij}^2$, where $R(\ell, i) = \sum_{k=1, k\neq i}^{m} |\sum_{j=1, j\neq\ell}^{n} a_{ij} a_{kj}|$.

Sketch of the proof: Please refer to [BBG+11] for the complete proof. One can represent the sequence of queries and their corresponding answers via a system of linear equations $\boldsymbol{Y} = A\boldsymbol{T}$, where $\boldsymbol{Y}$ is the output vector and A (called the the *design matrix*) is a $m \times n$ matrix. Each row A^i of the matrix A represents the coefficients of the i-th query. Note that we cannot hope to allow more than n linearly independent *linear* queries. Because in that case the adversary can extract the entire database T from the query responses.

We will prove the privacy of the ℓ^{th} data item, t_ℓ for some $\ell \in [n]$. 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 $\boldsymbol{v} = (y_1, \cdots, y_m) \in \mathbb{R}^m$ the following ratio r needs to be bounded by e^ϵ to guarantee Noiseless Privacy: $r = \frac{\text{pdf}(Y_1 = y_1, \cdots, Y_m = y_m | t_\ell = \alpha)}{\text{pdf}(Y_1 = y_1, \cdots, Y_m = y_m | t_\ell = \beta)}$. If we define $Z_i = \sum_{j=1, j\neq\ell}^{n} a_{ij} t_j$ for $i \in [m]$, $r = \frac{\text{pdf}(Z_1 = y_1 - a_{1\ell}\alpha, \cdots, Z_m = y_m - a_{m\ell}\alpha)}{\text{pdf}(Z_1 = y_1 - a_{1\ell}\beta, \cdots, Z_m = y_m - a_{m\ell}\beta)}$.

Let $\widetilde{A}$ denote the $m \times (n-1)$ matrix obtained by dropping ℓ^{th} column of A. We have $Z_i \sim \mathcal{N}(0, \sum_{j=1, j\neq\ell}^{n} a_{ij}^2)$ and the vector $\boldsymbol{Z} = (Z_1, \cdots, Z_m)$ 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 (see [BBG+11]), 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 at least two a_{ij}'s per query are strictly non-zero, Σ will have strictly positive eigenvalues, and since Σ is also real and symmetric, we know Σ is invertible. Hence, for a given vector $\boldsymbol{z} \in \mathbb{R}^m$, we can write $\text{pdf}(\boldsymbol{Z} = \boldsymbol{z}) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}\boldsymbol{z}^T \Sigma^{-1} \boldsymbol{z})$. Then, for $\boldsymbol{z_\alpha} = \boldsymbol{y} - \alpha \boldsymbol{A_\ell}$ and $\boldsymbol{z_\beta} = \boldsymbol{y} - \beta \boldsymbol{A_\ell}$ where $\boldsymbol{A_\ell}$ denotes the ℓ^{th} column of A, $r = exp\left(-\frac{1}{2}\left(\boldsymbol{z_\alpha}^T \Sigma^{-1} \boldsymbol{z_\alpha} - \boldsymbol{z_\beta}^T \Sigma^{-1} \boldsymbol{z_\beta}\right)\right)$ Let $\Sigma^{-1} = Q\Lambda Q^T$ be the eigen decomposition and let $\boldsymbol{z'_\alpha} = Q^T \boldsymbol{z_\alpha}$ and $\boldsymbol{z'_\beta} = Q^T \boldsymbol{z_\beta}$ under the eigen basis. Then, $r = exp\left(-\frac{1}{2}\sum_{i=1}^{m} \lambda_i \left((z'_{\alpha,i})^2 - (z'_{\beta,i})^2\right)\right)$, where $z'_{\alpha,i}$ is the i-th entry of $\boldsymbol{z'_\alpha}$, $z'_{\beta,i}$ is the i-th entry of $\boldsymbol{z'_\beta}$ and λ_i is the i-th eigen value of Σ^{-1}. Further it can be shown that,

$$r \leq exp\left(\frac{m\lambda_{\max}|\alpha - \beta|}{2} \sqrt{\sum_{i=1}^{m} (2y_i - a_{i\ell}(\alpha + \beta))^2} \sqrt{\sum_{i=1}^{m} a_{i\ell}^2}\right)$$

where $\lambda_{\max} = \arg\max_i \lambda_i$ and we have used the fact that L_1 norm $\leq \sqrt{m}$ L_2 norm and that L_2 norms of $\boldsymbol{z'_\alpha}$ and $\boldsymbol{z'_\beta}$ are equal to L_2 norms of $\boldsymbol{z_\alpha}$ and $\boldsymbol{z_\beta}$ respectively. Thus, this ratio will be less than e^ϵ if:

$$\sqrt{\sum_{i=1}^{m} (2y_i - a_{i\ell}(\alpha + \beta))^2} \leq \frac{2\epsilon}{m|(\alpha - \beta)|\lambda_{\max}\|A_\ell\|} \tag{7}$$

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 (7). Then, in the last step of the proof, we show (see [BBG$^+$11]) that the probability of the event G^c (compliment of G) is negligible in n for any ϵ and $m \leq n^{\frac{1}{5}}$. The above theorem is also true if the expected value of the database entries is a non-zero constant. This is our next claim (see [BBG$^+$11] for the proof).

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.*

The results of *Theorem 7* 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 8 (Privacy 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 $(n - \rho n) \leq r \leq n$*

$$\min_{S_r} \sum_{j \in S_r} \left(0.99 a_{ij}^2 - \sum_{k=1, k \neq l}^{m} a_{ij} a_{kj} \right) \geq 0 \tag{8}$$

where S_r is the collection of all possible $(r-1)$-size subsets of $[n] \setminus \{\ell\}$. The test in (8) 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. Thereafter, we use Theorem 1 to get 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

$$R = \frac{\text{pdf}(Y_1 = y_1, \ldots, Y_m = y_m \mid t_\ell = \alpha, \mathcal{A}ux)}{\text{pdf}(Y_1 = y_1, \ldots, Y_m = y_m \mid t_\ell = \beta, \mathcal{A}ux)} \tag{9}$$

$$= \frac{\text{pdf}(\widehat{Y}_i = y_i - \sum_{j \in \widehat{I}} a_{ij} t_j,\ i = 1, \ldots m \mid t_\ell = \alpha)}{\text{pdf}(\widehat{Y}_i = y_i - \sum_{j \in \widehat{I}} a_{ij} t_j,\ i = 1, \ldots, m \mid t_\ell = \beta)} \tag{10}$$

Applying *Theorem 7* to $\widehat{Y}_i$'s we get $(\epsilon, negl(n))$-Noiseless Privacy for any $m \leq \sqrt[5]{n}$, if $\forall i \in [m], \ell \in [n]$:

$$\sum_{j \in [n] \setminus \widehat{I}, j \neq \ell} 0.99 a_{ij}^2 - \sum_{k=1, k \neq i}^{m} \left| \sum_{j \in [n] \setminus \widehat{I}, j \neq \ell}^{n} a_{ij} a_{kj} \right| \geq 0 \tag{11}$$

Theorem 8 uses the stronger condition of $0 < a_{ij} \leq 1$ (compared to $|a_{ij}| \leq 1$ in *Theorem 7*). Hence, we can remove the mod signs and change order of summation to get the following equivalent test: For all $i \in [m], \ell \in [n]$,

$$\sum_{j \in [n] \setminus \widehat{I}, j \neq \ell} \left(0.99 a_{ij}^2 - \sum_{k=1, k \neq i}^{m} a_{ij} a_{kj} \right) \geq 0 \tag{12}$$

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 (8). We can first compute all entries inside the round braces of (12), 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 8* requires $0 < a_{ij} \leq 1$, we can obtain a very similar result for the $|a_{ij}| \leq 1$ case as well. This is because (11) is true even for $|a_{ij}| \leq 1$. However, unlike for $0 < a_{ij} \leq 1$ (when (12) could be derived), testing (11) for all $\widehat{I}$ becomes combinatorial and inefficient.

4.3 Privacy under multiple queries on changing databases

Theorems 6 & 8 provide (ϵ, δ)-privacy guarantees under leakage of constant fraction of data as auxiliary information. From *Theorem 2*, 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 growing, streaming and random replacement models. In addition, both the queries considered in this section are extendibile (see full version [BBG+11] for details) and thus, one can answer multiple *repeat* queries on a dynamic database (under growing data and streaming models) without degradation in privacy guarantee.

Acknowledgements. We thank Cynthia Dwork for suggesting the changing data model direction, among other useful comments. We also thank Adam Smith and Piyush Srivastava for many useful discussions and suggestions.

References

[BBG+11] Raghav Bhaskar, Abhishek Bhowmick, Vipul Goyal, Srivatsan Laxman, and Abhradeep Thakurta. Noiseless database privacy. Cryptology ePrint Archive, Report 2011/487, 2011. `http://eprint.iacr.org/`.

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

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

[Dwo08] Cynthia Dwork. Differential privacy: a survey of results. In *Proceedings of the 5th international conference on Theory and applications of models of computation*, TAMC'08, pages 1–19, Berlin, Heidelberg, 2008. Springer-Verlag.

[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.

[MGKV06] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. 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.

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