Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release

Boaz Barak[*] [†]
Princeton University
boaz@cs.princeton.edu

Kamalika Chaudhuri[*]
UC Berkeley
kamalika@cs.berkeley.edu

Cynthia Dwork
Microsoft Research, SVC
dwork@microsoft.com

Satyen Kale[*]
Princeton University
satyen@cs.princeton.edu

Frank McSherry
Microsoft Research, SVC
mcsherry@microsoft.com

Kunal Talwar
Microsoft Research, SVC
kunal@microsoft.com

ABSTRACT

The contingency table is a work horse of official statistics, the format of reported data for the US Census, Bureau of Labor Statistics, and the Internal Revenue Service. In many settings such as these privacy is not only ethically mandated, but frequently legally as well. Consequently there is an extensive and diverse literature dedicated to the problems of statistical disclosure control in contingency table release. However, all current techniques for reporting contingency tables fall short on at least one of privacy, accuracy, and consistency (among multiple released tables). We propose a solution that provides strong guarantees for all three desiderata simultaneously.

Our approach can be viewed as a special case of a more general approach for producing synthetic data: Any privacy-preserving mechanism for contingency table release begins with raw data and produces a (possibly inconsistent) privacy-preserving set of marginals. From these tables alone – and hence without weakening privacy – we will find and output the "nearest" consistent set of marginals. Interestingly, this set is no farther than the tables of the raw data, and consequently the additional error introduced by the imposition of consistency is no more than the error introduced by the privacy mechanism itself.

The privacy mechanism of [20] gives the strongest known privacy guarantees, with very little error. Combined with the techniques of the current paper, we therefore obtain excellent privacy, accuracy, and consistency among the tables. Moreover, our techniques are surprisingly efficient.

Our techniques apply equally well to the logical cousin of the contingency table, the OLAP cube.

Categories and Subject Descriptors

G.3 [**PROBABILITY AND STATISTICS**]: [Contingency table analysis]

General Terms

Algorithms, Theory

Keywords

Privacy, OLAP, Contingency table release

1. INTRODUCTION

Privacy-preserving data-mining, also known as statistical disclosure control, has historically been the purview of statisticians, both in practice, for example, at the Census Bureau, the Internal Revenue Service, and the Bureau of Labor Statistics, and in the research community. (see, for example, [12, 24, 17, 34, 1] and the references in Section 1.1 below). In recent years the topic has experienced a resurgence in the computer science community (see, for example, [2, 3, 4, 23, 14, 9, 10, 6, 20, 18, 37, 35]). In this work we focus on contingency tables, also known as frequency tables, and their logical cousins, On-Line Analytical Processing (OLAP) cubes.

1.1 Contingency Table Release

Informally, a contingency table is a table of counts. From a database consisting of n rows, each comprising values for a fixed set of, say, binary attributes $a_1, \ldots, a_k$, the contingency table is the histogram of counts for each of the 2^k possible settings of these attributes. Contingency tables are essentially equivalent to OLAP cubes, which cast traditional relational databases as a high-dimensional cube with dimensions corresponding to the attributes. While we stay with the notation of statisticians, we stress that this is simply notational, and the results can be directly mapped to privacy-preserving OLAP.[1]

What is commonly released is not a contingency table itself, but the projection of the cube onto a subset of the at-

[*]Research conducted while visiting Microsoft Research.

[†]Supported by NSF grants CNS-0627526 and CCF-0426582 and US-Israel BSF grant 2004288.

PODS'07, June 11–14, 2007, Beijing, China.

[1]Typically, attributes are non-binary. While our exposition uses binary attributes, any attribute with m possible values can be decomposed into $\log(m)$ binary attributes. This is even natural in many OLAP settings, where the attributes are hierarchically organized.

tributes: the counts for each of the possible settings of the restricted set of attributes. These counts are called marginals, each marginal associated with a subset of the attributes, and called and j-way marginals when at most $j \leq k$ attributes are used. 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.

At first glance, it might seem that low-order marginals are "naturally" privacy-preserving: after all, they are aggregations over many of database rows. This, however, is not the case. For example, small counts are considered disclosive: if a given pair of attribute values corresponds to a unique individual, then these fields can be used as a key in other databases to reveal further information about an individual. Empty cells are also potentially disclosive: while they do not point to a specific individual, they can permit the rejection of claims, eg: that student X received all 'A's by virtue of the fact that no student received such marks. Access to large counts over time can permit "differencing attacks", where the changes to the data set can serve as the basis for inference and privacy violation. Several papers examine the degree to which individual cell entries are revealed by marginals; see, eg, [15]. Finally, recent results of de Loera and Onn [30, 29, 31] are particularly discouraging.

The disclosure risks inherent in contingency tables have given rise to an extensive and diverse literature on techniques for altering the true tables. There are two broad classes of techniques: non-perturbative (specifically, cell suppression) and perturbative (eg, controlled rounding and controlled tabular adjustment). In cell suppression, so-called "sensitive" cells are identified (the *primary cells*; see [16, 33] for a discussion of sensitivity rules). These are suppressed, together with a set of complementary *secondary* cells (to avoid the disclosure of the primary cells). The typical goal is to suppress as few secondary cells as possible, leading to difficult combinatorial problems with impractical execution times on large instances. Controlled rounding, initially introduced in [5], also suffers from combinatorial explosion, and is NP-hard even for the case of three-dimensional tables [27]. Controlled Tabular Adjustment, due to Dandekar and Cox [13], and the use of quadratic interior-point methods, due to Castro [7], were introduced to address these difficulties. This is an active area of research; see for example, the discussion in [8].

Over the last 5 years or so, the database and cryptography communities have provided rigorous definitions of privacy and introduced techniques that provably satisfy the given definitions [23, 9, 4, 14, 22, 20].

The most general and robust of these, and the notion used in this work, captures the following intuition: the adversary learns nothing *more* about an individual when her data are included in the database than the adversary can learn about the individual when her data are not included in the database (see Section 1.2 for the formal definition and its motivation) [20, 18]. Combined with the algorithmic techniques developed in a series of papers ([14, 22, 6] and particularly [20]), these yield a simple approach to contingency table release, with excellent accuracy and strong privacy guarantees, independent of any auxiliary information available to the adversary and regardless of the adversary's computational power. At a high level, this approach involves adding a small amount of independently and identically distributed noise to each cell in the released marginals. However, the small errors introduced to ensure privacy will cause distinct breakdowns of the data to yield slightly different counts (not to mention possibly negative and non-integer cell counts). The current work addresses this, adding consistency (and positivity and integrality) to privacy and accuracy.

1.2 A Formal Statement of Our Contribution

Our contribution, as suggested by the paper's title, comes in the parts of privacy, accuracy, and consistency, each of which are critical components of any data analysis system. At an intuitive level, which we soon formalize, we are concerned with

- **Privacy**: The presence or absence of any one data element should not substantially influence the distribution over outcomes of the computation.
- **Accuracy**: The difference between the reported marginals and true marginals should be bounded, preferably independent of the size of the data set.
- **Consistency**: There should exist a contingency table whose marginals equal the reported marginals.

We now formally discuss each, in the context of prior work.

1.2.1 Privacy

Since a rigorous claim about privacy is integral to our result we begin by recalling the definition of differential privacy [18, 20], which our algorithms will ensure.

DEFINITION 1. *[18, 20]. A randomized function $\mathcal{K}$ gives* ϵ-differential privacy *if for all databases D_1 and D_2 differing on at most one element, and all measurable $S \subseteq Range(\mathcal{K})$,*

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

A randomized function satisfying this definition addresses *any* concern that a participant might have about the use of his or her data. In a formal sense, the distribution over outcomes is almost as if the participant had opted out of the data set; no event is made substantially more or less likely by the use of her data. These "events" can be viewed mathematically, perhaps as outputs leading to a substantial shift between prior and posterior probabilities, or pragmatically, as actual objectionable events, eg: outputs leading to telemarketing calls or denial of credit.

Remark: Differential privacy has several consequences that follow from the definition but may not be immediately apparent. Notably, the definition is agnostic to auxiliary information an adversary may possess, and provides guarantees against arbitrary attacks. Moreover, any function with ϵ-differential privacy also ensures (ϵt)-differential privacy for groups of size up to t, and the composition of s functions with ϵ-differential privacy ensures (ϵs)-differential privacy. See [20] for further discussion.

Comparison with Other Definitions.

Differential Privacy provides much stronger guarantees than other privacy definitions of which we are aware. For example, k-anonymity[35, 37, 36] and its extension l-diversity[32] impose syntactic constraints on the outputs, requiring that

many groups of tuples appear indistinguishable, or uninformative about specific values. Nonetheless, neither of these definitions protects against even simple background knowledge of the form "My colleague Mr. R., who works in zip code 2770*, is in the database". For example, if, cumulatively, the people in the database suffer from a small set of ailments, then the adversary learns that Mr. R suffers from one of these ailments. This may be worse than embarassing; it may result in Mr. R. being dropped from consideration for promotion, say. In addition, even two k-anonymous or even ℓ-diverse tables taken together may be completely disclosive.

Similarly, [23] promotes the concept of the (ρ_1, ρ_2)-privacy breach. Very roughly, such a breach represents a substantial change in the adversary's belief that an individual data item satisfies some particular property P. By definition, this notion is entangled with the adversary's prior knowledge about the data and seems to have forced some awkward assumptions (eg, independence among data items, adversary's knowledge of true prior). Comparable definitional awkwardness appears in [22, 9, 10, 6] (eg, the "informed adversary"). Interestingly, in hindsight we find that the *algorithmic* techniques in [23, 22, 6] yield stronger privacy than is proved in the papers themselves. This occurs because all three papers provide *statistical* guarantees regarding the outputs of the privacy mechanism, yielding, to differing extents, approximations to differential privacy.

Protection against (ρ_1, ρ_2)-privacy breaches in [23] comes as a *consequence* of the γ-amplification statistical guarantees. Other algorithmic approaches, such as in [4], prevent different sets of (ρ_1, ρ_2) privacy breaches than does ensuring at most γ-amplification, and are less satisfactory. For example, the guarantees in [4] may fail to protect Mr. R.'s recent purchase of "herbal supplements" – a not uncommon event for the population in general but again, embarassing to Mr. R.

One might – erroneously – conclude that no technique can protect individuals against adversaries with arbitrary background knowledge. After all, it is formally proven in [18] that, for essentially any non-trivial mechanism and definition of privacy compromise, there exists auxiliary information for which the output of the mechanism enables a privacy compromise that would not otherwise be possible[2].

This result, and its underlying intuition, led to Differential Privacy, which *does* provide guarantees against arbitrary auxiliary information. It succeeds because it makes the only fair comparison: the probabilities of disclosure (or any event at all, for that matter) with, versus without, the sensitive data. If a disclosure will happen even without a participant's data, perhaps because it is known beforehand that the participant is in the majority, say, then it is unfair to cast blame on the privacy mechanism: any mechanism that reports the majority would lead to the breach. This is a key distinction: comparing with and without *a participant's data*, rather than with and without *the output of the mechanism*, and it is what allows Differential Privacy to give such strong bounds. Since the definition talks about the statistical distribution of the outcome, it obviates any discussion of the adversary's auxiliary information.

[2]Intuitively, the *utility* of the database provides a cryptographic one-time pad which can be combined with auxiliary information to yield a devastating privacy compromise. The user of the system learns the utility and can therefore subtract out the one-time pad, revealing the privacy compromise. Anyone not having access to the system's utility cannot "decode" the auxiliary information.

1.2.2 Accuracy

Privacy guarantees are of course meaningless without accompanying accuracy guarantees. We could easily erase the data if the former were all we cared about. We now detail guarantees that our algorithm makes about the accuracy of the counts in the released marginals, while ensuring ϵ-differential privacy.

Our theorem statement is necessarily loose at the moment, for notational reasons. The full version appears as Theorem 7, and is tighter than what is presented now:

THEOREM 1. *(Rough Version): Let C be a set of marginals of the contingency table, each on at most j attributes. We compute marginals C' of a positive, integral contingency table, preserving ϵ-differential privacy, such that with probability $1-\delta$ for any marginal $c \in C$,*

$$\|c - c'\|_1 \quad \leq \quad 2^{j+3}|C|\log(|C|/\delta)/\epsilon + |C| \; . \qquad (2)$$

This result does not depend on the total number of attributes in the data set, nor on the total number of elements in the data set, but rather only on the "complexity" of the query, in terms of the number and order of the marginals. Our result is the first we are aware of where the error in the marginals falls below statistical error due to sampling. Note also that while one might be concerned that 2^j is a large number, it is the number of elements that are reported by each marginal, and a natural scale for the L1 norm.

The most natural comparison to make is with the recent work of [4], on privacy preserving OLAP. In this work, which provides a limited form of (ρ_1, ρ_2)-privacy, the data are randomized with a constant probability, resulting in each count being reconstructible to within roughly $\sqrt{|dataset|}$. Our approach improves the error by exploiting the property that it is the number of marginals requested, $|C|$, that determines a sufficient amount of noise.

Remark: Randomized response, or any other mechanism that allows the user to learn answers to too many counting queries of the form "How many of the subset S of tuples satisfy property P?" must necessarily introduce large amounts of noise. This follows from results of [14], originally obtained for the interactive case, but applying here as well. For example, given any mechanism for which the magnitude of the error on all 2^n counting queries is bounded by E, the adversary can produce a candidate vector that agrees with $(P(tuple_1), \ldots, P(tuple_n))$ on all but $4E$ entries. So for $E \in o(n)$, say, $E = n^{1-\epsilon}$, the adversary learns more than 99.99% of the P values. An efficient version of this requires only that the adversary obtain responses to $n \log^2 n$ randomly chosen subset counting queries with $o(\sqrt{n})$ error (in fact, an efficient attack may be carried out even if more than 20% of $O(n)$ queries have wild error, while the remaining suffer from at most $o(\sqrt{n})$ error [21]). In some sense the problem is that the randomized response mechanism reveals the (very roughly approximate) answers to many more queries than the user may actually want to pose. By focusing on interactive mechanisms we can add just enough noise to ensure privacy for a given number of queries. Whenever the curator knows the questions in advance, a "transcript" can be prepared with the desired queries and responses, so for

the case of contingency tables, or OLAP cubes, we are not restricting ourselves by focusing on the interactive model.

1.2.3 Consistency

The matter of consistency among the released marginals might appear trivial; indeed most previous approaches, which produced actual randomized data sets, it is a non-issue, as their tables are produced from these specific data sets. However, there is previous work, namely [20], that assures differential privacy and strong accuracy simply by adding noise to relased cell values. It is unlikely that there exists a single data set that yields all of the released marginals, and this potential inconsistency in the released data can be the source of many technical frustrations.

As we will base our privacy and accuracy around the techniques in [20], we take this section to introduce their results and approaches, while also distinguishing our current work from theirs.

DEFINITION 2. *[20]. For $f : \mathcal{D} \to R^d$, the $L1$-sensitivity of f is*

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \quad (3)$$

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 database. In the particular case of marginals of contingency tables, which integrate counts over disjoint regions of attribute space, the $L1$-sensitivity is always two: changing a single participant's data can alter at most two counts, one old and one new.

Our interest in sensitivity is summarized by Theorem 2 below, connecting sensitivity to the amount of noise that suffices to ensure ϵ-differential privacy.

THEOREM 2. *[20]. For any $f : \mathcal{D} \to R^d$, the addition of Laplace noise[3] with variance $2\sigma^2$ preserves $(\Delta f/\sigma)$-differential privacy.*

PROOF. Using the definition of the Laplace density, the density at any a is

$$\mu[a|D] \propto \exp(\|f(D) - a\|_1/\sigma) \quad (4)$$

Applying the triangle inequality, we bound the ratio

$$\frac{\mu[a|D_1]}{\mu[a|D_2]} = \frac{\exp(\|f(D_1) - a\|_1/\sigma)}{\exp(\|f(D_2) - a\|_1/\sigma)} \quad (5)$$

$$\leq \exp(\|f(D_1) - f(D_2)\|_1/\sigma) . \quad (6)$$

The last term is bounded by $\exp(\Delta f/\sigma)$, by the definition of Δf. Thus (1) holds for singleton sets $S = \{a\}$, and the theorem follows by integrating over S. □

Remark: To ensure ϵ-differential privacy for a query of sensitivity Δ we take $\sigma = \Delta/\epsilon$.

This perturbation approach directly leads to a mechanism for releasing approximations to the marginals of the contingency table: Assume the curator wishes to release the set of marginals C. One privacy-preserving approach applies Theorem 2 to the $|C|$ marginals (adding noise to each cell in the collection of tables independently), with sensitivity $\Delta f = |C|$. This yields ϵ-differential privacy, which is a very strong guarantee. When n (the number of rows in the database) is large compared to $|C|$ this also yields excellent accuracy. Thus we would be done, and there would be no need for the current 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. We know of one real-life test case in which poor communication with the system prototypers caused users who experienced inconsistencies among OLAP cubes to question the validity of the data.[19] And while such a problem can be solved with better communication and education, it may be difficult to arrange, say, when users are ordinary citizens accessing the United States Census public interface.

1.3 Key Steps in Our Solution

Apply Theorem 2 and Never Look Back.

In this paper we *always* obtain privacy by applying Theorem 2 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 2 is privacy-preserving, any quantity computed from the intermediate object is still safe: the curator could equally well release the privacy-protective intermediate object and the user can carry out the rest of the compuatations. The results would be the same.

Move to the Fourier Domain.

When adding noise, two natural approaches present themselves: add noise to entries of the source table and compromise on accuracy, or add noise to the reported marginals and violate consistency. A third approach transforms the data into the Fourier domain, which serves as a non-redundant encoding of the information in the marginals. Adding noise in this domain will not violate consistency, because any set of Fourier coefficients corresponds to a (fractional and possibly negative) contingency table. Morever, as we will show, very few Fourier coefficients are required to compute low-order marginals, and consequently the magnitude of the noise we must add to them is small.

Use Linear Programming.

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

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

[3]The Laplace distribution is centered at zero, with exponential tails in each direction.

data set, which can then be rounded as above. The accuracy in this case suffers slightly.

When k is Large.

The linear program requires time polynomial in 2^k, which is the size of the contingency table (because that is what the linear program is solving for). When k is large this is not satisfactory. However we show, somewhat surprisingly, that 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 2^k is an unbearable cost and one can live with non-integrality then this approach serves well. We note that non-integrality was a non-issue in the prototyped system mentioned above, since answers were anyway converted to percentages.

2. NOTATION AND PRELIMINARIES

Our formal treatment of contingency table release begins by casting our data set as a vector x in a high-dimensional space, indexed by attribute tuples. Formally, imagine k binary attributes, and for each $\alpha \in \{0,1\}^k$ there is a count x_α of the number of data elements with this setting of attributes. We let $n = \|x\|_1$ be the total number of tuples in our data set. As it is likely that x will be sparse, with $n \ll 2^k$, we will be mindful of n and 2^k independently.

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

2.1 The Marginal Operator

Central to our discussion are the operators $C^\alpha : \mathbb{R}^{2^k} \to \mathbb{R}^{2^{\|\alpha\|_1}}$ for $\alpha \in \{0,1\}^k$ mapping contingency tables to the marginals of the attributes that are positively set in α. 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:

$$(C^\alpha(x))_\beta \quad = \quad \sum_{\gamma : \gamma \wedge \alpha = \beta} x_\gamma \tag{7}$$

Notice that we are abusing notation, and only defining $C^\alpha x$ at those locations β for which $\beta \preceq \alpha$.

THEOREM 3. *The operator C^α is linear for all α.*

PROOF. As each output coordinate of C_i is a sum over predetermined input coordinates, scaling and addition of its inputs translate to equivalent scaling and addition of outputs. □

It is common to consider the ensemble of marginals C^α for all α with a fixed value of $\|\alpha\|_1 = i$, referred to as the i-way marginals.

2.2 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 that each aggregate across the table in various ways. Our motivation lies in the observation, made formally soon, that while a low-order marginal needs access to all coordinates of the contingency table, it will need 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

$$f^\alpha_\beta \quad = \quad (-1)^{\langle \alpha, \beta \rangle} / 2^{k/2} \, . \tag{8}$$

That is, each Fourier basis vector is comprised of coordinates of the form $\pm 1/2^{k/2}$, with the sign alternating based on the parity of the intersection between α and β. In fact, it will occosionally be helpful to view the vectors f^α as contingency tables themselves, as we will want to apply the marginal operators C^β to them.

THEOREM 4. *The f^α form an orthonormal basis for $\mathbb{R}^{2^k}$.*

PROOF. This is a standard result from the theory of Fourier analysis. See for example, [28] □

A change of basis allows us to rewrite a vector x as a sum of its projections onto the basis vectors, each of which is referred to as a Fourier coefficient. For our purposes, we will want to rewrite x in this basis just before it is supplied as input to a marginal computation C^β, which by linearity is

$$C^\beta x \quad = \quad C^\beta \sum_\alpha \langle f^\alpha, x \rangle f^\alpha \quad = \quad \sum_\alpha \langle f^\alpha, x \rangle C^\beta f^\alpha \, . \tag{9}$$

As promised, the motivation for this transformation comes from the following theorem, that any marginal over few attributes requires only a few Fourier coefficients.

THEOREM 5. *$C^\beta f^\alpha \neq \mathbf{0}$ if and only if $\alpha \preceq \beta$.*

PROOF. For any coordinate $\gamma \preceq \beta$ of the output

$$(C^\beta(f^\alpha))_\gamma \quad = \quad \sum_{\eta : \eta \wedge \beta = \gamma} f^\alpha_\eta \quad = \quad \sum_{\eta : \eta \wedge \beta = \gamma} (-1)^{\langle \alpha, \eta \rangle} / 2^{k/2} \, . \tag{10}$$

If $\alpha \not\preceq \beta$, then there is a coordinate for which α is one and β zero. For every η in the sum above, the same string with this bit flipped is also in the sum, as $\eta \wedge \beta$ is ignorant of this bit. However, their coordinates in f^α have opposite sign, and their contributions to the sum cancel exactly. This holds for all η, making the total sum zero.

If $\alpha \preceq \beta$, then $(C^\beta(f^\alpha))_\alpha$ is non-zero, as the sum is taken over η with $\eta \wedge \beta = \alpha$, causing $\langle \eta, \alpha \rangle = \langle \alpha, \alpha \rangle$. Thus all terms contributing to the summation are positive. □

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

$$C^\beta x \quad = \quad \sum_{\alpha \preceq \beta} \langle f^\alpha, x \rangle C^\beta f^\alpha \, . \tag{11}$$

The coefficients $\langle f^\alpha, x \rangle$ are necessary and sufficient data from x for the computation of $C^\beta x$.

3. ALGORITHMS AND THEOREMS

We now delve into the details of our algorithm, which comes in two parts. We first show how to create consistent marginals by applying a privacy-preserving mechanism to the Fourier coefficients rather than directly to the marginals. The resulting Fourier coefficients may correspond to a contingency table whose entries are negative and fractional, and we then give a linear program which, after rounding, returns a positive integral contingency table, from which we compute marginals.

3.1 Consistency

Rather than perturb the marginals, a naive, but effective, manner of ensuring privacy and consistency is to simply perturb and release each coordinate of the contingency table. As low-order marginals are sums over many entries in the contingency table, their entries will have noise that is Binomially distributed with variance $\Theta(2^k)$.

Instead, we will isolate and perturb those features of the data set relevant to the marginal computation, the Fourier coefficients. Because we are taking substantially fewer measurements, as compared with 2^k above, we can add substantially less noise to each measurement. For example, we need only 2^i coefficients for a i-way marginal, and only $\sum_{j \leq i} \binom{k}{j}$ coefficients for the full set of i-way marginals. While these numbers may seem large, recall that a i-way marginal releases 2^i counts, making this the natural scale.

We use the privacy mechanism of [20], based on the addition of additive noise, to ensure ϵ-differential privacy. Formally, we let $Lap(\sigma)$ be a random variable with density at y proportional to $\exp(-|y|/\sigma)$. The following theorem describes the amount of noise we must add to each Fourier coefficient, as a function of the number of coefficients we require.

THEOREM 6. *Let $A \subseteq \{0,1\}^k$ describe a set of Fourier basis vectors, and let x be the contingency table that results from a data set D. Releasing the set $\phi_\alpha = \langle f^\alpha, x \rangle + Lap(2|A|/\epsilon 2^{k/2})$ for $\alpha \in A$ preserves ϵ-differential privacy of D.*

PROOF. Each tuple of the data set D contributes exactly $\pm 1/2^{k/2}$ to each output coordinate, and consequently the $L1$ sensitivity of the set of $|A|$ outputs is at most $2|A|/2^{k/2}$. By Theorem 2, the addition of Laplace noise with parameter $2|A|/\epsilon 2^{k/2}$ gives ϵ-differential privacy. □

Remark: Note that $n = |D|$ does not appear in Theorem 6. To get a sense of scale for the error, we could achieve a similar perturbation to each coordinate by randomly relocating $4|A|^2/\epsilon$ individuals in the data set, which can be much smaller than n.

3.2 Non-Negative Integrality

While there is certainly a real valued contingency table whose Fourier coefficients equal the perturbed values, e.g.: by returning the perturbed values to the original space, it is unlikely that there is a non-negative, integral contingency table with these coefficients. We now 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 contingency table with little additional error.

Letting $B \subset \{0,1\}^k$, suppose that we observed Fourier coefficients ϕ_β for $\beta \in B$. The following linear program minimizes, over all contingency tables w, the largest error b error between its Fourier coefficients $\langle f^\beta, w \rangle$ and the observed ϕ_β:

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

This optimization occurs in a $2^k + 1$ dimensional space, and any vertex of the feasible polytope must intersect $2^k + 1$ constraints. At most $|B|$ of these can relate to Fourier coefficients (since for each β, only one of the two constraints corresponding to β can be satisfied by any point). Thus at least $2^k - |B| + 1$ must be non-negativity constraints. This means that at any vertex of the polytope, all but at most $|B|$ weights are zero. Without loss of generality, the linear program will return a vertex solution[25], and rounding to the nearest integral point will result in at most an $L1$ change of $|B|$.

3.3 Algorithmic Recap

To bring things together, we now collect the various steps we have taken into a single algorithm.

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

1. Let B be the downward closure of A under $\preceq$.
2. For $\beta \in B$, compute $\phi_\beta = \langle f^\beta, x \rangle + Lap(2|B|/\epsilon 2^{k/2})$.
3. Solve for w_α in the following linear program, and round to the nearest integral weights, w'_α.

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

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

THEOREM 7. *Using the notation of **Marginals**(A), for all $\delta \in [0,1]$ with probability $1-\delta$, for all $\alpha \in A$,*

$$\|C^\alpha x - C^\alpha w'\|_1 \leq 2^{\|\alpha\|_1} 8|B| \log(|B|/\delta)/\epsilon + |B| \quad (12)$$

PROOF. Each Fourier coefficient has Laplace noise with parameter $2|B|/\epsilon 2^{k/2}$ added to it, and with probability $1-\delta$ none of these exceeds $4|B|\log(|B|/\delta)/\epsilon 2^{k/2}$. In solving the linear program, the error associated with each Fourier coefficient is at most this bound as well, as the original contingency table x is at least as close. Mapping the perturbation of a single Fourier coefficient back to the contingency table domain increases the L1 norm of the perturbation by at most $2^{k/2}$, up to at most $8|B|\log(|B|/\delta)/\epsilon$.

Consequently, for any marginal C^α, the error $C^\alpha x - C^\alpha w'$ is a result of noise in the $2^{\|\alpha\|_1}$ Fourier coefficients that contribute to the table, as well as the rounding that occurs. Multiplying the number of coefficients, $2^{\|\alpha\|_1}$ by the bound above, and adding the $|B|$ error due to rounding, gives the stated bound. □

Even tighter bounds can be placed on sub-marginals of a marginal C^α, by noting that the bounds hold for the marginals C^β for $\beta \preceq \alpha$ at no additional cost. No more Fourier coefficients are used, so $|B|$ is not increased, but $\|\beta\|_1 \leq \|\alpha\|_1$.

4. ALTERNATE APPROACHES

We now describe a few variants on the previous approaches that trade some of the accuracy of the previous approach for some conceptual or computational simplicity.

4.1 Alternate Linear Programs

The linear program we chose to use minimizes the largest error in any Fourier coefficient. There are other linear programs that one could write, for example minimizing the total error in Fourier coefficients, the largest error in reported marginals, the total error in the reported marginals, or several hybrids thereof.

This flexibility allows the data analyst with more specific accuracy concerns (eg: per cell accuracy) to address them. The perturbed Fourier coefficients can be released, and the specific linear program can be run to arrive at an integral, non-negative solution. Bounds similar to Theorem 7 can be proven, using the same methodology: the noise added perturbs the measurements by some distance in the norm of choice, and the linear program finds a non-negative solution at no greater distance from the perturbed measurements.

4.2 Non-Fourier Linear Programming

Our conversion to the Fourier domain is done because the Fourier coefficients exactly describe the information required by the marinals. By measuring exactly what we need, we add the least amount of noise possible using the techniques of [20].

Instead, we could apply the techniques of [20] directly to the true marginals, producing a set of noisy marginals that preserve privacy but not consistency. To these noisy marginals we apply a linear program to find a non-negative contingency table with nearest marginals. Imagining we have observed the noisy marginals c^β, the linear program is

$$\begin{array}{rcll} \text{minimize} & & b & \\ \text{subject to:} & & & \\ w_\alpha & \geq & 0 & \forall \alpha \in \{0,1\}^k \\ (c^\beta - C^\beta w)_\gamma & \leq & b & \forall \beta \in A, \gamma \preceq \beta \\ (c^\beta - C^\beta w)_\gamma & \geq & -b & \forall \beta \in A, \gamma \preceq \beta \end{array}$$

As before, we are likely to discover a fractional contingency table w. However, the number of cell constraints is at most $2\sum_{\beta \in A} 2^{\|\beta\|_1}$, and at most $\sum_{\beta \in A} 2^{\|\beta\|_1}$ of the w_α variables are non-zero. By the reasoning above, any rounding to integers introduces error at most this much in the contingency table.

THEOREM 8. *Using the above approach, with probability at least $1-\delta$, for all $\beta \in A$, then $\|C^\beta x - C^\beta w'\|_1$ is at most*

$$2^{\|\beta\|_1} 8|A| \log(\textstyle\sum_{\beta \in A} 2^{\|\beta\|_1}/\delta)/\epsilon + \sum_{\beta \in A} 2^{\|\beta\|_1} . \quad (13)$$

PROOF. The reasoning is the same as before: the difference in the marginals is no more than twice the difference caused by the additive noise, which is a Laplacian with parameter $2|A|/\epsilon$. We introduce the $\log(\sum_{\beta \in A} 2^{\|\beta\|_1}/\delta)$ term to give the high probability guarantee, and the additive term to account for rounding error. □

Remark: This theorem mirrors Theorem 7, using $|A|$ and $\sum_{\beta \in A} 2^{\|\beta\|_1}$ in place of $|B|$. Depending on the situation, these bounds can actually be tighter than in Theorem 7, though only when a single multi-attribute marginal is desired. The tighter bounds given by Theorem 7 through subtables also would not apply here.

4.3 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 that 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. We construct the output marginals directly from the Fourier coefficients, rather than reconstructing the contingency table, which could take time 2^k.

To ensure the existence of a non-negative contingency table with the observed Fourier coefficients turns out to be a simple task, 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 9. *Let $B \subset \{0,1\}^k$, and let x be a non-negative contingency table with Fourier coefficients ϕ_β for $\beta \in B$. If the Fourier coefficients are perturbed to ϕ'_β, then the contingency table*

$$x' \quad = \quad x + \sum_\beta (\phi'_\beta - \phi_\beta) f^\beta + \|\phi' - \phi\|_1 f^{\overline{0}} \quad (14)$$

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

PROOF. Each of the coordinates of f^β are $\pm 1/2^{k/2}$, and the most negative an entry could become due to the perturbation is $-\|\phi' - \phi\|_1/2^{k/2}$. By increasing the Fourier coefficient of the zero vector by $\|\phi' - \phi\|_1$, we increase every entry in the contingency table by this much, making them all non-negative. □

Our perturbation to the Fourier coefficients has $L1$ norm distributed exponentially with standard deviation $2^{3/2}|B|^2/\epsilon$. It is *critical* that we not disclose the actual $L1$ norm of the perturbation, but we can add a value for which the negativity probability is arbitrarily low:

COROLLARY 1. *By adding $t \times 4|B|^2/\epsilon 2^{k/2}$ to the first Fourier coefficient, the resulting contingency table is non-negative with probability at least $1 - \exp(-t)$.*

The addition of $4t|B|^2/\epsilon 2^{k/2}$ to the first Fourier coefficient corresponds to the introduction of $4t|B|^2/\epsilon$ individuals at random locations in the table; a relatively minor accuracy compromise.

5. CONCLUSIONS

We have shown a holistic solution to the problem of contingency table release, that outputs an accurate and consistent set of tables while guaranteeing in a very strong sense

that privacy of individuals is preserved. We also show how to construct a positive and integral synthetic database that corresponds to these tables—thus, e.g., one can output a synthetic database that preserves all low-order marginals up to a small error. Moreover, we can get a gracefully degrading version of the results: we can compute a synthetic database such that the error in the low-order marginals is small, and increases smoothly with the order of the marginal.

One of the main algorithmic questions left open from this work is that of efficiency. In particular, solving the linear program could be a bottleneck when the number of attributes is large, and it seems possible that one could devise more efficient algorithms for this step. We remark that the simplex algorithm is already space efficient in this setting, since each vertex of the polytope that simplex traverses has a sparse description. We leave open the question of devising faster combinatorial algorithms for this problem.

We have optimized for a specific measure of data quality, the distance between the reported and true marginals. It would also be useful to analyze the effect of our techniques or variants thereof on statistical properties of the marginals, such as means, variances, covariances, regressions. See the related work on controlled tabular adjustment [11].

6. ACKNOWLEDGEMENTS

We would like to thank Larry Cox and Steve Fienberg for comments on an earlier version of this paper and general guidance though the statistics literature.

7. REFERENCES

[1] *Special Issue on Statistical Disclosure Control*, volume 14(4) of *Journal of Official Statistics*. 1998.

[2] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *PODS*. ACM, 2001.

[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. In W. Chen, J. F. Naughton, and P. A. Bernstein, editors, *SIGMOD Conference*, pages 439–450. ACM, 2000.

[4] R. Agrawal, R. Srikant, and D. Thomas. Privacy preserving OLAP. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 251–262, New York, NY, USA, 2005. ACM Press.

[5] M. Bacharach. Matrix rounding problems. *Management Science*, 9:732–742, 1966.

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

[7] J. Castro. Quadratic interior-point methods in statistical disclosure control. *Computational Management Science*, 2:pages 107–121, 2005.

[8] J. Castro. Minimum-distance controlled perturbation methods for large-scale tabular data protection. *Euorpean Journal of Operantional Research*, 171:39–52, 2006.

[9] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Toward privacy in public databases. In J. Kilian, editor, *TCC*, volume 3378 of *Lecture Notes in Computer Science*, pages 363–385. Springer, 2005.

[10] S. Chawla, C. Dwork, F. McSherry, and K. Talwar. On privacy-preserving histograms. In *Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, Arlington, Virginia, 2005. AUAI Press.

[11] L. Cox, J. Kelly, and R. Patil. Balancing quality and confidentiality in multivariate tabular data. *Privacy in Statistical Databases*, 3080:87–98, 2004.

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

[13] R. A. Dandekar and L. Cox. Synthetic tabular data: An alternative to complementary cell suppression, 2002. manuscript, energy Information Administration, US Department of Energy.

[14] I. Dinur and K. Nissim. Revealing information while preserving privacy. In Milo [26], pages 202–210.

[15] A. Dobra and S. Fienberg. Bounding entries in multi-way contingency tables given a set of marginal totals, 2002. Proceedings of Conference on Foundation of Statistical Inference and its Applicaitons.

[16] J. Domingo-Ferrer and V. Torra. A critique of the sensitivity rules usually employed for statistical table protection. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):545–556, 2002.

[17] G. Duncan. Confidentiality and statistical disclosure limitation. In N. Smelser and P. Baltes, editors, *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier, 2001.

[18] 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.

[19] C. Dwork, D. Lee, and F. McSherry. Privacy preserving histogram case study, 2007. Manuscript.

[20] 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.

[21] C. Dwork, F. McSherry, and K. Talwar. The price of privacy and the limits of LP decoding. In *Proceedings of the 39th annual Symposium on the Theory of Computation.*, 2007.

[22] 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.

[23] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In Milo [26], pages 211–222.

[24] I. Fellegi. On the question of statistical confidentiality. *Journal of the American Statistical Association*, pages 7–18, 1972.

[25] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization (Algorithms and Combinatorics)*. Springer, December 1994.

[26] T. Milo, editor. *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*. ACM, 2003.

[27] J. Kelly, A. Assad, and B. Golden. The controlled rounding problem: Relaxations and complexity issues. *OR Spektrum*, 12:129–138, 1990.

[28] T. W. Körner. *Fourier Analysis*. Cambridge University Press, Cambridge, UK, 1988.

[29] J. A. D. Loera and S. Onn. All rational polytopes are transportation polytopes and all polytopal integer sets are contingency tables. *Proceedings of the 10th Ann. Math. Prog. Soc. Symp. Integ. Prog. Combin. Optim., LNCS*, 3064:338–351, 2004.

[30] J. A. D. Loera and S. Onn. The complexity of three-way statistical tables. *SIAM J. Comput.*, 33:819–836, 2004.

[31] J. A. D. Loera and S. Onn. Markov bases of three-way tables are arbitrarily complicated. *J. Symb. Comput.*, 41(2):173–181, 2006.

[32] 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.

[33] D. A. Robertson and R. Ethier. Cell suppression: Experience and theory. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 8–20. Springer, 2002.

[34] D. Rubin. Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9:461 – 469, 1993.

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

[36] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.

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