Privacy-Preserving Datamining on Vertically Partitioned Databases

Cynthia Dwork and Kobbi Nissim

Microsoft Research, SVC, 1065 La Avenida, Mountain View CA 94043
{dwork, kobbi}@microsoft.com

Abstract. In a recent paper Dinur and Nissim considered a statistical database in which a trusted database administrator monitors queries and introduces noise to the responses with the goal of maintaining data privacy [5]. Under a rigorous definition of breach of privacy, Dinur and Nissim proved that *unless the total number of queries is sub-linear in the size of the database*, a substantial amount of noise is required to avoid a breach, rendering the database almost useless.
As databases grow increasingly large, the possibility of being able to query only a sub-linear number of times becomes realistic. We further investigate this situation, generalizing the previous work in two important directions: multi-attribute databases (previous work dealt only with single-attribute databases) and vertically partitioned databases, in which different subsets of attributes are stored in different databases. In addition, we show how to use our techniques for datamining on published noisy statistics.

Keywords: Data Privacy, Statistical Databases, Data Mining, Vertically Partitioned Databases.

1 Introduction

In a recent paper Dinur and Nissim considered a statistical database in which a trusted database administrator monitors queries and introduces noise to the responses with the goal of maintaining data privacy [5]. Under a rigorous definition of breach of privacy, Dinur and Nissim proved that *unless the total number of queries is sub-linear in the size of the database*, a substantial amount of noise is required to avoid a breach, rendering the database almost useless[1]. However, when the number of queries is limited, it is possible to simultaneously preserve privacy and obtain some functionality by adding an amount of noise that is a function of the number of queries. Intuitively, the amount of noise is sufficiently large that nothing specific about an individual can be learned from a relatively small number of queries, but not so large that information about sufficiently strong statistical trends is obliterated.

[1] For unbounded adversaries, the amount of noise (per query) must be linear in the size of the database; for polynomially bounded adversaries, $\Omega(\sqrt{n})$ noise is required.

As databases grow increasingly massive, the notion that the database will be queried only a sub-linear number of times becomes realistic. We further investigate this situation, significantly broadening the results in [5], as we describe below.

Methodology. We follow a cryptography-flavored methodology, where we consider a database access mechanism private only if it provably withstands any adversarial attack. For such a database access mechanism any computation over query answers clearly preserves privacy (otherwise it would serve as a privacy breaching adversary). We present a database access mechanism and prove its security under a strong privacy definition. Then we show that this mechanism provides utility by demonstrating a datamining algorithm.

Statistical Databases. A statistical database is a collection of samples that are somehow representative of an underlying population distribution. We model a database as a matrix, in which rows correspond to individual records and columns correspond to attributes. A query to the database is a set of indices (specifying rows), and a Boolean property. The response is a noisy version of the number of records in the specified set for which the property holds. (Dinur and Nissim consider one-column databases containing a single binary attribute.) The model captures the situation of a traditional, multiple-attribute, database, in which an adversary knows enough partial information about records to "name" some records or select among them. Such an adversary can target a selected record in order to try to learn the value of one of its unknown sensitive attributes. Thus, the mapping of individuals to their indices (record numbers) is not assumed to be secret. For example, we do not assume the records have been randomly permuted.

We assume each row is independently sampled from some underlying distribution. An analyst would usually assume the existence of a single underlying row distribution $\mathcal{D}$, and try to learn its properties.

Privacy. Our notion of privacy is a relative one. We assume the adversary knows the underlying distribution $\mathcal{D}$ on the data, and, furthermore, may have some a priori information about specific records, e.g., "p – the a priori probability that at least one of the attributes in record 400 has value 1 – is .38". We anlyze privacy with respect to any possible underlying (row) distributions $\{\mathcal{D}_i\}$, where the ith row is chosen according to D_i. This partially models a priori knowledge an attacker has about individual rows (i.e. $\mathcal{D}_i$ is $\mathcal{D}$ conditioned on the attacker's knowledge of the ith record). Continuing with our informal example, privacy is breached if the a posteriori probability (after the sequence of queries have been issued and responded to) that "at least one of the attributes in record 400 has value 1" differs from the a priori probability p "too much".

Multi-Attribute Sub-Linear Queries (SuLQ) Databases. The setting studied in [5], in which an adversary issues only a sublinear number of queries (SuLQ) to a single attribute database, can be generalized to multiple attributes in several natural ways. The simplest scenario is of a single k-attribute SuLQ database, queried by specifying a set of indices and a k-ary Boolean function. The response is a noisy version of the number of records in the specified set for which the function, applied to the attributes in the record, evaluates to 1. A more involved scenario is of multiple single-attribute SuLQ databases, one for each attribute, administered independently. In other words, our k-attribute database is *vertically partitioned* into k single-attribute databases. In this case, the challenge will be datamining: learning the statistics of Boolean functions of the attributes, using the single-attribute query and response mechanisms as primitives. A third possibility is a combination of the first two: a k-attribute database that is vertically partitioned into two (or more) databases with k_1 and k_2 (possibly overlapping) attributes, respectively, where $k_1 + k_2 \geq k$. Database i, $i = 1, 2$, can handle k_i-ary functional queries, and the goal is to learn relationships between the functional outputs, eg, "If $f_1(\alpha_{1,1}, \ldots, \alpha_{1,k_1})$ holds, does this increase the likelihood that $f_2(\alpha_{2,1} \ldots, \alpha_{2,k_2})$ holds?", where f_i is a function on the attribute values for records in the ith database.

1.1 Our Results

We obtain positive datamining results in the extensions to the model of [5] described above, while maintaining the strengthened privacy requirement:

1. Multi-attribute SuLQ databases: The statistics for every k-ary Boolean function can be learned[2]. Since the queries here are powerful (any function), it is not surprising that statistics for any function can be learned. The strength of the result is that statistics are learned while maintaining privacy.
2. Multiple single-attribute SuLQ databases: We show how to learn the statistics of any 2-ary Boolean function. For example, we can learn the fraction of records having neither attribute 1 nor attribute 2, or the conditional probability of having attribute 2 given that one has attribute 1. The key innovation is a procedure for testing the extent to which one attribute, say, α, implies another attribute, β, *in probability*, meaning that $\Pr[\beta|\alpha] = \Pr[\beta] + \Delta$, where Δ can be estimated by the procedure.
3. Vertically Partitioned k-attribute SuLQ Databases: The constructions here are a combination of the results for the first two cases: the k attributes are partitioned into (possibly overlapping) sets of size k_1 and k_2, respectively, where $k_1 + k_2 \geq k$; each of the two sets of attributes is managed by a multi-attribute SuLQ database. We can learn all 2-ary Boolean functions of the outputs of the results from the two databases.

We note that a single-attribute database can be simulated in all of the above settings; hence, in order to preserve privacy, the sub-linear upper bound on queries must be enforced. How this bound is enforced is beyond the scope of this work.

[2] Note that because of the noise, statistics cannot be learned exactly. An additive error on the order of $n^{1/2-\varepsilon}$ is incurred, where n is the number of records in the database. The same is true for single-attribute databases.

Datamining on Published Statistics. Our technique for testing implication in probability yields surprising results in the real-life model in which confidential information is gathered by a trusted party, such as the census bureau, who publishes aggregate statistics. Describing our results by example, suppose the bureau publishes the results of a large (but sublinear) number of queries. Specifically, for every, say, triple of attributes $(\alpha_1, \alpha_2, \alpha_3)$, and for each of the eight conjunctions of literals over three attributes $(\bar{\alpha}_1\bar{\alpha}_2\bar{\alpha}_3, \bar{\alpha}_1\bar{\alpha}_2\alpha_3, \ldots, \alpha_{k-2}\alpha_{k-1}\alpha_k)$, the bureau publishes the result of several queries on these conjunctions. We show how to construct approximate statistics for any binary function of *six* attributes. (In general, using data published for ℓ-tuples, it is possible to approximately learn statistics for any 2ℓ-ary function.) Since the published data are the results of SuLQ database queries, the total number of published statistics must be sublinear in n, the size of the database. Also, in order to keep the error down, several queries must be made for each conjunction of literals. These two facts constrain the values of ℓ and the total number k of attributes for which the result is meaningful.

1.2 Related Work

There is a rich literature on confidentiality in statistical databases. An excellent survey of work prior to the late 1980's was made by Adam and Wortmann [2]. Using their taxonomy, our work falls under the category of *output perturbation*. However, to our knowledge, the only work that has exploited the opportunities for privacy inherent in the fact that with massive of databases the actual number of queries will be sublinear is Sect. 4 of [5] (joint work with Dwork). That work only considered single-attribute SuLQ databases.

Fanconi and Merola give a more recent survey, with a focus on aggregated data released via web access [10]. Evfimievski, Gehrke, and Srikant, in the Introduction to [7], give a very nice discussion of work in randomization of data, in which data contributors (e.g., respondents to a survey) independently add noise to their own responses. A special issue (Vol.14, No. 4, 1998) of the *Journal of Official Statistics* is dedicated to disclosure control in statistical data. A discussion of some of the trends in the statistical research, accessible to the non-statistician, can be found in [8].

Many papers in the statistics literature deal with generating simulated data while maintaining certain quantities, such as marginals [9]. Other widely-studied techniques include cell suppression, adding simulated data, releasing only a subset of observations, releasing only a subset of attributes, releasing synthetic or partially synthetic data [13,12], data-swapping, and post-randomization. See Duncan (2001) [6].

R. Agrawal and Srikant began to address privacy in datamining in 2000 [3]. That work attempted to formalize privacy in terms of confidence intervals (intuitively, a small interval of confidence corresponds to a privacy breach), and also showed how to reconstruct an original distribution from noisy samples (i.e., each sample is the sum of an underlying data distribution sample and a noise sample), where the noise is drawn from a certain simple known distribution.

This work was revisited by D. Agrawal and C. Aggarwal [1], who noted that it is possible to use the outcome of the distribution reconstruction procedure to significantly diminish the interval of confidence, and hence breach privacy. They formulated privacy (loss) in terms of mutual information, taking into account (unlike [3]) that the adversary may know the underlying distribution on the data and "facts of life" (for example, that ages cannot be negative). Intuitively, if the mutual information between the sensitive data and its noisy version is high, then a privacy breach occurs. They also considered reconstruction from noisy samples, using the EM (expectation maximization) technique. Evfimievsky, Gehrke, and Srikant [7] criticized the usage of mutual information for measuring privacy, noting that low mutual information allows complete privacy breaches that happen with low but significant frequency. Concurrently with and independently of Dinur and Nissim [5] they presented a privacy definition that related the a priori and a posteriori knowledge of sensitive data. We note below how our definition of privacy breach relates to that of [7,5].

A different and appealing definition has been proposed by Chawla, Dwork, McSherry, Smith, and Wee [4], formalizing the intuition that one's privacy is guaranteed to the extent that one is not brought to the attention of others. We do not yet understand the relationship between the definition in [4] and the one presented here.

There is also a very large literature in secure multi-party computation. In secure multi-party computation, *functionality* is paramount, and privacy is only preserved to the extent that the function outcome itself does not reveal information about the individual inputs. In privacy-preserving statistical databases, *privacy* is paramount. Functions of the data that cannot be learned while protecting privacy will simply not be learned.

2 Preliminaries

Notation. We denote by $\mathsf{neg}(n)$ (read: negligible) a function that is asymptotically smaller than any inverse polynomial. That is, for all $c > 0$, for all sufficiently large n, we have $\mathsf{neg}(n) < 1/n^c$. We write $\tilde{O}(T(n))$ for $T(n) \cdot \mathbf{polylog}(n)$.

2.1 The Database Model

In the following discussion, we do not distinguish between the case of a vertically partitioned database (in which the columns are distributed among several servers) and a "whole" database (in which all the information is in one place).

We model a database as an $n \times k$ binary matrix $d = \{d_{i,j}\}$. Intuitively, the columns in d correspond to Boolean attributes $\alpha_1, \ldots, \alpha_k$, and the rows in d correspond to individuals where $d_{i,j} = 1$ iff attribute α_j holds for individual i. We sometimes refer to a row as a *record*.

Let $\mathcal{D}$ be a distribution on $\{0,1\}^k$. We say that a database $d = \{d_{i,j}\}$ is chosen according to distribution $\mathcal{D}$ if every row in d is chosen according to $\mathcal{D}$, independently of the other rows (in other words, d is chosen according to $\mathcal{D}^n$).

In our privacy analysis we relax this requirement and allow each row i to be chosen from a (possibly) different distribution $\mathcal{D}_i$. In that case we say that the database is chosen according to $\mathcal{D}_1 \times \cdots \times \mathcal{D}_n$.

Statistical Queries. A statistical query is a pair (q, g), where $q \subseteq [n]$ indicates a set of rows in d and $g : \{0,1\}^k \rightarrow \{0,1\}$ denotes a function on attribute values. The *exact* answer to (q, g) is the number of rows of d in the set q for which g holds (evaluates to 1):

$$a_{q,g} = \sum_{i \in q} g(d_{i,1}, \ldots, d_{i,k}) = |\{i : i \in q \text{ and } g(d_{i,1}, \ldots, d_{i,k}) \text{ holds}\}|.$$

We write (q, j) when the function g is a projection onto the jth element: $g(x_1, \ldots, x_k) = x_j$. In that case (q, j) is a query on a subset of the entries in the jth column: $a_{q,j} = \sum_{i \in q} d_{i,j}$. When we look at vertically partitioned single-attribute databases, the queries will all be of this form.

Perturbation. We allow the database algorithm to give perturbed (or "noisy") answers to queries. We say that an answer $\hat{a}_{q,j}$ is within perturbation $\mathcal{E}$ if $|\hat{a}_{q,j} - a_{q,j}| \leq \mathcal{E}$. Similarly, a database algorithm $\mathcal{A}$ is within perturbation $\mathcal{E}$ if for every query (q, g)

$$\Pr[|\mathcal{A}(q, g) - a_{q,g}| \leq \mathcal{E}] = 1 - \mathsf{neg}(n).$$

The probability is taken over the randomness of the database algorithm $\mathcal{A}$.

2.2 Probability Tool

Proposition 1. *Let $s_1, \ldots, s_t$ be random variables so that $|\mathbf{E}[s_i]| \leq \alpha$ and $|s_i| \leq \beta$ then*

$$\Pr[|\sum_{i=1}^{T} s_t| > \lambda(\alpha + \beta)\sqrt{t} + t\beta] < 2e^{-\lambda^2/2}.$$

Proof. Let $z_i' = s_i - \mathbf{E}[s_i]$, hence $|z_i'| \leq \alpha + \beta$. Using Azuma's inequality[3] we get that $\Pr[\sum_{i=1}^{T} z' \geq \lambda(\alpha + \beta)\sqrt{t}] \leq 2e^{-\lambda^2/2}$. As $|\sum_{i=1}^{T} s_t| = |\sum_{i=1}^{T} z' + \sum_{i=1}^{T} \mathbf{E}[s_i]| \leq |\sum_{i=1}^{T} z'| + t\beta$ the proposition follows.

3 Privacy Definition

We give a privacy definition that extends the definitions in [5,7]. Our definition is inspired by the notion of semantic security of Goldwasser and Micali [11]. We first state the formal definition and then show some of its consequences.

Let $p_0^{i,j}$ be the a priori probability that $d_{i,j} = 1$ (taking into account that we assume the adversary knows the underlying distribution $\mathcal{D}_i$ on row i. In

[3] Let $X_0, \ldots, X_m$ be a martingale with $|X_{i+1} - X_i| \leq 1$ for all $0 \leq i < m$. Let $\lambda > 0$ be arbitrary. Azuma's inequality says that then $\Pr[X_m > \lambda\sqrt{m}] < e^{\lambda^2/2}$.

general, for a Boolean function $f : \{0,1\}^k \to \{0,1\}$ we let $p_0^{i,f}$ be the a priori probability that $f(d_{i,1},\ldots,d_{i,k}) = 1$. We analyze the a posteriori probability that $f(d_{i,1},\ldots,d_{i,k}) = 1$ given the answers to T queries, *as well as* all the values in all the rows of d other than i: $d_{i',j}$ for all $i' \neq i$. We denote this a posteriori probability $p_T^{i,f}$.

Confidence. To simplify our calculations we follow [5] and define a monotonically-increasing 1-1 mapping $\mathrm{conf} : (0,1) \to \mathbb{R}$ as follows:

$$\mathrm{conf}(p) = \log \frac{p}{1-p}.$$

Note that a small additive change in conf implies a small additive change in p.[4] Let $\mathrm{conf}_0^{i,f} = \log \frac{p_0^{i,f}}{1-p_0^{i,f}}$ and $\mathrm{conf}_T^{i,f} = \log \frac{p_T^{i,f}}{1-p_T^{i,f}}$. We write our privacy requirements in terms of the random variables $\Delta\mathrm{conf}^{i,f}$ defined as:[5]

$$\Delta\mathrm{conf}^{i,f} = |\mathrm{conf}_T^{i,f} - \mathrm{conf}_0^{i,f}|.$$

Definition 1 (**(δ,T)-Privacy**). *A database access mechanism is (δ,T)-private if for every distribution $\mathcal{D}$ on $\{0,1\}^k$, for every row index i, for every function $f : \{0,1\}^k \to \{0,1\}$, and for every adversary $\mathcal{A}$ making at most T queries it holds that*

$$\Pr[\Delta\mathrm{conf}^{i,f} > \delta] \leq \mathsf{neg}(n).$$

The probability is taken over the choice of each row in d according to $\mathcal{D}$, and the randomness of the adversary as well as the database access mechanism.

A *target set* F is a set of k-ary Boolean functions (one can think of the functions in F as being selected by an adversary; these represent information it will try to learn about someone). A target set F is δ-*safe* if $\Delta\mathrm{conf}^{i,f} \leq \delta$ for all $i \in [n]$ and $f \in F$. Let F be a target set. Definition 1 implies that under a (δ,T)-private database mechanism, F is δ-safe with probability $1 - \mathsf{neg}(n)$.

Proposition 2. *Consider a (δ,T)-private database with $k = O(\log n)$ attributes. Let F be the target set containing all the 2^{2^k} Boolean functions over the k attributes. Then,* $\Pr[F$ *is* 2δ*-safe*$] = 1 - \mathsf{neg}(n)$.

Proof. Let F' be a target set containing all 2^k conjuncts of k attributes. We have that $|F'| = \mathbf{poly}(n)$ and hence F' is δ-safe with probability $1 - \mathsf{neg}(n)$.

To prove the proposition we show that F is safe whenever F' is. Let $f \in F$ be a Boolean function. Express f as a disjunction of conjuncts of k attributes:

[4] The converse does not hold – conf grows logarithmically in p for $p \approx 0$ and logarithmically in $1/(1-p)$ for $p \approx 1$.

[5] Our choice of defining privacy in terms of $\Delta\mathrm{conf}^{i,f}$ is somewhat arbitrary, one could rewrite our definitions (and analysis) in terms of the a priori and a posteriori probabilities. Note however that limiting $\Delta\mathrm{conf}^{i,f}$ in Definition 1 is a stronger requirement than just limiting $|p_T^{i,f} - p_0^{i,f}|$.

$f = c_1 \vee \ldots \vee c_\ell$. Similarly, express $\neg f$ as the disjunction of the remaining $2^k - \ell$ conjuncts: $\neg f = d_1 \vee \ldots \vee d_{2^k-\ell}$. (So $\{c_1, \ldots, c_\ell, d_1, \ldots, d_{2^k-\ell}\} = F$.)

We have:

$$\Delta\mathrm{conf}^{i,f} = \left|\log\left(\frac{p_T^{i,f}}{p_0^{i,f}} \cdot \frac{p_0^{i,\neg f}}{p_T^{i,\neg f}}\right)\right| = \left|\log\left(\frac{\sum p_T^{i,c_j}}{\sum p_0^{i,c_j}} \cdot \frac{\sum p_0^{i,d_j}}{\sum p_T^{i,d_j}}\right)\right|.$$

Let k maximize $|\log(p_T^{i,c_k}/p_0^{i,c_k})|$ and k' maximize $|\log(p_0^{i,d_{k'}}/p_T^{i,d_{k'}})|$. Using $|\log(\sum a_i/\sum b_i)| \leq \max_i |\log(a_i/b_i)|$ we get that $\Delta\mathrm{conf}^{i,f} \leq |\Delta\mathrm{conf}^{i,c_k}| + |\Delta\mathrm{conf}^{i,d_{k'}}| \leq 2\delta$, where the last inequality holds as $c_k, d_{k'} \in F'$.

(δ, T)-Privacy vs. Finding Very Heavy Sets. Let f be a target function and $\delta = \omega(\sqrt{n})$. Our privacy requirement implies $\delta' = \delta'(\delta, \Pr[f(\alpha_1, \ldots, \alpha_k])$ such that it is infeasible to find a "very" heavy set $q \subseteq [n]$, that is, a set for which $a_{q,f} \geq |q|\,(\delta' + \Pr[f(\alpha_1, \ldots, \alpha_k)])$. Such a δ'-heavy set would violate our privacy requirement as it would allow guessing $f(\alpha_1, \ldots, \alpha_k)$ for a random record in q.

Relationship to the privacy definition of [7] Our privacy definition extends the definition of p_0-to-p_1 privacy breaches of [7]. Their definition is introduced with respect to a scenario in which several users send their sensitive data to a center. Each user randomizes his data prior to sending it. A p_0-to-p_1 privacy breach occurs if, with respect to some property f, the a priori probability that f holds for a user is at most p_0 whereas the a posteriori probability may grow beyond p_1 (i.e. in a worst case scenario with respect to the coins of the randomization operator).

4 Privacy of Multi-Attribute SuLQ databases

We first describe our SuLQ Database algorithm, and then prove that it preserves privacy.

Let $T(n) = O(n^c)$, $c < 1$, and define $R = \big(T(n)/\delta^2\big) \cdot \log^\mu n$ for some $\mu > 0$ (taking $\mu = 6$ will work). To simplify notation, we write d_i for $(d_{i,1}, \ldots, d_{i,k})$, $g(i)$ for $g(d_i) = g(d_{i,1}, \ldots, d_{i,k})$ (and later $f(i)$ for $f(d_i)$).

SuLQ Database Algorithm $\mathcal{A}$
Input: a query (q, g).

1. Let $a_{q,g} = \sum_{i \in q} g(i) \left(= \sum_{i \in q} g(d_{i,1}, \ldots, d_{i,k})\right)$.
2. Generate a perturbation value: Let $(e_1, \ldots, e_R) \in_R \{0,1\}^R$ and $\mathcal{E} \leftarrow \sum_{i=1}^R e_i - R/2$.
3. Return $\hat{a}_{q,g} = a_{q,g} + \mathcal{E}$.

Note that $\mathcal{E}$ is a binomial random variable with $\mathbf{E}[\mathcal{E}] = 0$ and standard deviation $\sqrt{R}$. In our analysis we will neglect the case where $\mathcal{E}$ largely deviates from zero, as the probability of such an event is extremely small: $\Pr[|\mathcal{E}| > \sqrt{R}\log^2 n] = \mathsf{neg}(n)$. In particular, this implies that our SuLQ database algorithm $\mathcal{A}$ is within $\tilde{O}(\sqrt{T(n)})$ perturbation.

We will use the following proposition.

Proposition 3. *Let B be a binomially distributed random variable with expectation 0 and standard deviation $\sqrt{R}$. Let L be the random variable that takes the value $\log\left(\frac{\Pr[B]}{\Pr[B+1]}\right)$. Then*

1. $\log\left(\frac{\Pr[B]}{\Pr[B+1]}\right) = \log\left(\frac{\Pr[-B]}{\Pr[-B-1]}\right)$. *For $0 \leq B \leq \sqrt{R}\log^2 n$ this value is bounded by $O(\log^2 n/\sqrt{R})$).*
2. $\mathbf{E}[L] = O(1/R)$, *where the expectation is taken over the random choice of B.*

Proof. 1. The equality follows from the symmetry of the Binomial distribution (i.e. $\Pr[B] = \Pr[-B]$).
To prove the bound consider $\log(\Pr[B]/\Pr[B+1]) = \log(\binom{R}{R/2+B}/\binom{R}{R/2+B+1}) = \log\frac{R/2+B+1}{R/2-B-1}$. Using the limits on B and the definition of R we get that this value is bounded by $\log(1+O(\log^2 n/\sqrt{R})) = O(\log^2 n/\sqrt{R})$.
2. Using the symmetry of the Binomial distribution we get:

$$\mathbf{E}[L] = \sum_{0\leq B\leq R/2}\binom{R}{R/2+B}2^{-R}\left[\log\frac{R/2+B+1}{R/2-B} + \log\frac{R/2-B+1}{R/2+B}\right]$$
$$= \sum_{0\leq B\leq \log^2 n\sqrt{R}}\binom{R}{R/2+B}2^{-R}\log\left(1+\frac{R+1}{R^2/4-B^2}\right) + \mathsf{neg}(n) = O(1/R)$$

Our proof of privacy is modeled on the proof in Section 4 of [5] (for single attribute databases). We extend their proof (i) to queries of the form (q,g) where g is any k-ary Boolean function, and (ii) to privacy of k-ary Boolean functions f.

Theorem 1. *Let $T(n) = O(n^c)$ and $\delta = 1/O(n^{c'})$ for $0 < c < 1$ and $0 \leq c' < c/2$. Then the SuLQ algorithm $\mathcal{A}$ is $(\delta, T(n))$-private within $\tilde{O}(\sqrt{T(n)}/\delta)$ perturbation.*

Note that whenever $\sqrt{T(n)}/\delta < \sqrt{n}$ bounding the adversary's number of queries to $T(n)$ allows privacy with perturbation magnitude less than $\sqrt{n}$.

Proof. Let $T(n)$ be as in the theorem and recall $R = \left(T(n)/\delta^2\right)\cdot\log^{\mu} n$ for some $\mu > 0$.

Let the $T = T(n)$ queries issued by the adversary be denoted $(q_1, g_1), \ldots, (q_T, g_T)$. Let $\hat{a}_1 = \mathcal{A}(q_1, g_1), \ldots, \hat{a}_t = \mathcal{A}(q_T, g_T)$ be the perturbed answers to these queries. Let $i \in [n]$ and $f : \{0,1\}^k \to \{0,1\}$.

We analyze the *a posteriori* probability p_ℓ that $f(i) = 1$ given the answers to the first ℓ queries $(\hat{a}_1, \ldots, \hat{a}_\ell)$ *and* $d^{\{-i\}}$ (where $d^{\{-i\}}$ denotes the entire database except for the ith row). Let $\mathrm{conf}_\ell = \log_2 p_\ell/(1-p_\ell)$. Note that $\mathrm{conf}_T = \mathrm{conf}_T^{i,f}$ (of Section 3), and (due to the independence of rows in d) $\mathrm{conf}_0 = \mathrm{conf}_0^{i,f}$.

By the definition of conditional probability [6] we get

$$\frac{p_\ell}{1-p_\ell} = \frac{\Pr[f(i)=1|\hat{a}_1,\ldots,\hat{a}_\ell,d^{\{-i\}}]}{\Pr[f(i)=0|\hat{a}_1,\ldots,\hat{a}_\ell,d^{\{-i\}}]} = \frac{\Pr[\hat{a}_1,\ldots,\hat{a}_\ell \wedge f(i)=1|d^{\{-i\}}]}{\Pr[\hat{a}_1,\ldots,\hat{a}_\ell \wedge f(i)=0|d^{\{-i\}}]} = \frac{\mathbf{Num}}{\mathbf{Denom}}.$$

Note that the probabilities are taken over the coin flips of the SuLQ algorithm and the choice of d. In the following we analyze the numerator (the denominator is analyzed similarly).

$$\begin{aligned}\mathbf{Num} &= \sum_{\sigma\in\{0,1\}^k, f(\sigma)=1} \Pr[\hat{a}_1,\ldots,\hat{a}_\ell \wedge d_i=\sigma|d^{\{-i\}}] \\ &= \sum_{\sigma\in\{0,1\}^k, f(\sigma)=1} \Pr[\hat{a}_1,\ldots,\hat{a}_\ell|d_i=\sigma,d^{\{-i\}}]\Pr[d_i=\sigma]\end{aligned}$$

The last equality follows as the rows in d are chosen independently of each other. Note that given both d_i and $d^{\{-i\}}$ the random variable $\hat{a}_\ell$ is independent of $\hat{a}_1,\ldots,\hat{a}_{\ell-1}$. Hence, we get:

$$\mathbf{Num} = \sum_{\sigma\in\{0,1\}^k, f(\sigma)=1} \Pr[\hat{a}_1,\ldots,\hat{a}_{\ell-1}|d_i=\sigma,d^{\{-i\}}]\Pr[\hat{a}_\ell|d_i=\sigma,d^{\{-i\}}]\Pr[d_i=\sigma].$$

Next, we observe that although $\hat{a}_\ell$ depends on d_i, the dependence is weak. More formally, let $\sigma_0,\sigma_1 \in \{0,1\}^k$ be such that $f(\sigma_0)=0$ and $f(\sigma_1)=1$. Note that whenever $g_\ell(\sigma)=g_\ell(\sigma_1)$ we have that $\Pr[\hat{a}_\ell|d_i=\sigma,d^{\{-i\}}] = \Pr[\hat{a}_\ell|d_i=\sigma_1,d^{\{-i\}}]$. When, instead, $g_\ell(\sigma)\neq g_\ell(\sigma_1)$, we can relate $\Pr[\hat{a}_\ell|d_i=\sigma,d^{\{-i\}}]$ and $\Pr[\hat{a}_\ell|d_i=\sigma_1,d^{\{-i\}}]$ via Proposition 3:

Lemma 1. *Let σ,σ_1 be such that $g_\ell(\sigma)\neq g_\ell(\sigma_1)$. Then $\Pr[\hat{a}_\ell|d_i=\sigma,d^{\{-i\}}] = 2^\epsilon\Pr[\hat{a}_\ell|d_i=\sigma_1,d^{\{-i\}}]$ where $|\mathbf{E}[\epsilon]| = O(1/R)$ and*

$$\epsilon = \begin{cases} -(-1)^{g_\ell(\sigma_1)}O(\log^2 n/\sqrt{R}) & \text{if } \mathcal{E}\le 0 \\ (-1)^{g_\ell(\sigma_1)}O(\log^2 n/\sqrt{R}) & \text{if } \mathcal{E} > 0\end{cases}$$

and $\mathcal{E}$ is noise that yields $\hat{a}_\ell$ when $d_i=\sigma$.

Proof. Consider the case $g_\ell(\sigma_1)=0$ ($g_\ell(\sigma)=1$). Writing $\Pr[\hat{a}_\ell|d_i=\sigma,d^{\{-i\}}] = \Pr[\mathcal{E}=k]$ and $\Pr[\hat{a}_\ell|d_i=\sigma_1,d^{\{-i\}}] = \Pr[\mathcal{E}=k-1]$ the proof follows from Proposition 3. Similarly for $g_\ell(\sigma_1)=1$.

Note that the value of ϵ does not depend on σ.

Taking into account both cases ($g_\ell(\sigma)=g_\ell(\sigma_1)$ and $g_\ell(\sigma)\neq g_\ell(\sigma_1)$) we get

$$\mathbf{Num} = \sum_{\sigma\in\{0,1\}^k, f(\sigma)=1} \Pr[\hat{a}_1,\ldots,\hat{a}_{\ell-1}|d_i=\sigma,d^{\{-i\}}]2^\epsilon\Pr[\hat{a}_\ell|d_i=\sigma_1,d^{\{-i\}}]\Pr[d_i=\sigma].$$

[6] I.e. $\Pr[E_1|E_2]\cdot\Pr[E_2] = \Pr[E_1\wedge E_2] = \Pr[E_2|E_1]\cdot\Pr[E_1]$.

Let $\hat{\gamma}$ be the probability, over d_i, that $g(\sigma) \neq g(\sigma_1)$. Letting $\gamma \geq 1$ be such that $2^{1/\gamma} = \hat{\gamma}$, we have

$$\begin{aligned}
\mathbf{Num} &= 2^{\epsilon/\gamma} \Pr[\hat{a}_\ell | d_i = \sigma_1, d^{\{-i\}}] \sum_{\sigma \in \{0,1\}^k, f(\sigma)=1} \Pr[\hat{a}_1, \ldots, \hat{a}_{\ell-1} | d_i = \sigma, d^{\{-i\}}] \Pr[d_i = \sigma] \\
&= 2^{\epsilon/\gamma} \Pr[\hat{a}_\ell | d_i = \sigma_1, d^{\{-i\}}] \sum_{\sigma \in \{0,1\}^k, f(\sigma)=1} \Pr[\hat{a}_1, \ldots, \hat{a}_{\ell-1} \wedge d_i = \sigma | d^{\{-i\}}] \\
&= 2^{\epsilon/\gamma} \Pr[\hat{a}_\ell | d_i = \sigma_1, d^{\{-i\}}] \Pr[\hat{a}_1, \ldots, \hat{a}_{\ell-1} \wedge f(i) = 1 | d^{\{-i\}}] \\
&= 2^{\epsilon/\gamma} \Pr[\hat{a}_\ell | d_i = \sigma_1, d^{\{-i\}}] \Pr[f(i) = 1 | \hat{a}_1, \ldots, \hat{a}_{\ell-1}, d^{\{-i\}}] \Pr[\hat{a}_1, \ldots, \hat{a}_{\ell-1} | d^{\{-i\}}] \\
&= 2^{\epsilon/\gamma} \Pr[\hat{a}_\ell | d_i = \sigma_1, d^{\{-i\}}] p_{\ell-1} \Pr[\hat{a}_1, \ldots, \hat{a}_{\ell-1} | d^{\{-i\}}]
\end{aligned}$$

and similarly

$$\mathbf{Denom} = 2^{\epsilon'/\gamma'} \Pr[\hat{a}_\ell | d_i = \sigma_0, d^{\{-i\}}](1 - p_{\ell-1}) \Pr[\hat{a}_1, \ldots, \hat{a}_{\ell-1} | d^{\{-i\}}].$$

Putting the pieces together we get that

$$\mathrm{conf}_\ell = \log_2 \frac{\mathbf{Num}}{\mathbf{Denom}} = \mathrm{conf}_{\ell-1} + (\epsilon/\gamma - \epsilon'/\gamma') + \log_2 \frac{\Pr[\hat{a}_\ell | d_i = \sigma_1, d^{\{-i\}}]}{\Pr[\hat{a}_\ell | d_i = \sigma_0, d^{\{-i\}}]}.$$

Define a random walk on the real line with $\mathrm{step}_\ell = \mathrm{conf}_\ell - \mathrm{conf}_{\ell-1}$. To conclude the proof we show that (with high probability) T steps of the random walk do not suffice to reach distance δ. From Proposition 3 and Lemma 1 we get that

$$|\mathbf{E}[\mathrm{step}_\ell]| = O(1/R) = O\left(\frac{\delta^2}{T \log^\mu n}\right)$$

and

$$|\mathrm{step}_\ell| = O(\log^2 n / \sqrt{R}) = O\left(\frac{\delta}{\sqrt{T} \log^{\mu/2-2} n}\right).$$

Using Proposition 1 with $\lambda = \log n$ we get that for all $t \leq T$,

$$\Pr[|\mathrm{conf}_t - \mathrm{conf}_0| > \delta] = \Pr[|\sum_{\ell \leq t} \mathrm{step}_\ell| > \delta] \leq \mathsf{neg}(n).$$

5 Datamining on Vertically Partitioned Databases

In this section we assume that the database is chosen according to $\mathcal{D}^n$ for some underlying distribution $\mathcal{D}$ on rows, where $\mathcal{D}$ is independent of n, the size of the database. We also assume that n, is sufficiently large that the true database statistics are representative of $\mathcal{D}$. Hence, in the sequel, when we write things like "$\Pr[\alpha]$" we mean the probability, over the entries in the database, that α holds.

Let α and β be attributes. We say that α *implies* β *in probability* if the conditional probability of β given α exceeds the unconditional probability of β. The ability to measure implication in probability is crucial to datamining. Note that since $\Pr[\beta]$ is simple to estimate well, the problem reduces to obtaining a good estimate of $\Pr[\beta|\alpha]$. Moreover, once we can estimate the $\Pr[\beta|\alpha]$, we can use Bayes' Rule and de Morgan's Laws to determine the statistics for any Boolean function of attribute values.

Our key result for vertically partitioned databases is a method, given two single-attribute SuLQ databases with attributes α and β respectively, to measure $\Pr[\beta|\alpha]$.

For more general cases of vertically partitioned data, assume a k-attribute database is partitioned into $2 \leq j \leq k$ databases, with $k_1, \ldots, k_j$ (possibly overlapping) attributes, respectively, where $\sum_i k_i \geq k$. We can use functional queries to learn the statistics on k_i-ary Boolean functions of the attributes in the ith database, and then use the results for two single-attribute SuLQ databases to learn binary Boolean functions of any two functions f_{i_1} (on attributes in database i_1) and f_{i_2} (on attributes in database i_2), where $1 \leq i_1, i_2 \leq j$.

5.1 Probabilistic Implication

In this section we construct our basic building block for mining vertically partitioned databases.

We assume two SuLQ databases d_1, d_2 of size n, with attributes α, β respectively. When α implies β in probability with a gap of Δ, we write $\alpha \overset{\Delta}{\to} \beta$, meaning that $\Pr[\beta|\alpha] = \Pr[\beta] + \Delta$. We note that $\Pr[\alpha]$ and $\Pr[\beta]$ are easily computed within error $O(1/\sqrt{n})$, simply by querying the two databases on large subsets. Our goal is to determine Δ, or equivalently, $\Pr[\beta|\alpha] - \Pr[\beta]$; the method will be to determine if, for a given Δ_1, $\Pr[\beta|\alpha] \geq \Pr[\beta] + \Delta_1$, and then to estimate Δ by binary search on Δ_1.

Notation. We let $p_\alpha = \Pr[\alpha]$, $p_\beta = \Pr[\beta]$, $p_{\beta|\alpha} = \Pr[\beta|\alpha]$ and $p_{\beta|\bar{\alpha}} = \Pr[\beta|\neg\alpha]$.

Let X be a random variable counting the number of times α holds when we take N samples from $\mathcal{D}$. Then $\mathbf{E}[X] = Np_a$ and $\mathbf{Var}[X] = Np_a(1 - p_a)$.

Let

$$p_{\beta|\alpha} = p_\beta + \Delta. \tag{1}$$

Note that $p_\beta = p_\alpha p_{\beta|\alpha} + (1 - p_\alpha)p_{\beta|\bar{\alpha}}$. Substituting $p_\beta + \Delta$ for $p_{\beta|\alpha}$ we get

$$p_{\beta|\bar{\alpha}} = p_\beta - \Delta \frac{p_\alpha}{1 - p_\alpha}, \tag{2}$$

and hence (by another application of Eq. (1))

$$p_{\beta|\alpha} - p_{\beta|\bar{\alpha}} = \frac{\Delta}{1 - p_\alpha}. \tag{3}$$

We define the following testing procedure to determine, given Δ_1, if $\Delta \geq \Delta_1$. Step 1 finds a heavy (but not very heavy) set for attribute α, that is, a set q for which the number of records satisfying α exceeds the expected number by more than a standard deviation. Note that since $T(n) = o(n)$, the noise $|\hat{a}_{q,1} - a_{q,1}|$

is $o(\sqrt{n})$, so the heavy set really has $Np_\alpha + \Omega(\sqrt{N})$ records for which α holds. Step 2 queries d_2 on this heavy set. If the incidence of β on this set sufficiently (as a function of Δ_1) exceeds the expected incidence of β, then the test returns "1" (ie, success). Otherwise it returns 0.

Test Procedure $\mathcal{T}$
Input: $p_\alpha, p_\beta, \Delta_1 > 0$.

1. Find $q \in_R [n]$ such that $a_{q,1} \geq Np_\alpha + \sigma_\alpha$ where $N = |q|$ and $\sigma_\alpha = \sqrt{Np_\alpha(1-p_\alpha)}$.
Let $\text{bias}_\alpha = a_{q,1} - Np_\alpha$.
2. If $a_{q,2} \geq Np_\beta + \text{bias}_\alpha \frac{\Delta_1}{1-p_\alpha}$ return 1, otherwise return 0.

Theorem 2. *For the test procedure $\mathcal{T}$:*

1. *If $\Delta \geq \Delta_1$, then $\Pr[\mathcal{T}$ outputs 1$] \geq 1/2$.*
2. *If $\Delta \leq \Delta_1 - \varepsilon$, then $\Pr[\mathcal{T}$ outputs 1$] \leq 1/2 - \gamma$,*

where for $\varepsilon = \Theta(1)$ the advantage $\gamma = \gamma(p_\alpha, p_\beta, \varepsilon)$ is constant, and for $\varepsilon = o(1)$ the advantage $\gamma = c \cdot \varepsilon$ with constant $c = c(p_\alpha, p_\beta)$.

In the following analysis we neglect the difference between $a_{q,i}$ and $\hat{a}_{q,i}$, since, as noted above, the perturbation contributes only low order terms (we neglect some other low order terms). Note that it is possible to compute all the required constants for Theorem 2 explicitly, in polynomial time, without neglecting these low-order terms. Our analysis does not attempt to optimize constants.

Proof. Consider the random variable corresponding to $a_{q,2} = \sum_{i \in q} d_{i,2}$, given that q is biased according to Step 1 of $\mathcal{T}$. By linearity of expectation, together with the fact that the two cases below are disjoint, we get that

$$\begin{aligned}\mathbf{E}[a_{q,2}|\text{bias}_\alpha] &= (Np_\alpha + \text{bias}_\alpha)p_{\beta|\alpha} + (N(1-p_\alpha) - \text{bias}_\alpha)p_{\beta|\bar{\alpha}} \\ &= Np_\alpha p_{\beta|\alpha} + N(1-p_\alpha)p_{\beta|\bar{\alpha}} + \text{bias}_\alpha(p_{\beta|\alpha} - p_{\beta|\bar{\alpha}}) \\ &= Np_\beta + \text{bias}_\alpha \frac{\Delta}{1-p_\alpha}.\end{aligned}$$

The last step uses Eq. (3). Since the distribution of $a_{q,2}$ is symmetric around $\mathbf{E}[a_{q,2}|\text{bias}_\alpha]$ we get that the first part of the claim, i.e. if $\Delta \geq \Delta_1$ then

$$\Pr[\mathcal{T} \text{ outputs } 1] = \Pr[a_{q,2} > Np_\beta + \text{bias}_\alpha \frac{\Delta_1}{1-p_\alpha} | \text{bias}_\alpha] \geq 1/2.$$

To get the second part of the claim we use the de Moivre-Laplace theorem and approximate the binomial distribution with the normal distribution so that we can approximate the variance of the sum of two distributions (when α holds and when α does not hold) in order to obtain the variance of $a_{q,2}$ conditioned on bias_α. We get:

$$\mathbf{Var}[a_{q,2}|\text{bias}_\alpha] \approx (Np_\alpha+\text{bias}_\alpha)p_{\beta|\alpha}(1-p_{\beta|\alpha})+(N(1-p_\alpha)-\text{bias}_\alpha)p_{\beta|\bar{\alpha}}(1-p_{\beta|\bar{\alpha}}).$$

Assuming N is large enough, we can neglect the terms involving bias_α. Hence,

$$\begin{aligned}\mathbf{Var}[a_{q,2}|\text{bias}_\alpha] &\approx N[p_\alpha p_{\beta|\alpha} + (1-p_\alpha)p_{\beta|\bar{\alpha}}] - N[p_\alpha p^2_{\beta|\alpha} + (1-p_\alpha)p^2_{\beta|\bar{\alpha}}] \\ &\approx Np_\beta - N[p_\alpha p^2_{\beta|\alpha} + (1-p_\alpha)p^2_{\beta|\bar{\alpha}}] \\ &= N[p_\beta - p_\beta^2] - N\Delta^2 \frac{p_\alpha}{1-p_\alpha} < N[p_\beta - p_\beta^2] = \mathbf{Var}_\beta.\end{aligned}$$

The transition from the second to third lines follows from $[p_\alpha p^2_{\beta|\alpha}+(1-p_\alpha)p^2_{\beta|\bar{\alpha}}] - p_\beta^2 = \Delta^2 \frac{p_\alpha}{1-p_\alpha}$. [7]

We have that the probability distribution on $a_{q,2}$ is a Gaussian with mean and variance at most $Np_\beta + \text{bias}_\alpha(\Delta_1 - \varepsilon)/(1-p_\alpha)$ and $\mathbf{Var}_\beta$ respectively. To conclude the proof, we note that the conditional probability mass of $a_{q,2}$ exceeding its own mean by $\varepsilon \cdot \text{bias}_\alpha/(1-p_\alpha) > \varepsilon\sigma_\alpha/(1-p_\alpha)$ is at most

$$\frac{1}{2} - \gamma = \Phi\left(-\frac{\varepsilon\sigma_\alpha/(1-p_\alpha)}{\sqrt{\mathbf{Var}_\beta}}\right)$$

where Φ is the cumulative distribution function for the normal distribution. For constant ε this yields a constant advantage γ. For $\varepsilon = o(1)$, we get that $\gamma \geq \frac{\varepsilon}{2} \frac{\sigma_\alpha/(1-p_\alpha)}{\sqrt{\mathbf{Var}_\beta}\sqrt{2\pi}}$.

By taking $\varepsilon = \omega(1/\sqrt{n})$ we can run the Test procedure enough times to determine with sufficiently high confidence which "side" of the interval $[\Delta_1 - \varepsilon, \Delta_1]$ Δ is on (if it is not inside the interval). We proceed by binary search to narrow in on Δ. We get:

Theorem 3. *There exists an algorithm that invokes the test $\mathcal{T}$*

$$O_{p_\alpha,p_\beta}(\log(1/\epsilon)\frac{\log(1/\delta)+\log\log(1/\epsilon)}{\epsilon^2})$$

times and outputs $\hat{\Delta}$ such that $\Pr[|\hat{\Delta} - \Delta| < \varepsilon] \geq 1 - \delta$.

6 Datamining on Published Statistics

In this section we apply our basic technique for measuring implication in probability to the real-life model in which confidential information is gathered by

[7] In more detail: $[p_\alpha p^2_{\beta|\alpha} + (1-p_\alpha)p^2_{\beta|\bar{\alpha}}] - p^2_\beta = p^2_{\beta|\alpha}p_\alpha(1-p_\alpha) + p^2_{\beta|\bar{\alpha}}(1-p_\alpha)p_\alpha - 2p_\alpha(1-p_\alpha)p_{\beta|\alpha}p_{\beta|\bar{\alpha}} = p_\alpha(1-p_\alpha)[p^2_{\beta|\alpha}+p^2_{\beta|\bar{\alpha}}-2p_{\beta|\alpha}p_{\beta|\bar{\alpha}}] = p_\alpha(1-p_\alpha)(p_{\beta|\alpha}-p_{\beta|\bar{\alpha}})^2 = \Delta^2\frac{p_\alpha}{1-p_\alpha}$.

a trusted party, such as the census bureau, who publishes aggregate statistics. The published statistics are the results of queries to a SuLQ database. That is, the census bureau generates queries and their noisy responses, and publishes the results.

Let k denote the number of attributes (columns). Let $\ell \leq k/2$ be fixed (typically, ℓ will be small; see below). For every ℓ-tuple of attributes $(\alpha_1, \alpha_2, \ldots, \alpha_\ell)$, and for each of the 2^ℓ conjunctions of literals over these ℓ attributes, ($\bar{\alpha}_1\bar{\alpha}_2 \ldots \bar{\alpha}_\ell$, $\bar{\alpha}_1\bar{\alpha}_2 \ldots \alpha_\ell$, and so on), the bureau publishes the result of some number t of queries on these conjunctions. More precisely, a query set $q \subseteq [n]$ is selected, and noisy statistics for all $\binom{k}{\ell}2^\ell$ conjunctions of literals are published for the query. This is repeated t times.

To see how this might be used, suppose $\ell = 3$ and we wish to learn if $\alpha_1\alpha_2\alpha_3$ implies $\bar{\alpha}_4\bar{\alpha}_5\alpha_6$ in probability. We know from the results in Section 4 that we need to find a heavy set q for $\alpha_1\alpha_2\alpha_3$, and then to query the database on the set q with the function $\bar{\alpha}_4\bar{\alpha}_5\alpha_6$. Moreover, we need to do this several times (for the binary search). If t is sufficiently large, then with high probability such query sets q are among the t queries. Since we query all triples (generally, ℓ-tuples) of literals for each query set q, all the necessary information is published. The analyst need only follow the instructions for learning the strength Δ of the implication in probability $\alpha_1\alpha_2\alpha_3 \stackrel{\Delta}{\rightarrow} \bar{\alpha}_4\bar{\alpha}_5\alpha_6$, looking up the results of the queries (rather than randomly selecting the sets q and submitting the queries to the database).

As in Section 4, once we can determine implication in probability, it is easy to determine (via Bayes' rule) the statistics for the conjunction $\alpha_1\alpha_2\alpha_3\bar{\alpha}_4\bar{\alpha}_5\alpha_6$. In other words, *we can determine the approximate statistics for any conjunction of 2ℓ literals of attribute values.* Now the procedure for arbitrary 2ℓ-ary functions is conceptually simple. Consider a function of attribute values $\beta_1 \ldots \beta_{2\ell}$. The analyst first represents the function as a truth table: for each possible 2ℓ-tuple of literals over $\beta_1 \ldots \beta_{2\ell}$ the function has value either zero or one. Since these conjunctions of literals are mutually exclusive, the probability (overall) that the function has value 1 is simply the sum of the probabilities that each of the positive (one-valued) conjunctions occurs. Since we can approximate each of these statistics, we obtain an approximation for their sum. Thus, we can approximate the statistics for each of the $\binom{k}{2\ell}2^{2^{2\ell}}$ Boolean functions of 2ℓ attributes. It remains to analyze the quality of the approximations.

Let $T = o(n)$ be an upper bound on the number of queries permitted by the SuLQ database algorithm, e.g., $T = O(n^c), c < 1$. Let k and ℓ be as above: k is the total number of attributes, and statistics for ℓ-tuples will be published. Let ε be the (combined) additive error achieved for all $\binom{k}{2\ell}2^{2\ell}$ conjuncts with probability $1 - \delta$.

Input: a database $d = \{d_{i,j}\}$ of dimensions $n \times k$.
Repeat t times:

1. Let $q \in_R [n]$. Output q.
2. For all selections of ℓ indices $1 \le j_1 < j_2 < \ldots < j_\ell \le k$, output $\hat{a}_{q,g}$ for all the 2^ℓ conjuncts g over the literals $\alpha_{j_1}, \ldots, \alpha_{j_\ell}$.

Privacy is preserved as long as $t \cdot \binom{k}{2\ell} 2^{2\ell} \le T$ (Theorem 1). To determine utility, we need to understand the error introduced by the summation of estimates. Let $\varepsilon' = \varepsilon/2^{2\ell}$. If our test results in a ε' additive error for each possible conjunct of 2ℓ literals, the truth table method described above allows us to compute the frequency of every function of 2ℓ literals within additive error ε (a lot better in many cases). We require that our estimate be within error ε' with probability $1 - \delta'$ where $\delta' = \delta/\binom{k}{2\ell}2^{2\ell}$. Hence, the probability that a 'bad' conjunct exists (for which the estimation error is not within ε') is bounded by δ.

Plugging δ' and ε' into Theorem 3, we get that for each conjunction of ℓ literals, the number of subsets q on which we need to make queries is

$$t = O\left(2^{4\ell}(\log(1/\epsilon) + \ell)(\log(1/\delta) + \ell \log k + \log\log(1/\epsilon))/\epsilon^2\right).$$

For each subset q we query each of the $\binom{k}{\ell}2^\ell$ conjuncts of ℓ attributes. Hence, the total number of queries we make is

$$t \cdot \binom{k}{\ell} 2^\ell = O\left(k^\ell 2^{5\ell}(\log(1/\epsilon) + \ell)(\log(1/\delta) + \ell \log k + \log\log(1/\epsilon))/\epsilon^2\right).$$

For constant ϵ, δ we get that the total number of queries is $O(2^{5\ell}k^\ell\ell^2 \log k)$. To see our gain, compare this with the naive publishing of statistics for all conjuncts of 2ℓ attributes, resulting in $\binom{k}{2\ell}2^{2\ell} = O(k^{2\ell}2^{2\ell})$ queries.

7 Open Problems

Datamining of 3-ary Boolean Functions. Section 5.1 shows how to use two SuLQ databases to learn that $\Pr[\beta|\alpha] = \Pr[\beta] + \Delta$. As noted, this allows estimating $\Pr[f(\alpha, \beta)]$ for any Boolean function f. Consider the case where there exist three SuLQ databases for attributes α, β, γ. In order to use our test procedure to compute $\Pr[f(\alpha, \beta, \gamma)]$, one has to either to find heavy sets for $\alpha \wedge \beta$ (having bias of order $\Omega(\sqrt{n})$), or, given a heavy set for γ, to decide whether it is also heavy w.r.t. $\alpha \wedge \beta$. It is not clear how to extend the test procedure of Section 5.1 in this direction.

Maintaining Privacy for all Possible Functions. Our privacy definition (Definition 1) requires for every function $f(\alpha_1, \ldots, \alpha_k)$ that with high probability the confidence gain is limited by some value δ. If k is small (less than $\log\log n$), then, via the union bound, we get that with high probability the confidence gain is kept small for all the 2^{2^k} possible functions.

For large k the union bound does not guarantee simultaneous privacy for all the 2^{2^k} possible functions. However, the privacy of a randomly selected function is (with high probability) preserved. It is conceivable that (e.g. using cryptographic measures) it is possible to render infeasible the task of finding a function f whose privacy was breached.

Dependency Between Database Records. We explicitly assume that the database records are chosen independently from each other, according to some underlying distribution $\mathcal{D}$. We are not aware of any work that does not make this assumption (implicitly or explicitly). An important research direction is to come up with definition and analysis that work in a more realistic model of weak dependency between database entries.

References

1. D. Agrawal and C. Aggarwal, *On the Design and Quantification of Privacy Preserving Data Mining Algorithms*, Proceedings of the 20th Symposium on Principles of Database Systems, 2001.
2. N. R. Adam and J. C. Wortmann, *Security-Control Methods for Statistical Databases: A Comparative Study*, ACM Computing Surveys 21(4): 515-556 (1989).
3. R. Agrawal and R. Srikant, *Privacy-preserving data mining*, Proc. of the ACM SIGMOD Conference on Management of Data, pp. 439–450, 2000.
4. S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, *Toward Privacy in Public Databases*, submitted for publication, 2004.
5. I. Dinur and K. Nissim, *Revealing information while preserving privacy*, Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 202-210, 2003.
6. G. Duncan, Confidentiality and statistical disclosure limitation. In N. Smelser & P. Baltes (Eds.), International Encyclopedia of the Social and Behavioral Sciences. New York: Elsevier. 2001
7. A. V. Evfimievski, J. Gehrke and R. Srikant, *Limiting privacy breaches in privacy preserving data mining*, Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 211-222, 2003.
8. S. Fienberg, Confidentiality and Data Protection Through Disclosure Limitation: Evolving Principles and Technical Advances, IAOS Conference on Statistics, Development and Human Rights September, 2000, available at `http://www.statistik.admin.ch/about/international/fienberg_final_paper.doc`
9. S. Fienberg, U. Makov, and R. Steele, *Disclosure Limitation and Related Methods for Categorical Data, Journal of Official Statistics, 14*, pp. 485–502, 1998.
10. L. Franconi and G. Merola, *Implementing Statistical Disclosure Control for Aggregated Data Released Via Remote Access*, Working Paper No. 30, United Nations Statistical Commission and European Commission, joint ECE/EUROSTAT work session on statistical data confidentiality, April, 2003, available at `http://www.unece.org/stats/documents/2003/04/confidentiality/wp.30.e.pdf`
11. S. Goldwasser and S. Micali, *Probabilistic Encryption and How to Play Mental Poker Keeping Secret All Partial Information*, STOC 1982: 365-377
12. T.E. Raghunathan, J.P. Reiter, and D.B. Rubin, *Multiple Imputation for Statistical Disclosure Limitation*, *Journal of Official Statistics 19*(1), pp. 1 – 16, 2003
13. D.B. Rubin, *Discussion: Statistical Disclosure Limitation*, *Journal of Official Statistics 9*(2), pp. 461 – 469, 1993.
14. A. Shoshani, *Statistical databases: Characteristics, problems and some solutions*, Proceedings of the 8th International Conference on Very Large Data Bases (VLDB'82), pages 208–222, 1982.