Differential Privacy for Protecting Multi-dimensional Contingency Table Data: Extensions and Applications

Xiaolin Yang[1], Stephen E. Fienberg[1,2], and Alessandro Rinaldo[1]

[1] Department of Statistics,
Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA
[2] Machine Learning Department, Cylab, and i-Lab
Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA

Abstract. The methodology of differential privacy has provided a strong definition of privacy which in some settings, using a mechanism of doubly-exponential noise addition, also allows for extraction of informative statistics from databases. In a recent paper, Barak et al. Barak et al. [1] extend this approach to the release of a specified set of margins from a multi-way contingency table. Privacy protection in such settings implicitly focuses on small cell counts that might allow for the identification of units that are unique in the database. We explore how well the mechanism works in the context of a series of examples, and the extent to which the proposed differential-privacy mechanism allows for sensible inferences from the released data. We conclude that the methodology, as it is currently formulated, is problematic in the context of the types of large sparse contingency tables encountered in statistical practice.

Keywords and phrases: Efron-Stein decomposition; Infeasible tables; Log-linear models; Privacy-protected marginals; Risk-Utility tradeoff.

1 Introduction

Contingency tables providing the cross-classification of a sample or a population according to a collection of categorical variables are among the most prevalent forms of statistical data, especially in the context of official statistics and sample surveys. When the data displayed are a random sample from a population, the most widely used statistical methods for analyzing the data are log-linear model methods. A key feature of log-linear models applied to multi-dimensional contingency tables is the fact that the minimal sufficient statistics are sets of possibly overlapping marginals, from which one can compute maximum likelihood estimates, e.g., see the books by Bishop et al. [2], Edwards [11], Lauritzen [17], and Whittaker [21].

Fienberg and Slavkovic [16] reviewed the statistical literature on privacy protection of results from contingency tables focusing on the exact release of minimal sufficient marginals under a well-fitting log-linear model and they discuss this method in the context of the Risk-Utility (RU) trade-off initially proposed by Duncan et al. [5], who defined risk in terms of protection of small counts in the table. Dobra et al. [4] provided further insight into the RU-trade-off problem for large sparse tables using recent results from algebraic statistics. Winkler [22] proposed a method to reduce re-identification risk while preserving analytic properties by placing upper and lower bounds on margins, the key aggregates needed for log-linear modeling, and also on large sets of small cells and sampling zeros. He generated synthetic data from the model in appropriate fashion.

The methodology of differential privacy [6,7] has provided a clear and very strong definition of privacy which, in many settings, uses a mechanism of doubly-exponential noise addition. Differential privacy also allows for extraction of informative statistics from databases. A recent paper by Barak et al. [1] extended the differential privacy approach to the release of a pre-specified set of margins from a 2^k contingency table, for $k \geq 3$, using a Fourier basis expansion. Adding non-integer noise in such contexts poses a variety of problems: violation of non-negativity cell probabilities, incompatible margins, and infeasible tables. The proposed methodology in [1] purports to handle all of these problems. In Fienberg et al. [15] we provided an initial report on how well the mechanism works in the context of a series of three examples, and the extent to which the proposed differential-privacy mechanism allows for sensible inferences from the released data. In the present paper, we extend the Barak et al. method to non-binary multi-way tables using the Efron-Stein [13] decomposition and expand the empirical results from our earlier work to demonstrate the problems we encountered earlier.

In the following sections, we briefly describe differential privacy, our notation for contingency tables, the Barak et al. [1] approach to multi-way binary tables and our extension to it using the Efron-Stein decomposition. Then we evaluate the usefulness of these versions of the differential privacy methodology using a variant of the RU-tradeoff. We conclude that, for the type of large sparse contingency tables tables often encountered in statistical practice, that the current variations on differential privacy either protect too little in real terms or obscure the data by adding too much noise and thus impair realistic statistical data analysis.

2 Differential Privacy

Let $\mathcal{D}$ denote the set of databases. A privacy protecting mechanism is a randomized function $K\colon \mathcal{D} \to \mathcal{D}$. The output of K is a random database called the sanitized database.

Definition 1. *The privacy protecting mechanism K satisfies ϵ-differential privacy if, for all databases D_1 and D_2 in $\mathcal{D}$ differing on at most one record, and all measurable subsets S of the range of K,*

$$Pr[K(D_1) \in S] \le \exp(\epsilon) Pr[K(D_2) \in S].$$

The smaller the value of ϵ, the greater the privacy provided by the mechanism, in the sense that the probability distribution of the sanitized database is rather insensitive to a one-record change in the input database. Wasserman and Zhou [20, Theorem 2.4] provide a related statistical interpretation of differential privacy based on the theoretical perspective of hypothesis testing.

3 Notation for Contingency Tables

A k-way contingency table arises from the cross classification of n units according to k categorical variables $(X_1, \dots, X_k)$, with the j-th variable taking on $k_j \ge 2$ possible values, $j = 1, \dots, k$. For any positive integer k, let $[k] = \{1, \dots, k\}$ and set $\Omega = \prod_{j=1}^{k} [k_j]$. Every coordinate point $x \in \Omega$ is called a cell, and it is convenient to think of a contingency table as a vector $f \in \mathbb{R}^{\Omega}$ whose x coordinate, denoted with $f(x)$, corresponds to the number of times the x combination of the k variables occurred in the sample.

For a given subset $\alpha \subset \{1, \dots, k\}$, let $\Omega_\alpha = \prod_{j\in\alpha} [k_j]$. We will write $x_\alpha = \{x_j, j \in \alpha\} \in \Omega_\alpha$ for the α-coordinate projection of x. The α-marginal table of the contingency table f is the $|\alpha|$-dimensional array $f_\alpha = \{f_\alpha(x_\alpha), x_\alpha \in \Omega_\alpha\}$, whose x_α entry is obtained by summing over the cells $y \in \Omega$ of the original table f whose α-coordinate projection is x_α:

$$f_\alpha(x_\alpha) = \sum_{y \in \mathbb{R}^{\Omega} : y_\alpha = x_\alpha} f(y). \tag{1}$$

With a slight abuse of notation, we refer to both α and f_α as margins. Finally, for any margin α, we will write compactly $f_\alpha = C^\alpha f$, were C^α is the $|\Omega_\alpha| \times |\Omega|$ matrix realizing the sums in equation (1).

For vectors $f, g \in \mathbb{R}^{\Omega}$, we will denote the L_1 norm as $\|f\|_1 = \sum_x |f(x)|$ and the standard inner product as $\langle f, g\rangle = \sum_x f(x)g(x)$.

Example 1. A 2^k contingency table arises from the cross classification of n individuals according to k binary categorical variables, where each cell of the table corresponds to the number of times a given combination of the k variables occurred in the sample. It is convenient for us to think of a table f as a vector in R^{2^k}.

Example 2. A more general form of contingency table involves multiple categories for one or more attributes. Instead of having 0 or 1 as the attributes' values, they may have more than two possible values. For example, in a $3 \times 3 \times 2$ table the three attributes can take $\{0, 1, 2\}$, $\{0, 1, 2\}$ and $\{0, 1\}$ and we represent the 3^3 table as a vector in R^{3^3}. We obtain the margins using similar methods described above.

4 The Risk-Utility Trade-off

Let $\mathcal{A} \subset 2^{\{0,1\}^k}$ be a collection of margins, such that $\cup_{\alpha \in \mathcal{A}} = \{1, \ldots, k\}$ and $\alpha_1 \not\subset \alpha_2$ for any $\alpha_1, \alpha_2 \in \mathcal{A}$.

From the theory of log-linear models [2,17], we know that each such collection $\mathcal{A} \subset 2^{\{0,1\}^k}$ encode a statistical model for the the probabilistic dependence among the k attributes, each of which as a categorical random variable. Specifically, each $\mathcal{A}$ specify a collection of positive probability distributions over $\{0, 1\}^k$ obeying a set of rules known as Markov properties. Each probability distribution is a point p in the simplex in R^{Ω} such that $p(x)$ denotes the probability of observing the cell x. The corresponding marginal tables $\{f_\alpha, \alpha \in \mathcal{A}\}$ are minimal sufficient statistics for the model determined by $\mathcal{A}$. This means that, from an inferential standpoint, the $\mathcal{A}$-margins of f contains as much statistical information as f itself. Furthermore, they determine the maximum likelihood estimator (MLE) $\hat{p}$, which is the unique probability distribution in the model encoded by $\mathcal{A}$ that makes f the "most likely" sample that we could have observed. The MLE possesses many optimal properties and, in particular, and we can use it to assess the fit of the model $\mathcal{A}$ using the likelihood ratio test statistic

$$G^2 = 2 \sum_{x \in \Omega} f(x) \log \left(\frac{f(x)}{n\hat{p}(x)} \right). \tag{2}$$

From a privacy protection perspective the table x contains potentially sensitive information whose public release would entail a violation of privacy. Because the release of some information from such databases is a public utility, a database curator overseeing the table seeks to implement a mechanism of partial data release that are safe from the privacy standpoint. While the $\mathcal{A}$-margins contain only aggregate (partial) information about x and thus appear to be a natural candidates for a data release [16,4], marginal releases may not in general correspond to a private-preserving mechanism, especially when the data base is sparse and contains many small counts [1]. By titrating the privacy mechanism we might also be able to apply some form of perturbation to the data and yet also produce statistical useful results.

5 A Differential Privacy Mechanism for Contingency Tables

5.1 Differential Privacy Mechanism for Binary Tables

We first review the theory and the algorithms developed in [1] for differential privacy for binary tables, i.e. tables for which $k_j = 2$, for all $j \in [k]$. In this special and simple setting, the set $\Omega = \{0,1\}^k$ consists of the vertices of the k-dimensional unit hypercube and [1] used the Fourier basis $\mathbb{R}^\Omega = \mathbb{R}^{2^k}$. To this end, we represent a set $\alpha \subset \{1,\ldots,k\}$ as a vector in $\{0,1\}^k$ whose positive coordinates are precisely α. In particular, when we speak of α-margin, we are treating α as a point in $\{0,1\}^k$. Thus, in this binary setting, both the cell coordinates x and the margins α are described by points in $\{0,1\}^k$.

Let $\{\chi^S, S \in \{0,1\}^k\}$ be the Fourier basis for R^{2^k}, whose S element is the vector $\chi^S = \{\chi^S(x), x \in \Omega\}$, where

$$\chi^S(x) = \frac{1}{2^{k/2}}(-1)^{\langle S,x\rangle}.$$

Barak et al. [1] show that, for every marginal α, the orthonormal Fourier basis yields a basis for $R^{2^{|\alpha|}}$, in the sense that

$$C^\alpha f = \sum_{S \preceq \alpha} \langle f, \chi^S\rangle C^\alpha \chi^S,$$

where for $S, \alpha \in \{0,1\}^k$, $S \preceq \alpha$ signifies that every non-zero coordinate of S is also a non-zero coordinate of α. The Fourier basis representation is exactly the traditional u-parametrization of log-linear models e.g., as described in [2]; equivalently, it gives the direct sum decomposition of R^{2^k} in terms of the subspaces of interaction, e.g., see [17, Appendix B]. Based on the Fourier basis representation of the marginal tables, Barak et al. [1] proposed a differentially private mechanism for releasing a prescribed set of margins $\mathcal{A}$ from a binary table f, which we reproduce in Table 1. They showed that the algorithm possesses the following properties.

Theorem 1. *Let $\mathcal{A}$ denote a set of margins and $\mathcal{B}$ its downward closure with respect to $\preceq$. Then, the privacy mechanism of Table 1 satisfies differential privacy and, for each $\delta \in (0,1)$, with probability at least $(1-\delta)$,*

$$\|C^\alpha f - C^\alpha w'\|_1 \leq 2^{|\alpha|} 8 \frac{|\mathcal{B}|}{\epsilon} \log\left(\frac{|\mathcal{B}|}{\delta}\right) + |\mathcal{B}|,$$

uniformly over all $\alpha \in \mathcal{A}$.

Barak et al. [1] argue that the above mechanism is simultaneously (i) private (since it satisfies the strong requirement of differential privacy), (ii) accurate (as it provides probabilistic guarantees on the maximal L_1 distance between the observed and release margins), and (iii) consistent, (as it release margins that can be realized by an integer-valued table (namely w')).

Table 1. The differentially private mechanism for binary contingency tables.

1. Inputs: a differential privacy parameter ϵ, a binary k-dimensional table f and a set of margins $\mathcal{A}$.
2. Let $\mathcal{B}$ the downward closure of $\mathcal{A}$ with respect to $\preceq$.
3. Generate $\{X_S, S \in B\}$ as independent random variables with common distribution $\mathrm{Lap}\left(\frac{2|\mathcal{B}|}{\epsilon 2^{k/2}}\right)$.
4. For each $S \in \mathcal{B}$, compute the perturbed S-marginal $\phi_S = \langle f, \chi^S \rangle + X_S$
5. Solve for $w = \{w(x), x \in \{0,1\}^k\}$ the linear program

$$\begin{gathered} \min b \\ \text{subject to:} \\ w(x) \geq 0, \quad \forall x \\ \phi_S - \textstyle\sum_x w(x)\chi^S(x) \leq b, \quad \forall S \in \mathcal{B} \\ \phi_S - \textstyle\sum_x w(x)\chi^S(x) \geq -b, \quad \forall S \in \mathcal{B}. \end{gathered}$$

6. Round w to w', where $w'(x)$ is the nearest integer to $w(x)$.
7. Return the $\mathcal{A}$-margins of w'.

Remarks

1. The result is independent of the sample size, and the accuracy guarantees depend only on the model complexity $|\mathcal{B}|$ and the differential privacy parameter ϵ.
2. The linear program described above may return a solution for which $b > 0$ (in fact, we have often observed this phenomenon in our computations). This implies that there does not exist any real-valued non-negative table with $\mathcal{B}$-margins given by $\{\phi_S, S \in \mathcal{B}\}$.
3. The linear program has typically many (in fact infinite) solutions.
4. The proof of Theorem 1 implicitly assumes that $b = 0$, which, as we mentioned, does not hold in general.

5.2 Differential Privacy Mechanism for Non-binary Contingency Tables

The method proposed in [1] to achieve differential privacy by adding Laplacian noise to the Fourier coefficients only works with binary tables. Here we outline a similar methodology for non-binary tables using a different orthogonal basis, known as the Efron-Stein decomposition (see, for instance, [13]).

We associate with each of the k categorical variables its own finite probability space: $(\Omega_1, \mathcal{F}_1, \mu_1), \ldots, (\Omega_k, \mathcal{F}_k, \mu_k)$ with $\Omega_j = [k_j]$ and μ_j a measure on $(\Omega, \mathcal{F}_j)$. With denote with μ the corresponding product measure on $\Omega = \prod_j \Omega_j$. The Efreon-=Stein decomposition of any function on Ω is given below.

Definition 2. *Let f a real-valued function on Ω. The Efron-Stein decomposition of f is given by*

$$f(x) = \sum_{S \subseteq [k]} f^S(x_S), \quad x \in \Omega, \tag{3}$$

where the functions $f^S \colon \Omega \to \mathbb{R}$ satisfy:

1. *f^S only depends on S in the sense that $f^S(x) = f^S(x_S)$;*
2. *for $S \nsubseteq S'$, $E[f^S(X)|X_{S'} = x_{S'}] = 0$, where the expectation is with respect to the product measure μ.*

Explicitly, each component function f^S can be written as

$$f^S(x) = \sum_{S' \subseteq S} (-1)^{|S \setminus S'|} E[f(X)|X_{S'} = x_{S'}]. \tag{4}$$

In particular, choosing μ to be the uniform probability measure on Ω and identifying, as we did above, the function f with the vector $(f(x), x \in \Omega) \in \mathbb{R}^{\Omega}$, the conditional expectations in (4) can be written as

$$E[f(X)|X_{S'} = x_{S'})] = \left\langle f, v_{S',x_{S'}} \right\rangle,$$

where $v_{S',x_{S'}} \in \mathbb{R}^{\Omega}$ is the conditional probability of X given $X_{S'} = x_{S'}$. Notice that, since the conditional distributions depend both on the coordinates in S' and the value of x, $v_{S',x_{S'}}$ is indexed by both S' and $x_{S'}$.

Therefore, by linearity, (4) can be written explicitly as

$$f^S(x_S) = \sum_{S' \subseteq S} (-1)^{|S \setminus S'|} \left\langle f, v_{S',x_{S'}} \right\rangle = \left\langle f, f^{S,x_S} \right\rangle$$
$$\text{where } f^{S,x_S} = \sum_{S' \subseteq S} (-1)^{|S \setminus S'|} v_{S',x_{S'}} \tag{5}$$

It is not hard to see that, for a fixed S', the vectors $\{v_{S',x_{S'}}, x \in \Omega\}$ are orthogonal each other. Also f^{S,x_S}'s for different S form an orthogonal basis for $\mathbb{R}^{\Omega}$, furthermore, their entries are $\pm \frac{1}{\prod_{j \in [k] \setminus S} n_j}$, where n_j is the number of values each variable X_j can take.

Example 3. In our first example, we consider a $3 \times 2 \times 2$ table, $S' = \{1\}$, which means we only condition on the first variable X_1. Since X_1 takes three values 0, 1 and 2, we obtain 3 vectors for $v_{S',x_{S'}}$:

$$v_{\{1\},0} = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, 0, 0, 0, 0, 0, 0, 0\right]^T,$$
$$v_{\{1\},1} = \left[0, 0, 0, 0, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, 0, 0, 0\right]^T$$
$$v_{\{1\},2} = \left[0, 0, 0, 0, 0, 0, 0, 0, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right]^T.$$

In our second example, we assume a $2 \times 2 \times 2$ table with $S' = \{2, 3\}$. In this case, there are four vectors for $v_{S',x_{S'}}$:

$$v_{\{2,3\},00} = \left[\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0, 0, 0\right]^T$$
$$v_{\{2,3\},01} = \left[0, \frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0, 0\right]^T$$

$$v_{\{2,3\},10} = \left[0, 0, \frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0\right]^T$$
$$v_{\{2,3\},11} = \left[0, 0, 0, \frac{1}{2}, 0, 0, 0, \frac{1}{2}\right]^T$$

For this example, we also verify the second condition in definition (2). First we compute f^{S,x_S} for $S = \{3\}$. The downward closure of S is $\Big\{\emptyset, \{3\}\Big\}$, so $f^{\{3\},x_3}$ depends on the vectors

$$v_{\emptyset} = \left[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right]^T$$
$$v_{\{3\},0} = \left[\frac{1}{4}, 0, \frac{1}{4}, 0, \frac{1}{4}, 0, \frac{1}{4}, 0\right]^T$$
$$v_{\{3\},1} = \left[0, \frac{1}{4}, 0, \frac{1}{4}, 0, \frac{1}{4}, 0, \frac{1}{4}\right]^T$$

Then, the vectors f^{S,x_S} in (5) for $x_{\{3\}}$ equal to 0 and 1 are

$$f^{\{3\},0} = \left[\frac{1}{8}, -\frac{1}{8}, \frac{1}{8}, -\frac{1}{8}, \frac{1}{8}, -\frac{1}{8}, \frac{1}{8}, -\frac{1}{8}\right]^T$$

and

$$f^{\{3\},1} = \left[-\frac{1}{8}, \frac{1}{8}, -\frac{1}{8}, \frac{1}{8}, -\frac{1}{8}, \frac{1}{8}, -\frac{1}{8}, \frac{1}{8}\right]^T,$$

respectively. Choosing $S = \{1\}$, it is easy to see that $E[f^S(X)|X_{S'} = x_{S'}] = \frac{1}{4}(f^S(000) + f^S(001) + f^S(010) + f^S(011)) = 0$.

Example 4. For the case of binary tables, the Efron-Stein decomposition coincides with the Fourier representation used in [1], in the sense that, for any $S \subseteq [k]$,

$$f^S(x_S) = \left\langle f, \chi^S \right\rangle \chi^S(x),$$

where χ^S is the fourier basis element described as described in the previous section, with S subsets of $[k]$ and x only related to subsets of $[k]$. Indeed, for a $2 \times 2 \times 2$ table and $S = \{3\}$, by the previous calculations using Efron-Stein decomposition, we obtain

$$f^{\{3\}}(0) = \frac{1}{8}f_1 - \frac{1}{8}f_2 + \frac{1}{8}f_3 - \frac{1}{8}f_4 + \frac{1}{8}f_5 - \frac{1}{8}f_6 + \frac{1}{8}f_7 - \frac{1}{8}f_8$$

and

$$f^{\{3\}}(1) = -\frac{1}{8}f_1 + \frac{1}{8}f_2 - \frac{1}{8}f_3 + \frac{1}{8}f_4 - \frac{1}{8}f_5 + \frac{1}{8}f_6 - \frac{1}{8}f_7 + \frac{1}{8}f_8,$$

where we ordered the entries of f lexicographically. On the other hand, using Fourier basis,

$$\langle f, \chi^S \rangle = \frac{1}{2^{\frac{1}{2}}}f_1 - \frac{1}{2^{\frac{1}{2}}}f_2 + \frac{1}{2^{\frac{1}{2}}}f_3 - \frac{1}{2^{\frac{1}{2}}}f_4 + \frac{1}{2^{\frac{1}{2}}}f_5 - \frac{1}{2^{\frac{1}{2}}}f_6 + \frac{1}{2^{\frac{1}{2}}}f_7 - \frac{1}{2^{\frac{1}{2}}}f_8.$$

Thus,

$$\langle f, \chi^S \rangle \chi^S(**0) = \frac{1}{8}f_1 - \frac{1}{8}f_2 + \frac{1}{8}f_3 - \frac{1}{8}f_4 + \frac{1}{8}f_5 - \frac{1}{8}f_6 + \frac{1}{8}f_7 - \frac{1}{8}f_8 = f_{\{3\}}(0)$$

and

$$\langle f, \chi^S \rangle \chi^S(**1) = -\frac{1}{8}f_1 + \frac{1}{8}f_2 - \frac{1}{8}f_3 + \frac{1}{8}f_4 - \frac{1}{8}f_5 + \frac{1}{8}f_6 - \frac{1}{8}f_7 + \frac{1}{8}f_8 = f_{\{3\}}(1),$$

where in the above expressions $(**1)$ denotes any binary string of length three terminating in a "1".

Suppose we want to release a set of margins $\mathcal{A} \subseteq [k]$ and let $\mathcal{B}$ be the downward closures of $\mathcal{A}$. Then, using equation (3), we can write

$$f(x) = \sum_{S \in \mathcal{B}} f^S(x_S) + \sum_{S \notin \mathcal{B}} f^S(x_S).$$

By Theorem 2 in [1], the following perturbation f' of the function f will preserve ϵ-differentia privacy:

$$f'(x) = \sum_{S \in \mathcal{B}} (f^S(x_S) + \mathrm{Lap}(\Delta f/\epsilon)) + \sum_{S \notin \mathcal{B}} f^S(x_S). \tag{6}$$

The term Δf is the L_1 sensitivity of f (see definition 2 in [1]). Notice that, just like with binary tables, we only add noise to the downward closure of released margins.

The exact value of the noise level needed to preserve ϵ-differential privacy is given in the following theorem.

Theorem 2. *Suppose we wish to release the margin $\mathcal{A}$ of a contingency table and $\mathcal{B}$ is the downward closure of $\mathcal{A}$. When using Efron-Stein decomposition, the addition of Laplace noise with variance $\sum_{S \in \mathcal{B}} \frac{2}{\epsilon \prod_{j \in [k] \setminus S} n_j}$ to each term $f^S(x_S)$, where $S \in \mathcal{B}$, preserves ϵ-differential privacy.*

Proof. The proof follows similar procedures as using Fourier basis for binary tables. Suppose two database D_1 and D_2 differ only in one data point, for each $S \in \mathcal{B}$ and x_S, each data point contributes at most $\pm \frac{1}{\prod_{j\in[k]} k_j}$ to the output $f^S(x_S)$. The L_1 sensitivity of $f^S(x_S)$ is $\frac{2}{\prod_{j\in[k]} k_j}$. The total number of terms of the form $f^S(x_S)$ is $\sum_{S\in\mathcal{B}}(\prod_{j\in S} k_j)$. So the L_1 sensitivity of all outputs is bounded by $\sum_{S\in\mathcal{B}} \frac{2}{\prod_{j\in[k]\setminus S} k_j}$. Then adding Laplace noise $\mathrm{Lap}\left(\sum_{S\in\mathcal{B}} \frac{2}{\epsilon \prod_{j\in[k]\setminus S} k_j}\right)$ preserves ϵ-differential privacy.

Theorem 3 below generalizes Theorem 1 (Theorem 7 of [1]) by providing probabilistic bound on the change in the L_1 norm of the margins due to the addition of Laplace noise.

Theorem 3. *Let $\mathcal{A}$ denote a set of margins and $\mathcal{B}$ its downward closure with respect to $\preceq$. For all $\delta \in [0, 1]$ with probability $1 - \delta$,*

$$||C^\alpha f - C^\alpha w^{'}|| \le \frac{2}{\epsilon}\left(\prod_{i\in\alpha} k_i\right) \sum_{S\in\mathcal{B}} \frac{1}{\prod_{j\in[k]\setminus S} k_j} \log\left(\frac{N}{\delta}\right) + N$$

where $N = \sum_{S\in\mathcal{B}} \prod_{j\in S} k_j$, uniformly over all $\alpha \in \mathcal{A}$.

Proof. We add Laplacian noise with variance $\sigma = \sum_{S\in\mathcal{B}} \frac{2}{\epsilon \prod_{j\in[k]\setminus S} k_j}$ to each term $f^S(x_S)$. With probability $1 - \delta$, the maximum of these $f^S(x_S)$ never exceeding λ is equivalent to the fact that each $f^S(x_S)$ will not exceed λ with probability $\frac{\delta}{N}$. Using the property of Laplacian distribution, we get, for $X \sim \mathrm{Lapl}(\lambda)$,

$$P(|X| > \lambda) = \frac{\delta}{N} \Leftrightarrow P(X > \lambda) = \frac{\delta}{2N} = \frac{1}{2} \exp^{-\lambda/\sigma}.$$

Then $\lambda = \sum_{S\in\mathcal{B}} \frac{2}{\epsilon \prod_{j\in[k]\setminus S} k_j} \log(\frac{N}{\delta})$. For $\alpha \in \mathcal{A}$ the number of $f^S(x_S)$ is $\prod_{i\in\alpha} k_i$. So the total error introduced is

$$\frac{2}{\epsilon}\left(\prod_{i\in\alpha} k_i\right) \sum_{S\in B} \frac{1}{\prod_{j\in[k]\setminus S} k_j} \log\left(\frac{N}{\delta}\right)$$

Then adding N to the bound due to the rounding error we got the total error bound.

From Equation (6), we get the perturbed $f^S(x_S)$. We hope to solve $f(x)$ from the perturbed $f^S(x_S)$. According to Equation (4), we know

Table 2. The differentially private mechanism for non-binary contingency tables.

1. Inputs: a differential privacy parameter ϵ, a binary k-dimensional table f and a set of margins $\mathcal{A}$.
2. Let $\mathcal{B}$ be the downward closure of $\mathcal{A}$ with respect to $\preceq$.
3. Generate $\{X_S, S \in B\}$ as independent random variables with common distribution $\mathrm{Lap}\left(\sum_{S \in B} \frac{2}{\epsilon \prod_{j \in [k] \setminus S} k_j}\right)$.
4. For each $S \in B$, compute the perturbed S-marginal $f^S(x_S)' = \langle f(x), f^{S,x_S} \rangle + X_S$
5. Solve for $w = \{w(x), x \in \Omega\}$ the linear program

$$\begin{gathered} \min b \\ \text{subject to:} \\ w(x) \geq 0, \quad \forall x \\ f^S(x_S)' - \sum_x w(x) f^{S,x_S} \leq b, \quad \forall S \in \mathcal{B} \text{ and } \forall x_S \\ f^S(x_S)' - \sum_x w(x) f^{S,x_S} \geq -b, \quad \forall S \in \mathcal{B} \text{ and } \forall x_S. \end{gathered}$$

6. Round w to w', where $w'(x)$ is the nearest integer to $w(x)$.
7. Return the $\mathcal{A}$-margins of w'.

how to compute $f^S(x_S)$ given the conditional expectation of $f(x)$ and the downward closure of S. Then solving $f(x)$ is equivalent to solving a linear programming problem.

Following the "holistic" algorithm in Table 1, in Table 2 we provide an algorithm for computing perturbed margins using the Efron-Stein decomposition.

6 Empirical Evaluation of the Differential Privacy Mechanisms

We now analyze the statistical properties of the privacy preserving mechanism of [1] on three real-life datasets. We also analyze a non-binary table using the method we propose in section 5.2. We study empirically whether the algorithms in Tables 1 and 2 for producing differentially private results, is also statistically sound, in the sense that the results of statistical analyses of the sanitized margins do not deviate significantly from the results obtained using the original database. In particular, we are interested in the rather basic question of whether a model that fits the original database well will also fit the perturbed data.

Table 3. Cell counts 2^6 table involving genetic linkage in barley powder mildew fungus. Source: Edwards [10].

			1				2				D
			1		2		1		2		E
			1	2	1	2	1	2	1	2	F
1	1	1	0	0	0	0	3	0	1	0	
		2	0	1	0	0	0	1	0	0	
	2	1	1	0	1	0	7	1	4	0	
		2	0	0	0	2	1	3	0	11	
2	1	1	16	1	4	0	1	0	0	0	
		2	1	4	1	4	0	0	0	1	
	2	1	0	0	0	0	0	0	0	0	
		2	0	0	0	0	0	0	0	0	
A	B	C									

1. Table 3 is a sparse 6-dimensional binary contingency table obtained from the cross-classification of six dichotomous categorical variables, labeled with the letters A-F, recording the parental alleles corresponding to six loci along a chromosome strand of a barley powder mildew fungus, for a total of 70 offspring. The data were originally described by [3] and further analyzed by [10]. Based on the model selection analysis described in [11], the model $[AD][AB][BE][CE][CF]$ fits the data well and has also a biological foundation. Out of 64 cells, only 22 are non-zero and most the entries are small counts.
2. The data in Table 4 were collected in a prospective epidemiological study of 1841 workers in a Czechoslovakian car factory, as part of an investigation of potential risk factors for coronary thrombosis. See [12]. In the left-hand panel of Table 1, A indicates whether or not the worker "smokes", B corresponds to "strenuous mental work", C corresponds to "strenuous physical work", D corresponds to "systolic blood pressure", E corresponds to "ratio of and lipoproteins" and F represents "family anamnesis of coronary heart disease". The model $[BF][ABCE][ADE]$ fits the data well. The cell counts are fairly large, with 14 cells having values of 5 or less.

Table 4. Cell counts for Czech autoworker 2^6 table. Source: Edwards and Havranek [12].

F	E	D									
			C	1				2			
			B	1		2		1		2	
			A	1	2	1	2	1	2	1	2
1	1	1		44	40	112	67	129	145	12	23
		2		35	12	80	33	109	67	7	9
	2	1		23	32	70	66	50	80	7	13
		2		24	25	73	57	51	63	7	16
2	1	1		5	7	21	9	9	17	1	4
		2		4	3	11	8	14	17	5	2
	2	1		7	3	14	14	9	16	2	3
		2		4	0	13	11	5	14	4	4

3. The data in Table 5 involve 8 binary variables (Yes/No) relating women's economic activity and husband's unemployment from a survey of households in Rochdale [21, see page 279]. The 8 variables are: wife economically active (A); wife older than 38 (B); husband unemployed (D); child of age less than 4 (D); wife's education, high-school or higher (E); husband's education, high-school or higher (F); Asian origin (G); other household member working (H). The sample size is 665, and 165 of the 256 cells contain zero counts and 58 cells have positive counts of 4 or less.
4. The data in Table 6 correspond to 3 categorical variables with 4 zones of origin (Home), 4 zones of destination (Work) and 16 income categories, respectively. The sparseness of the data is due to the fact that some neighborhoods do not contain any low income workers since they could not afford to live there. Similarly some destinations do not have highly paid positions. The sample size is 2291 and 183 out of 256 cells contain zero counts.

Table 7 provides a quick summary of the dimensions and sample sizes of the four datasets, along with the selections of margins corresponding to log-linear models fitting the data adequately. In addition, we report the LR statistic and corresponding degrees of freedom. All the datasets have small dimensions and, except for the dataset in Table 3, relatively large sample sizes.

Table 5. Rochdale table. Source: Whittaker [21].

	Y								N								H
	Y				N				Y				N				G
	Y		N		Y		N		Y		N		Y		N		F
	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	E
Y Y Y Y	5	0	2	1	5	1	0	0	4	1	0	0	6	0	2	0	
N	8	0	11	0	13	0	1	0	3	0	1	0	26	0	1	0	
N Y	5	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	
N	4	0	8	2	6	0	1	0	1	0	1	0	0	0	1	0	
N Y Y	17	10	1	1	16	7	0	0	0	2	0	0	10	6	0	0	
N	1	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	
N Y	4	7	3	1	1	1	2	0	1	0	0	0	1	0	0	0	
N	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	
N Y Y Y	18	3	2	0	23	4	0	0	22	2	0	0	57	3	0	0	
N	5	1	0	0	11	0	1	0	11	0	0	0	29	2	1	1	
N Y	3	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0	
N	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
N Y Y	41	25	0	1	37	26	0	0	15	10	0	0	43	22	0	0	
N	0	0	0	0	2	0	0	0	0	0	0	0	3	0	0	0	
N Y	2	4	0	0	2	1	0	0	0	1	0	0	2	1	0	0	
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
A B C D																	

Table 8 reports, for each of the four datasets under study, the variances of the Laplace additive noise corresponding to values of ϵ of 0.1, 1 and 2, and also the bounds on the L_1 distances between observed and perturbed margins as predicted by Theorems 1 and 3, as functions of the probability parameter $\delta \in (0,1)$. It is immediately clear that the variance of the additive Laplace noise decreases very rapidly as ϵ gets larger, suggesting a significant sensitivity of the privacy mechanism to the differential privacy guarantee as measured by the parameter ϵ. Another striking feature that emerges from Table 8 is the magnitude of the constants in the upper bound on the L_1 distances between observed and perturbed margins. As these constants are decreasing in ϵ, when ϵ is even moderately small, the corresponding values end up being larger than the sample size, a clearly undesirable feature.

Table 6. Synthetic journey to work by income table developed using an ad hoc privacy approach for data extracted from a 2000 census database. Source: Fienberg and Love [14].

Home Zone	Work Zone	Income Category 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	C
a	a	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
a	b	46	34	0	23	0	0	0	0	0	0	0	0	0	0	0	0	
a	c	243	200	0	0	45	0	0	0	70	0	0	80	0	0	0	0	
a	d	0	0	0	0	0	0	0	45	60	0	0	0	0	0	0	0	
b	a	4	9	15	14	18	17	0	0	17	18	22	44	33	0	16	16	
b	b	0	0	0	0	0	0	0	0	0	0	0	0	0	78	0	0	
b	c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
b	d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
c	a	14	24	36	34	14	16	17	18	0	18	12	0	44	34	33	33	
c	b	0	0	14	0	16	18	18	34	12	16	44	22	16	18	12	14	
c	c	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	
c	d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
d	a	12	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
d	b	14	12	67	9	22	66	14	14	34	37	38	12	24	22	16	18	
d	c	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	
d	d	0	0	0	0	0	0	18	0	0	22	0	0	0	0	0	0	
A	B																	

In order to investigate the effect and statistical implications of the privacy protecting mechanisms described in Tables 1 and 2, we conducted a series of simulation experiments which we summarize through Figures 1 - 5.

Specifically, we considered a grid of values for the differential privacy parameter ϵ ranging from 0.005 (strong privacy guarantee) to 2 (weak privacy guarantee) with grid size 0.01 and, for each such value, we applied the privacy algorithms of Tables 1 and 2 50 times (because these are randomized algorithms, their outputs are random). For the binary tables we used the algorithm of [1], summarized in Table 1, while for the non-binary table 6 we used the algorithm we described in Section 5.2. For clarity, we have produced two separate plots for each experiment, one for the values ϵ up to 1 and the second one from values between 1 and 2.

We first consider the effect of the privacy protecting mechanism on the sample size of the perturbed table. Figure 1 shows the sample size of

Table 7. Table dimension, sample size, chosen model, likelihood ratio statistic (2) and associated number of degrees of freedom for the four tables analyzed.

Table	Dimension	Sample Size	Model	LR	d.f.
Edwards	$k = 6$	$n = 70$	[AD][AB][BE][CE][CF]	22.96	52
Czech	$k = 6$	$n = 1841$	[BF][ADE][ABCE]	48.18	42
Rochdale	$k = 8$	$n = 665$	[ACE][ACG][ADG][BDH] [BF][BE][CEF][CFG]	238.18	225
Journey to work	$k = 3$	$n = 2291$	[AB][AC][BC]	365.82	134

Table 8. Variance of the additive noise and L_1 bounds on the margins for the four datasets considered and three different values of ϵ.

	ϵ		
	0.01	1	2
Edwards	Lap(300) $38400 \log(12/\delta) + 12$	Lap(3) $384 \log(12/\delta) + 12$	Lap(1.5) $192 \log(12/\delta) + 12$
Czech	Lap(550) $70400 \log(22/\delta) + 22$	Lap(5.5) $704 \log(22/\delta) + 22$	Lap(2.75) $352 \log(22/\delta) + 22$
Rochdale	Lap(362.5) $185600 \log(29/\delta) + 29$	Lap(3.625) $1856 \log(29/\delta) + 29$	Lap(1.8125) $928 \log(29/\delta) + 29$
Journey to work	Lap(132) $8450 \log(169/\delta) + 169$	Lap(1.32) $84.5 \log(169/\delta) + 169$	Lap(0.66) $42.25 \log(169/\delta) + 169$

the perturbed tables as a function of ϵ. It is easy to see that the smaller ϵ the more variable the sample sizes of the perturbed tables become. In particular, when ϵ is very small, the sample size become unrealistically large, order of magnitudes larger than the true sample sizes. In fact, even for values of ϵ as large as 2 (which is a rather weak privacy guarantee) the sample size is highly variable—we deem this to be a serious problem for statistical analysis.

Figure 2 shows the maximal L_1 distance between the margins of the true and perturbed tables as a function of ϵ. Similarly to what we pointed out above, for a wide spectrum of values of ϵ, which provides good privacy guarantees, these discrepancies are significantly larger than the sample size, so that the perturbations induced by privacy protecting mechanism may mask or even destroy any underlying statistical signal. We see similar patterns in Figure 3 which shows the L_2 distance between the Likelihood Ratio Statistics(LR) of the original table and perturbed tables.

Figure 4 shows the L_1 distance between the MLE of the cell probabilities computed using the original table with the MLE obtained from the perturbed margins, as a function of ϵ. We recall that this value has a well-known probabilistic interpretation, as it is twice the total variation distance between the probability distribution over the cells specified by the MLE of the original table and the probability distribution specified by the MLE of the perturbed table. The maximal value of this distance is 2, which corresponds to mutually singular probability distributions (i.e. having disjoint supports). As we expected, Figure 4 shows that this distance gets increasing larger as the privacy parameter ϵ gets smaller, with values that are quite high even when ϵ is large, thus providing only weak privacy guarantees. To get a sense of how much the privacy mechanism effects the total variation distance, we computed this distance between the MLE of the cell probabilities based on the original table and the uniform distribution over the cells for each of our four tables: Edwards–0.83, Czech–0.86, Rochdale–1.43 and Journey to work–1.43. Thus we conclude that, when ϵ is small, the MLE of the perturbed table will be at roughly the same probabilistic distance from the true MLE than a uniform distribution over the cells. While this may lead to a satisfactory privacy protection, it will essentially disrupt any possibility of a meaningful statistical analysis.

Finally, in our last experiment we investigated whether the linear programing optimization problem compromising step 5 in the algorithms of Table 1 and 2 is feasible. The reason why we consider this as an important issue is that unfeasibility of the program implies that there does not exist any real-valued table with margins matching the perturbed margins. In this case, the optimization problem will return a real-value non-negative table whose margins are closest to the perturbed margins. This additional approximation in effect constitutes an additional perturbation to the original table that is completely unaccounted for by the theory. Even though this extra perturbation is likely to strengthen the privacy guarantees even further, the statistical consequences are rather negative. In fact, not only is it extremely hard to quantify directly the magnitude of such approximation, but it almost certainly will make the perturbed table even more statistically dissimilar from the true table. Figure 5 shows the proportion of times, out the 50 simulations and as a function of ϵ, the optimal values of b in the linear programming part of the algorithm of Table 1 is larger than 0 for the Edward's fungus data. We recall that a positive value of b implies unfeasibility. It is immedate from the plots that a large proportion of the simulations are resulted in an unfeasible problem. Figure 6 shows instead the actual optimal values of b for the Edward's fungus data. As we see from this figure, not not only is the optimization problem frequently unfeasible, but the optimal values of b can be extremely large.

Based on the experiments we summarize above, we see a clear pattern even for the non-sparse Czech autoworkers example. As the noise level, controlled by the parameter ϵ, increases, the deviance between the generated tables and their MLEs get smaller. This means that if we add too much noise, we get strong privacy guarantees but inadequate and potentially misleading statistical inference. On the other hand, when we add little noise, the statistical inference is better but the differential privacy guarantees appear to have little practical value.

7 Conclusions

We have re-examined the differential privacy approach to the protection of pre-specified margins from a multi-way binary contingency tables proposed by Barak et al. [1], and we have extended their methodology using the Efron-Stein decomposition so that it is directly applicable to non–binary tables. Then we analyzed the theoretical claims in the original Barak et al. paper and we discovered clear shortcomings. In order to understand how the choice of the key noise parameter ϵ situates the methodology from the perspective of the risk-utility trade-off developed in the statistical literature on confidentiality, we applied the methodology in a systematic fashion to three binary tables (Edwards' fungus data, the Czech autoworkers data, and the Rochdale survey extract), and to the non-binary journey-to-work table. Through a simulation study for each of the four examples, we demonstrated what we deem to be serious problems with the methodology as originally proposed and with our related extension.

Differential privacy remains an attractive methodology because of its clear definition of privacy and the strong guarantees that it promises. But much is hidden in the noise parameter, ϵ, especially in the context of the proposed implementations to date. Moreover, because differential privacy provides guarantees for the method and not for the specific data at hand, we do not believe the methodology is suitable for the type of large sparse tables often produced by statistics agencies and sampling organizations. Our preference remains for the less formal but seemingly effective approach described by Fienberg and Slavkovic [16], Dobra et al. [4] and Winkler [22].

References

1. Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K. (2007) "Privacy, accuracy, and consistency too: A holistic solution to contingency table release." *Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*.
2. Bishop, Y. M. M., Fienberg, S. E., Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA. Reprinted: Springer, New York (2007).
3. Christiansen, S. K., Giese, H. (1991) "Genetic analysis of obligate barley powdery mildew fungus based on RFLP and virulence loci." *Theoretical and Applied Genetics*, 79, 705–712.
4. Dobra, A., Fienberg, S. E., Rinaldo, A., Slavkovic, A. B., Zhou, Y. (2008) "Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation, and disclosure limitation." In M. Putinar and S. Sullivant, eds., *Emerging Applications of Algebraic Geometry*, IMA Series in Applied Mathematics, Springer, New York, 63–88.
5. Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R., Roehrig, S. F. (2001) "Disclosure limitation methods and information loss for tabular data." In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds., *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier, Amsterdam, 135–166.
6. Dwork, C. (2006) "Differential privacy." In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, eds., *ICALP (2)*, Volume 4052, Lecture Notes in Computer Science, Springer, Berlin,1–12.
7. Dwork, C., McSherry, F., Nissim, K., Smith, A. (2006) "Calibrating noise to sensitivity in private data analysis." In S. Halevi and T. Rabin, eds., *TCC*, Volume 3876, Lecture Notes in Computer Science, Springer, Berlin, 265–284.
8. Dwork, C. and Naor, M. (2010) "On the difficulties of disclosure prevention in statistical databases or the case for differential privacy." *Journal of Privacy and Confidentiality*, 2 (1), Article 8.
9. Dwork, C. and Smith, A. (2009) "Differential privacy for statistics: What we know and what we want to learn." *Journal of Privacy and Confidentiality*, 2 (1), Article 2.
10. Edwards, D (1992) "Linkage analysis using log-linear models." *Computational Statistics and Data Analysis,* 13, 281–290.
11. Edwards, D(2000) *Introduction to Graphical Modeling,* (2nd ed.). Springer, New York.
12. Edwards, D. and Havranek, T. (1985) "Fast procedure for model search in multidimensional contingency tables." *Biometrika*, 72, 339–351.
13. Efron, B. and Stein, C. (1996) "The jackknife estimate of variance." *Annals of Statistics*, 9(3), 586–596.
14. Fienberg, S. E. and Love, T. (2009) "Disclosure avoidance techniques to improve ACS data availability for transportation planners," National Cooperative Highway Research Program Project 08-36, Task 71 Final Report (May 2009). Available online at: `http://onlinepubs.trb.org/onlinepubs/archive/NotesDocs/NCHRP08-36(71)_FR.pdf`
15. Fienberg, S. E., Rinaldo, A., and Yang, X. (2010) "Differential privacy and The risk-utility tradeoff for multi-dimensional contingency tables," In J. Domingo-Ferrer and E. Magkos, eds., *Privacy in Statistical Databases 2010 (PSD 2010)*, LNCS Vol. 6344, Springer, Berlin (2010), pp. 187–199.
16. Fienberg, S. E. and Slavkovic, A. B. (2008) "A survey of statistical approaches to preserving confidentiality of contingency table entries." In C. Aggarwal and P. S. Yu,

eds., *Privacy Preserving Data Mining: Models and Algorithms*, Springer, New York, 289–310.
17. Lauritzen, S. L. (1996) *Graphical Models.* Oxford University Press, Oxford.
18. Nabar, S. U. and Mishra, N. (2010) "Releasing Private Contingency Tables." *Journal of Privacy and Confidentiality*, 2 (1), Article 9.
19. Mossel, E. (2009) "Gaussian bounds for noise correlation of functions." *Geometric and Functional Analysis*, 19 (6), 1713–1756.
20. Wasserman, L. and Shuheng, Z. (2010) "A statistical framework for differential privacy." *Journal of the American Statistical Association*, 105, 375–389.
21. Whittaker, J. (1990) "Graphical Models in Applied Multivariate Statistics". Wiley, New York.
22. Winkler, W. (2008) "General Discrete-data Modeling Methods for Producing Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties." Statistics Research Report Series, 2010-02, U.S. Bureau of the Census, Washington, DC.

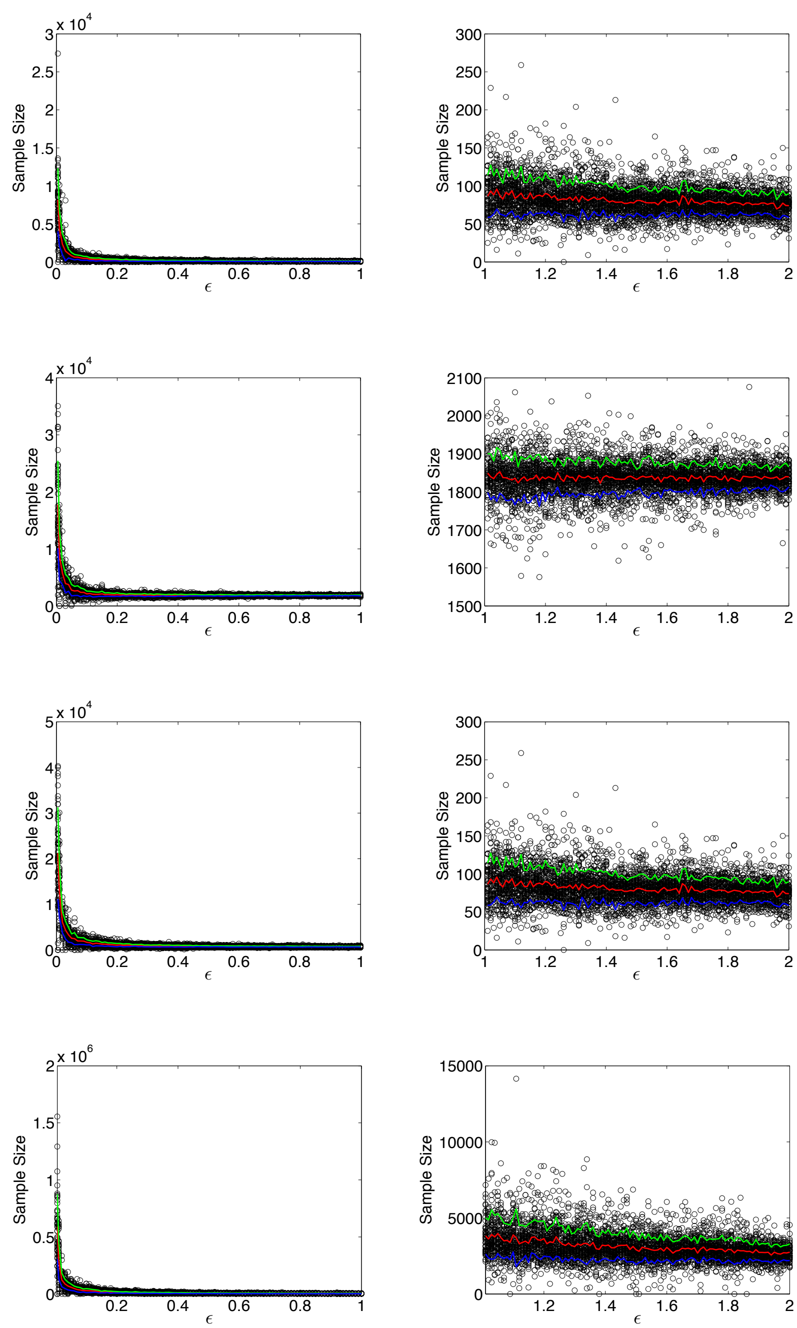


Fig. 1. Sample sizes for the Fungus table (top row), Czech autoworker table (second row), Rochdale table (third row) and Journey to work table (bottom row). To improve readability, for each table,we split the plot in two parts, for $\epsilon < 1$ (left) and $\epsilon \geq 1$ (right). The three lines represent the mean plus or minus one standard deviation.

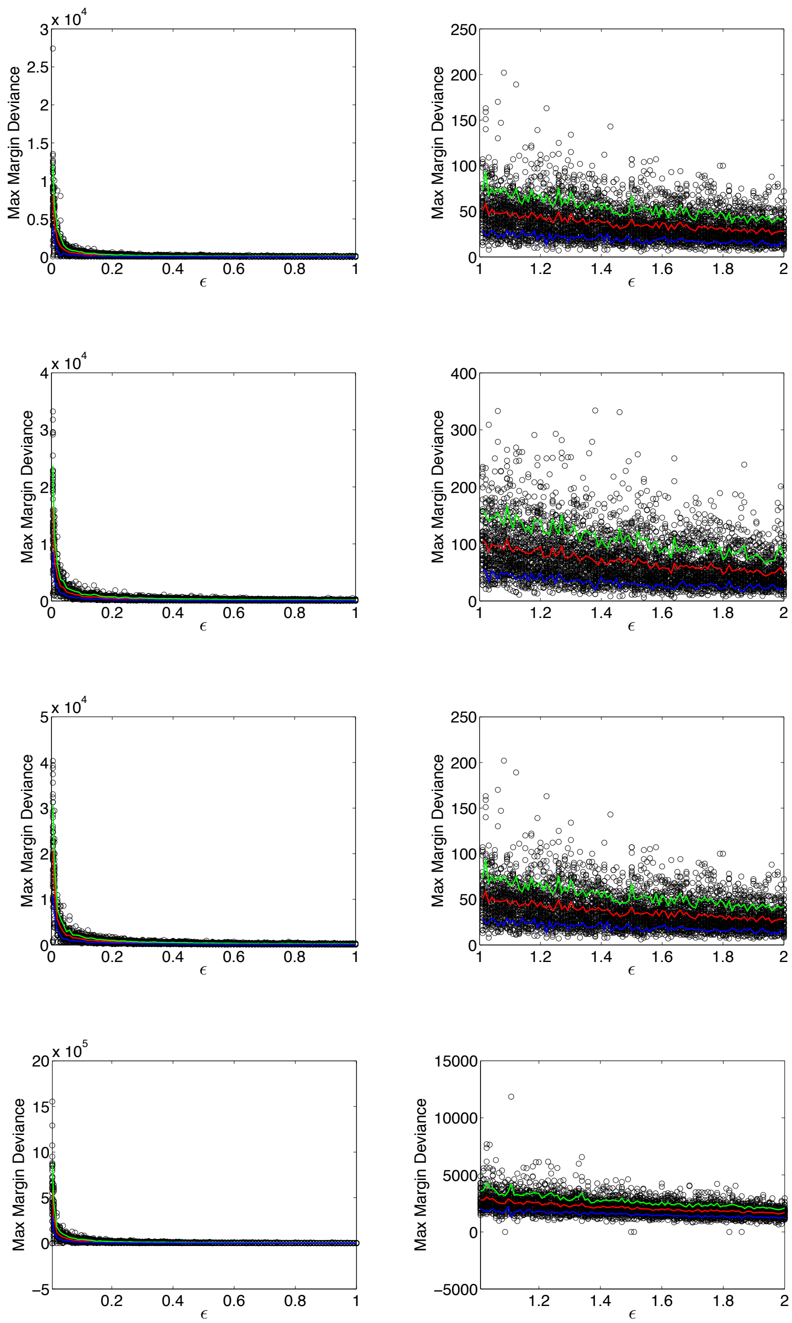


Fig. 2. Maximal L_1 difference between the true and perturbed margins for the Fungus table (top row), Czech autoworker table (second row), Rochdale table (third row) and Journey to work table. To improve readability, for each table, we split the plot in two parts, for $\epsilon < 1$ (left) and $\epsilon \geq 1$ (right). The three lines represent the mean plus or minus one standard deviation.

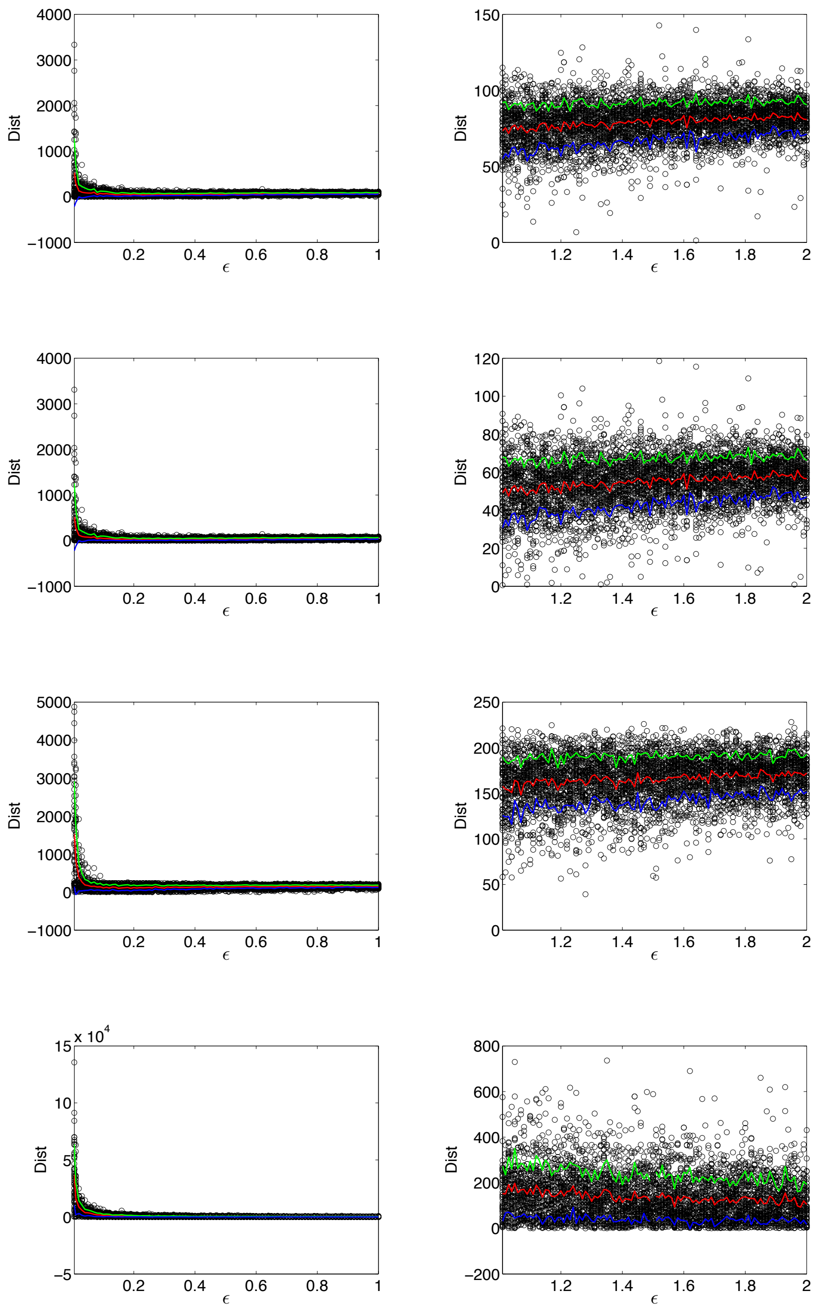
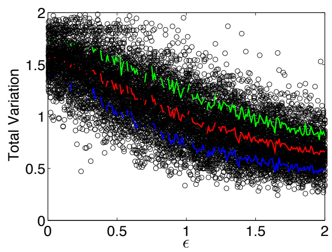
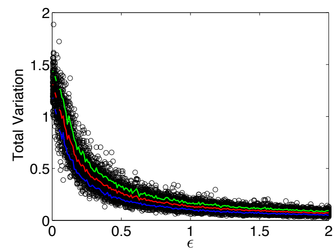


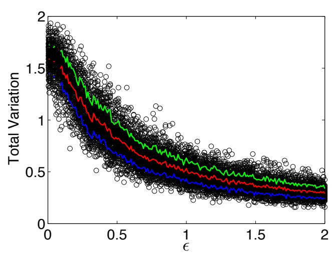
Fig. 3. The $L2$ distance of likelihood ratio statistics between the perturbed tables and the original tables: Fungus table (top row), Czech autoworker table (second row), Rochdale table (third row) and Journey to work table. To improve readability, for each table, we split the plot in two parts, for $\epsilon < 1$ (left) and $\epsilon \geq 1$ (right). The three lines represent the mean plus or minus one standard deviation.



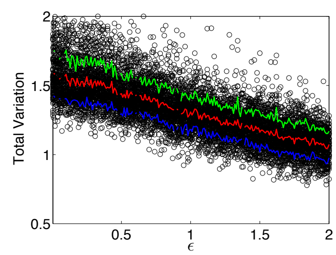
(a)



(b)



(c)



(d)

Fig. 4. Total variation distance between the MLE of the chosen model based on the original table and the MLE based on the perturbed tables as a function of ϵ the Fungus table (a), Czech autoworker table (b), Rochdale table (c) and Journey to work table (d). The three lines represent the mean plus or minus one standard deviation.

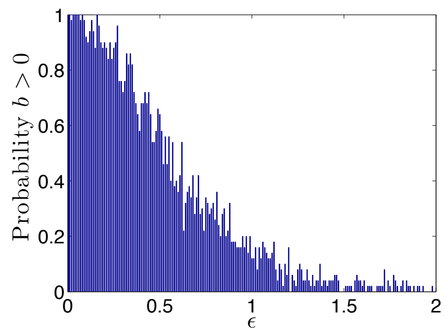


Fig. 5. Fraction of times the optimal value of b in the linear programming part of the algorithm of Table 1 was larger than 0 as a function of ϵ for the fungus table.

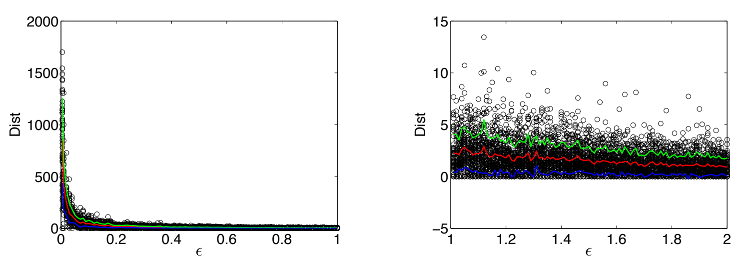


Fig. 6. Optimal values of b for the linear programming part of the algorithm of Table 1 as a function of ϵ for the fungus table.