WP.19
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (ii): Tabular data protection

DIFFERENTIALLY PRIVATE MARGINALS RELEASE WITH MUTUAL CONSISTENCY AND ERROR INDEPENDENT OF SAMPLE SIZE

Invited Paper

Prepared by Cynthia Dwork, Frank McSherry, Kunal Talwar, Microsoft Research

Differentially Private Marginals Release with Mutual Consistency and Error Independent of Sample Size

Cynthia Dwork, Frank McSherry, Kunal Talwar
Microsoft Research, Silicon Valley,
1065 La Avenida Mountain View, CA, 94043, USA.
{dwork,mcsherry,kunal}@microsoft.com

Abstract. We report on a result of Barak *et al.* on a privacy-preserving technology for release of mutually consistent multi-way marginals [1]. The result ensures *differential privacy*, a mathematically rigorous notion for privacy-preserving statistical data analysis capturing the intuition that essentially no harm can befall a respondent who accurately reports her data beyond that which would befall her should she refuse to respond, or respond completely inaccurately [7, 5].

In addition to differential privacy, the techniques described herein ensure consistency among released tables and, in many cases, excellent accuracy.

1 Introduction

In this note we describe a method developed by Barak *et al.* for privacy-preserving contingency table release [1]. This was part of an ongoing project in Microsoft Research on privacy-preserving data analysis, and relies on earlier project contributions both for the definition of privacy and for some of the techniques used. An important feature of this ongoing effort is that domain-specific knowledge or expertise is *not* required for ensuring privacy. Thus, using our techniques, one does not need to be an expert on the American population, or an expert on the availability of other databases produced by official and commercial parties, and so on, to “safely” compute and release useful statistics.

Our approach to privacy-preserving data mining has its roots in cryptography. In specifying a cryptographic primitive one must formally define what it means to break the primitive – intuitively, what is the adversary’s goal? – and delineate to what resources – computataional power and auxiliary information – the adversary may have access.A rigorous pursuit of an *ad omnia* definition of privacy, taking into consideration auxiliary information, led to the discovery that (at least one natural formalization of) Dalenius’ goal, to wit, that anything learnable about a respondent, given access to a statistical database, should be learnable without access to the database [3], is provably not achievable [5]. This led us to an alternative, but still *ad omnia*, goal, *differential privacy*, which captures the intuition that essentially no harm can befall a respondent who accurately reports her data beyond what would befall her should she refuse to respond, or should she respond completely inaccurately [7, 5].

Since our language may be non-standard in the statistics community, we begin with an informal description of the problem, clarifying what *we* mean by the terms "contingency table" and "marginal."

1.1 Contingency Table Release

Informally, a contingency table is a table of counts. In the context of a census or other survey, we think of the data of an individual as a *row* in a database. For the present, each row consists of k bits describing the values of k binary attributes $a_1, \ldots, a_k$.[1] Formally, the contingency table is a vector in $\mathbb{R}^{2^k}$ describing, for each setting of the k attributes, the number of rows in the database with this setting of the attribute values.

Commonly, the contingency table itself is not released. Instead, for various sets of attributes, one releases the projection of the contingency table onto each such subset of the attributes, *i.e.*, the counts for each of the possible settings of the restricted set of attributes. These counts are called marginals, each marginal being named by a subset of the attributes. A marginal named by a set of j attributes, $j \leq k$, is called a *j-way* marginal. The data curator will typically release many sets of low-order marginals for a single contingency table, with the goal of revealing correlations between many different, and possibly overlapping, sets of attributes.

2 Differential Privacy

Many papers in the literature attempt to formalize Dalenius' goal by requiring that the adversary's prior and posterior views about an individual (*i.e.*, before and after having access to the statistical database) shouldn't be "too different," or that access to the statistical database shouldn't change the adversary's views about any individual "too much." Of course, this is clearly silly, if the statistical database teaches us anything at all. For example, suppose the adversary's (incorrect) prior view is that everyone has 2 left feet. Access to the statistical database teaches that almost everyone has one left foot and one right foot. The adversary now has a very different view of whether or not any given respondent has two left feet. Even when used correctly, in a way that is decidedly not silly, this prior/posterior approach suffers from definitional awkwardness [9, 8, 2].

The real difficulty in achieving Dalenius' goal is posed by what cryptographers call *auxiliary information.* This is any information available to the adversary/user

[1] Typically, attributes are non-binary. Any attribute with m possible values can be decomposed into $\log(m)$ binary attributes; see the discussion in Section 5.

other than what is in the statistical database. Statisticians worry about auxiliary information too: this is precisely what is exploited in a linkage attack.

Suppose we have a statistical database that teaches average heights of population subgroups, and suppose further that it is infeasible to learn this information (perhaps for financial reasons) any other way (say, by conducting a new study). Consider the auxiliary information "The radio talk show host Terry Gross is two inches shorter than the average Lithuanian woman." Access to the statistical database teaches Terry Gross' height. In contrast, someone without access to the database, knowing only the auxiliary information, learns much less about Terry Gross' height.[2]

This brings us to an important observation: Terry Gross did not have to be a member of the database for the attack described above to be prosecuted against her. This suggests a new notion of privacy: minimize the increased risk to an individual incurred by joining (or leaving) the database. That is, we move from comparing an adversary's prior and posterior views of an individual to comparing the risk to an individual when included in, versus when not included in, the database. This new notion is called *differential privacy*.

2.1 The Definition

In the sequel, the randomized function $\mathcal{K}$ is the algorithm applied by the curator (*e.g.*, census bureau) when releasing information. So the input is the data set, and the output is the released information, or *transcript*.

Recall that we think of a database as a set of rows. We say databases D_1 and D_2 *differ in at most one element* if one is a subset of the other and the larger database contains just one additional row.

Definition 1 *A randomized function $\mathcal{K}$ gives ε-*differential privacy *if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{K})$,*

$$\Pr[\mathcal{K}(D_1) \in S] \quad \leq \quad \exp(\varepsilon) \times \Pr[\mathcal{K}(D_2) \in S], \tag{1}$$

where the probability space in each case is over the coin flips of the mechanism $\mathcal{K}$.

A mechanism $\mathcal{K}$ satisfying this definition addresses all concerns that a respondent might have about accurately contributing her personal information: even if the participant removed her data from the data set, no outputs (and thus consequences of outputs) would become significantly more or less likely. For example, if the database were to be consulted by an insurance provider before deciding whether or not to insure a given individual, then the presence or absence of that individual's data in the database will not significantly affect her chance of receiving coverage.

[2] A rigorous impossibility result generalizes and formalizes this argument, extending to essentially any notion of privacy compromise [5].

Differential privacy is therefore an *ad omnia* guarantee. It is also a very strong guarantee, since it is a statistical property about the behavior of the mechanism and therefore is independent of the computational power and auxiliary information available to the adversary/user.

Remarks: *1. The parameter ε is public. The choice of ε is essentially a social question and is beyond the scope of our work. That said, we tend to think of ε as, say,* 0.01, 0.1, *or in some cases,* $\ln 2$ *or* $\ln\ 3$. *If the probability that some bad event will occur is very small, it might be tolerable to increase it by such factors as 2 or 3, while if the probability is already felt to be close to unacceptable, then an increase of* $\mathrm{e}^{0.01} \approx 1.01$ *might be tolerable, while an increase of* e, *or even only* $\mathrm{e}^{0.1}$, *would not.*

2. Definition 1 extends to group privacy as well (and to the case in which an individual contributes more than a single row to the database). A group of c participants might be concerned that their collective data might leak information, even when a single participant's does not. Using this definition, we can bound the dilation of any probability by at most $\exp(\varepsilon c)$, *which may be tolerable for small c. Of course, the point of the statistical database is to disclose aggregate information about large groups (while simultaneously protecting individuals), so we should expect privacy bounds to disintegrate with increasing group size.*

Differential privacy provides much stronger guarantees than other privacy definitions of which we are aware [11, 12, 13, 10] (see [1] for a discussion). Differential privacy also rules out the practice of publishing a random subsample of the database. For a given row x, consider two datasets D_1 and $D_2 = D_1 \setminus \{x\}$. A sampling procedure that conceivably chooses x when the dataset is D_1 can *never* choose x when the dataset is D_2, since $x \notin D_2$, yielding a zero in the denominator in 1. Even if the subsampled row is somewhat altered, its release could violate differential privacy.

Our techniques work best, that is, introduce the least relative error, when n, the size of the dataset, is large. This is because the amount of distortion introduced for protecting privacy depends only on the set of marginals requested, and not on n. However, privacy is *always* guaranteed. This is in contrast with the case of subsampling, where the unexpressed but implicit privacy guarantee depends on n.

Strong as it is, differential privacy is not an absolute guarantee of privacy. As we have seen, any statistical database with any non-trivial utility can be used to compromise privacy, even of people not in the database. However, in a society that has decided that the benefits of certain databases outweigh the costs, differential privacy ensures that only a limited amount of additional risk is incurred by participating in the (socially beneficial) databases.

3 Differentially Private Data Analysis

The *sensitivity* of a function f acting on a data set is defined in [7] as

Definition 2 [7]. *For $f : \mathcal{D} \rightarrow \mathrm{I\!R}^d$, the* L_1-sensitivity *of f is*

$$\Delta f \quad = \quad \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \tag{2}$$

for all D_1, D_2 differing in at most one element.

Note that sensitivity is a property of the function alone, and is independent of the actual database held by the curator.

Intuitively, the sensitivity of the function f describes the degree of uncertainty that must be present in the released approximation to the value of f when applied to the specific database held by the curator, in order to hide the presence or absence of any individual. This is captured by Theorem 1 below, connecting sensitivity to the amount of noise that suffices to ensure ε-differential privacy.

Theorem 1 [7]. *For any $f : \mathcal{D} \rightarrow \mathrm{I\!R}^d$, the addition of independent, symmetric, exponential noise with variance $2\sigma^2$ ensures $(\Delta f/\sigma)$-differential privacy.*

Remark: Taking $\sigma = \Delta/\epsilon$ ensures ϵ-differential privacy for a query of sensitivity Δ.

Note that application of the theorem yields a *process* for releasing statistics in general, and marginals in particular. The *process* ensures differential privacy. It makes no sense to speak of the privacy of a specific set of marginals that are released.

An application of Theorem 1 that is of particular interest in the context of contingency tables is to the class of *histogram* queries. A histogram query is an arbitrary partitioning of the domain of database rows into disjoint “cells,” and the true answer is the set of counts describing, for each cell, the number of database rows in this cell. Although a histogram query with d cells may be viewed as d individual counting queries, the addition or removal of a single database row can affect the entire d-tuple of counts in at most one location (the count corresponding to the cell to (from) which the row is added (deleted); moreover, the count of this cell is affected by at most 1, so by Definition 2, every histogram query has sensitivity 1.

Since a contingency table is a histogram, this means that we can add independently generated noise proportional to ε^{-1} to each cell of the contingency table to obtain an ε-differentially private (non-integer and not necessarily non-negative) table. We will address the question of integrality and non-negativity later. For now, we simply note that any desired set of marginals can be computed directly from this noisy table, and consistency among the different marginals is immediate. A drawback of this approach, however, is that while the noise in each cell of the contingency table is relatively small, the noise in the computed marginals may be large. For example, the variance in the 1-way table describing attribute a_1 is $2^{k-1}\varepsilon^{-2}$. We consider this unacceptable, especially when $n \ll 2^k$.

Marginals are also histograms. A second approach, with much less noise, but not offering consistency of marginals, works as follows. Let C be the set of marginals to be released. We can think of a function f that, when applied to the database, yields the desired marginals. Now apply Theorem 1 with this choice of f, (adding noise to each cell in the collection of tables independently), as directed in Theorem 1, with sensitivity $\Delta f = |C|$. When n (the number of rows in the database) is large compared to $|C|\varepsilon^{-1}$, this also yields excellent accuracy. Thus we would be done, and there would be no need for the Barak *et al.* paper, if the small table-to-table inconsistencies caused by independent randomization of each (cell in each) table are not of concern, and if the user is comfortable with occasionally negative and typically non-integer cell counts.

We have no philosophical or mathematical objection to these artifacts of the privacy-enhancing technology, but in practice they can be problematic. For example, the cell counts may be used as input to other, possibly off-the-shelf, programs that anticipate positive integers, giving rise to type mismatch. It may also be confusing to lay users, say, ordinary citizens accessing the American FactFinder website.

4 Release of Marginals

The material in this Section appears in [1]. We first highlight key elements of the approach, then introduce formal notation, and finally state the results. Proofs may be found in the original paper.

4.1 Key Steps in The Solution

Apply Theorem 1 and Never Look Back. We *always* obtain privacy by applying Theorem 1 to the raw data or a possibly reversible transformation of the raw data. This gives us an intermediate object, on which we operate further, but we never again access the raw data. Since anything obtained via Theorem 1 is differentially private, any quantity computed from the intermediate object also enjoys differential privacy.

Move to the Fourier Domain. When adding noise, two natural solutions present themselves: adding noise to entries of the source table (this was our first proposal; accuracy is poor when k is large), or adding noise to the reported marginals (our second proposal; consistency is violated). A third approach begins by transforming the data into the Fourier domain. This is just a change of basis. Were we to compute all 2^k Fourier coefficients we would have a non-redundant encoding of the entire consistency table. If we were to perturb the Fourier coefficients and then convert back to the contingency table domain, we would get a (different, possibly non-integer, possibly negative) contingency table, whose "distance" (for example, L_2 distance) from the original is determined by the magnitude of the perturbations. The advantage of moving to the Fourier domain is that if only a set C of marginals is desired then we do not need the full complement of Fourier coefficients. For example, if C is the set of all 3-way marginals, then we need only the Fourier coefficients of weight at most 3 (see Section 4.4), of which there are $\binom{k}{3} + \binom{k}{2} + k + 1$. This will translate into a much less noisy set of marginals.

The Fourier coefficients needed to compute the marginals C form a model of the dataset that captures everything that can be learned from the set C of marginals. Adding noise to these coefficients as indicated by Theorem 1 and then converting back to the contingency table domain yields a procedure for generating synthetic datasets that ensures differential privacy and yet to a great (and measurable) extent captures the information in the model. This is an example of a concrete method for generating synthetic data with provable differential privacy.

Strictly speaking, we don't really need to move to the Fourier domain: we can perturb the marginals directly and then use linear programming to find a positive fractional data set, which can then be rounded. See [1] for a discussion.

Use Linear Programming We employ linear programming to obtain a non-negative, but likely non-integer, data set with (almost) the given Fourier coefficients, and then round the results to obtain an integer solution. Interestingly, the marginals obtained from the linear program are no "farther" (made precise below) from those of the noisy measurements than are the true marginals of the raw data. Consequently, the additional error introduced by the imposition of consistency is no more than the error introduced by the privacy mechanism itself.

When k is Large The linear program requires time polynomial in 2^k. When k is large this is not satisfactory. However, somewhat surprisingly, non-negativity (but not integrality) can be achieved by adding a relatively small amount to the first Fourier coefficient before moving back to the data domain. No linear program is required, and the error introduced is pleasantly small. Thus if polynomial in 2^k is an unbearable cost and one can live with non-integrality then this approach serves well. (In this case we construct the output marginals directly from the Fourier coefficients, rather than reconstructing the contingency table. See [1] for additional details.) We remark that non-integrality was a non-issue in the prototyped system mentioned above, since answers were anyway converted to percentages.

4.2 Notation and Preliminaries

For all positive integers d, $\forall x \in \mathbb{R}^d$, the L_1 norm of x is $\|x\|_1 = \sum_{i=1}^{d} |x_i|$. As noted above, and letting k denote the number of (binary) attributes, we think of the data set as a vector $x \in \mathbb{R}^{2^k}$, indexed by attribute tuples. For each $\alpha \in \{0,1\}^k$ the quantity x_α is the number of data elements with this setting of attributes. We let $n = \|x\|_1$ be the total number of tuples, or rows, in our data set.

For any $\alpha \in \{0,1\}^k$, we use $\|\alpha\|_1$ for the number of non-zero locations. We write $\beta \preceq \alpha$ for $\alpha, \beta \in \{0,1\}^k$ if every zero location in α is also a zero in β.

4.3 The Marginal Operator

We think of the computation of a set of marginals as the result of applying a *marginal operator* to the contingency table vector x. The operator $C^\alpha : \mathbb{R}^{2^k} \to \mathbb{R}^{2^{\|\alpha\|_1}}$ for $\alpha \in \{0,1\}^k$ maps contingency tables to the marginal of the attributes that are positively set in α (there are $2^{\|\alpha\|_1}$ possible settings of these attributes). We abuse notation, and only define $C^\alpha x$ at those locations β for which $\beta \preceq \alpha$: for any $\beta \preceq \alpha$, the outcome of $C^\alpha x$ at position β is the sum over those coordinates of x that agree with β on the coordinates described by α:

$$(C^\alpha(x))_\beta \;=\; \sum_{\gamma:\gamma\wedge\alpha=\beta} x_\gamma \tag{3}$$

Notice that the operator C^α is linear for all α.

It is common to consider the ensemble of marginal operators C^α for all α with a fixed value of $\|\alpha\|_1 = j$, referred to as the j-way marginals. For example, when $j = 3$ this is the ensemble of marginal operators C^α for all $\alpha \in \{0,1\}^k$ containing exactly 3 ones, *i.e.*, the ensemble of all 3-way marginals.

4.4 The Fourier Basis

We will find it helpful to view our contingency table x in an alternate basis; rather than a value for each position α, we will project onto a set of 2^k so-called *Fourier basis* vectors, each of which aggregates across the table in a different way. Our motivation lies in the observation, made formally soon, that while a marginal depends on all coordinates of the contingency table, a low-order marginals (that is, C^α when $\|\alpha\|_1$ is small) depends on only a few of the new coordinates in the Fourier basis.

The Fourier basis for real vectors defined over the Boolean hypercube is the set of vectors f^α for each $\alpha \in \{0,1\}^k$, defined coordinate-wise as

$$f^\alpha_\beta \;=\; (-1)^{\langle\alpha,\beta\rangle}/2^{k/2}\ . \tag{4}$$

That is, each vector comprises coordinates of the form $\pm 1/2^{k/2}$, with the sign determined by the parity of the intersection between α and β.

The following theorem is well known.

Theorem 2 *The f^α form an orthonormal basis for $\mathbb{R}^{2^k}$.*

The projection $\langle f^\alpha, x\rangle f^\alpha$ of a vector x onto a Fourier basis vector f^α is referred to as a Fourier coefficient. The following theorem says that any marginal over few attributes requires only a few Fourier coefficients.

Theorem 3 *$C^\beta f^\alpha \neq \boldsymbol{0}$ if and only if $\alpha \preceq \beta$.*

Consequently, we are able to write any marginal as the small summation over relevant Fourier coefficients:

$$C^\beta x \;=\; \sum_{\alpha\preceq\beta} \langle f^\alpha, x\rangle C^\beta f^\alpha\ . \tag{5}$$

4.5 Algorithms

To apply Theorem 1 in the Fourier domain, we need only bound the sensitivity of each Fourier coefficient, since a straightforward argument shows that the sensitivity of a collection of coefficients is bounded by the number of coefficients in the collection times the sensitivity of any one coefficient. Formally, let $Lap(\sigma)$ be a random variable with density at x proportional to $\exp(-|x|/\sigma)$.

Theorem 4 *Let $B \subseteq \{0,1\}^k$ describe a set of Fourier basis vectors. Releasing the set $\phi_\beta = \langle f^\beta, x\rangle + Lap(|B|/\epsilon 2^{k/2})$ for $\beta \in B$ preserves ϵ-differential privacy.*

Proof: Each tuple contributes exactly $\pm 1/2^{k/2}$ to each output coordinate, and consequently the L_1 sensitivity of the set of $|B|$ outputs is at most $|B|/2^{k/2}$. By Theorem 1, the addition of symmetric exponential noise with standard deviation $|B|/\epsilon 2^{k/2}$ gives ϵ-differential privacy. QED.

Remark: To get a sense of scale, we could achieve a similar perturbation to each coordinate by randomly adding or deleting $|B|^2/\epsilon$ individuals in the data set, which can be much smaller than n.

4.5.1 Non-Negative Integrality

Consider the set of (now noisy) Fourier coefficients $\{\phi_\beta \,|\, \beta \in B\}$ released in Theorem 4. While there certainly exists a real valued contingency table whose Fourier coefficients equal these released values, it is unlikely that there is a non-negative, integral contingency table with these coefficients. We use linear programming to find a non-negative, but likely fractional, contingency table with nearly the correct Fourier coefficients, which we round to an integral table with little additional error.

Imagining that we observed (noisy) Fourier coefficients ϕ_β, the linear program in the next section minimizes, over all contingency tables w, the largest error b between its Fourier coefficients $\langle f^\beta, w\rangle$ and the observed ϕ_β's.

4.5.2 Putting the Steps Together

We now collect the various steps. Recall that to compute a marginal $\alpha \in \{0,1\}^k$ we require the Fourier coefficients $\{f^\beta \,|\, \beta \preceq \alpha\}$. Thus, to compute a set A of marginals, we need all the Fourier coefficients f^β for β in the downward closure of A uner $\preceq$.

Marginals($A \subseteq \{0,1\}^k, D$):

1. Let B be the downward closure of A under $\preceq$.

2. For $\beta \in B$, compute $\phi_\beta = \langle f^\beta, D\rangle + Lap(|B|/\epsilon 2^{k/2})$.

3. Solve for w_α in the following linear program, and round to the nearest integral weights, w'_α.

$$\begin{array}{rcl}
\text{minimize} & b & \\
\text{subject to:} & & \\
w_\alpha & \geq & 0 \quad \forall\alpha \\
\phi_\beta - \sum_\alpha w_\alpha f_\alpha^\beta & \leq & b \quad \forall\beta \in B \\
\phi_\beta - \sum_\alpha w_\alpha f_\alpha^\beta & \geq & -b \quad \forall\beta \in B
\end{array}$$

4. Using the contingency table w'_α, compute and return the marginals for A.

Theorem 5 *Using the notation of **Marginals**(A), with probability $1-\delta$, for all $\alpha \in A$,*

$$\|C^\alpha x - C^\alpha w'\|_1 \quad \leq \quad 2^{\|\alpha\|_1} 2|B| \log(|B|/\delta)/\epsilon + |B| \; . \tag{6}$$

Why the *Fourier* Basis? The Fourier coefficients exactly describe the information required by the marginals. By measuring exactly what we need, we add the least amount of noise possible using the techniques of [7]. Moreover, the Fourier basis is particularly attractive because of the natural decomposition according to sets of attribute values. In particular, even tighter bounds than in Theorem 5 can be placed on sub-marginals (that is, lower order marginals) of a marginal C^α, by noting that the bounds for the marginals C^β, where $\beta \preceq \alpha$, are obtained at no additional cost. No more Fourier coefficients are required, so $|B|$ is not increased, but $\|\beta\|_1 \leq \|\alpha\|_1$.

4.5.3 Simple Non-Negativity

The solution of the linear programs we have described is an expensive process, taking time polynomial in 2^k. In many settings, but not all, this is an excessive amount, and must be avoided. We now describe a very simple technique for arriving at Fourier coefficients corresponding to a non-negative, but fractional, contingency table with high probability, without the solution of a linear program. As noted above, we construct the output marginals directly from the Fourier coefficients, rather than reconstructing the contingency table.

Ensuring the existence of a non-negative contingency table with the observed Fourier coefficients turns out to be easy: we simply add a small amount to the first Fourier coefficient. Intuitively, any negativity due to the small perturbation we have made to the Fourier coefficients is spread uniformly across all elements of the contingency table. Consequently, very little needs to be added to make the elements non-negative.

Theorem 6 *Let x be a non-negative contingency table with d Fourier coefficients ϕ_α. If the Fourier coefficients are perturbed to ϕ'_α, then the contingency table*

$$x' \; = \; x + \sum_\alpha (\phi'_\alpha - \phi_\alpha) f^\alpha + \|\phi' - \phi\|_1 f^{\overline{0}} \tag{7}$$

is non-negative, and has $\langle f^\alpha, x' \rangle = \phi'_\alpha$ for $\alpha \neq \overline{0}$.

It is *critical* that we not disclose the actual L_1 norm of the perturbation, but we can add a value for which the negativity probability is arbitrarily low:

Corollary 1 *By adding $t \times d^2/\epsilon 2^{k/2}$ to the first Fourier coefficient, the resulting contingency table is non-negative with probability at least $1 - \exp(-t)$.*

5 Discussion

Non-Binary Attributes. We have assumed that attributes are binary-valued. While it is possible to convert a non-binary valued attribute to vector of binary attributes, this introduces some inefficiency, for example, we need 2 bits to describe a three-valued attribute, and 5 bits to describe a 17-valued attribute. The linear program can be modified to force an outcome of 0 in these "structurally zero" locations, but for each such attribute we may be increasing the size of the linear program by a factor of almost 2. Even without linear programming, some care must be taken to redistribute any non-zero weight assigned to these locations (after adding noise to the Fourier coefficients and converting back to the space of marginals) to the "real" locations.

When the number of attributes is large, the error introduced in each cell in the set of all low-order marginals, although independent of n, the size of the population, is still substantial. For example, in the case of the set of all 3-way marginals, we get a value in $\Omega(k^3)/\varepsilon$, even if the attributes are binary. When k is on the order of 30, 50, or 100 the distortion is on the order of thousands or (a small number of) millions. There are several possible approaches to improving the accuracy.

Gaussian Noise. First, we can use a Gaussian distribution on noise. The analysis in this case is more involved, and the nature of the privacy guarantee is slightly different. Instead of ε-differential privacy we obtain (ε, δ)-differential privacy (see [6], where this is called *δ-approximate ε-differential privacy*). Roughly speaking, this says that, with all but probability δ, ε-differential privacy is ensured. Using Gaussian noise with distribution $G(x) \propto \exp(-x^2/2R)$ we get (ε, δ) differential privacy when $\varepsilon \geq [2\log(1/\delta)/R]^{1/2}$. The advantage of this approach is that it allows us to improve the dependence on k of the distortion in each cell to $O(k^{3/2})$.

Partitioning into "Sensitive" and "Insensitive" Attributes. The literature frequently discusses "sensitive" and "insensitive" attributes. We prefer to avoid such distinctions; they are often flawed and they rely on domain-specific information. However, there is a way to exploit the distinction, were one to accept its validity: only add noise to Fourier coefficients in the downward closure of any requested marginal α containing at least one sensitive attribute. For example, suppose only attribute a_1 is sensitive. If the requested set C is the ensemble of all 3-way marginals, we need only add noise to $\binom{k-1}{2} + \binom{k}{2} + k + 1$ Fourier coefficients, rather than to all $\binom{k}{3} + \binom{k}{2} + k + 1$, and we may scale down the magnitude of the noise accordingly.

Lower Bounds on Noise. Finally we remark that, at least theoretically, there is no way to avoid the dependence on k in the presence of even one sensitive attribute. In the admittedly contrived setting in which there are $k = n \log^2 n + 1$ attributes, and (for some reason) for each tuple independently, for each attribute independently, the attribute is set with probability 1/2, it is possible to compute, in time polynomial in n, a candidate vector c that agrees with the vector v of values of the secret, kth, attribute, in all but $o(n)$ locations, assuming only that the magnitude of the noise added to each marginal is $o(\sqrt{n})$ [4]. If the adversary is not restricted to polynomial time then the attack works whenever the noise in each marginal is $o(n)$.

References

[1] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In L. Libkin, editor, *PODS*, pages 273–282. ACM, 2007.

[2] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In C. Li, editor, *PODS*, pages 128–138. ACM, 2005.

[3] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistisk. tidskrift*, 3:213–225, 1977.

[4] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *PODS '03: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, New York, NY, USA, 2003. ACM.

[5] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *ICALP (2)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

[6] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In S. Vaudenay, editor, *EUROCRYPT*, volume 4004 of *Lecture Notes in Computer Science*, pages 486–503. Springer, 2006.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *TCC*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

[8] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In M. K. Franklin, editor, *CRYPTO*, volume 3152 of *Lecture Notes in Computer Science*, pages 528–544. Springer, 2004.

[9] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS '03: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222, New York, NY, USA, 2003. ACM.

[10] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In L. Liu, A. Reuter, K.-Y. Whang, and J. Zhang, editors, *ICDE*, page 24. IEEE Computer Society, 2006.

[11] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

[12] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[13] X. Xiao and Y. Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In C. Y. Chan, B. C. Ooi, and A. Zhou, editors, *SIGMOD Conference*, pages 689–700. ACM, 2007.