Differential Privacy and the Risk-Utility Tradeoff for Multi-dimensional Contingency Tables

Stephen E. Fienberg[1,2,3], Alessandro Rinaldo[1,2], and Xiaolin Yang[1,⋆]

[1] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
fienberg@stat.cmu.edu, arinaldo@stat.cmu.edu, xyang@stat.cmu.edu
[2] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] Cylab, and i-Lab, Carnegie Mellon University, Pittsburgh, PA 15213, 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. A recent paper extends 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.

1 Introduction

Contingency tables, databases arising from 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 [2,9,12]. Fienberg and Slavkovic [11] review 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 in Duncan et al. [5], where risk was defined in terms of protection of small counts in the table. Dobra et al. [4] further insight into the RU-trade-off problem for large sparse tables using recent results from algebraic statistics. Winkler [15] proposed a method to reduce re-identification risk while preserving analytic properties by placing upper and lower bounds on key aggregates needed for loglinear modeling and also on large sets of small cells and sampling zeros.

The methodology of differential privacy [6,7] 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. A recent paper by

[⋆] Corresponding author.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 187–199, 2010.

Barak et al. [1] extends this approach to the release of a specified set of margins from a multi-way contingency table. Adding non-integer noise in such contexts poses a variety of additional problems: violation of non-negativity, incompatible margins, and infeasible tables. The proposed methodology purports to handle all of these problems. In this paper, 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.

2 Differential Privacy

Let $\mathcal{D}$ denote the set of databases. A privacy protecting mechanism is a randomized function $K\colon\mathcal{D} \rightarrow \mathcal{D}$. 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,*

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

The smaller 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 [13, Theorem 2.4] provide a related statistical interpretation of differential privacy based on hypothesis testing theory.

3 Notation for Binary Contingency Tables

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 x as a vector in R^{2^k}. We represent each cell i of the table x as a vertex of the k-dimensional unit hypercube: $x = \left\{x_i, i \in \{0,1\}^k\right\}$. For a given subset $\alpha \subset \{1,\ldots,k\}$, we write $i_\alpha = \{i_j, j \in \alpha\} \in \{0,1\}^{|\alpha|}$ for the α-coordinate projection of i. The α-marginal table of x is the $|\alpha|$-dimensional binary array $x^\alpha = \left\{x_{i_\alpha}, i_\alpha \in \{0,1\}^{|\alpha|}\right\}$, whose i_α entry is obtained by summing over the cells j of x having identical α-coordinate projection:

$$x_{i_\alpha} = \sum_{j : j_\alpha = i_\alpha} x_j. \tag{1}$$

We will write compactly $x^\alpha = C^\alpha x$, were C^α is the $2^{|\alpha|} \times 2^k$ matrix realizing the sums in equaition (1). Also, with a slight abuse of notation, we refer to both α and x^α as margins.

4 The Risk-Utility Trade-Off

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

From the theory of log-linear models [2,12], we know that each such collection $A \subset 2^{\{0,1\}^k}$ encode a statistical model for the the probabilistic dependence among the k attributes, each of which as a categorical random variable. Specifically, each 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 in the simplex in R^{2^k} such that p_i denotes the probability of observing cell i. The corresponding marginal tables $\{x^\alpha, \alpha \in A\}$ are minimal sufficient statistics for the model determined by A. This means that, from an inferential standpoint, the A-margins of x contains as much statistical information as x itself. Furthermore, they determine the maximum likelihood estimator (MLE) $\hat{p}$, which is the unique probability distribution in the model encoded by A that makes x the "most likely" sample that we could have observed. The MLE possess many optimal properties and, in particular, and we can use it to assess the fit of the model A using the likelihood ratio test statistic

$$\sum_i x_i \log\left(\frac{x_i}{n\hat{p}_i}\right). \tag{2}$$

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 A-margins contain only aggregate (partial) information about x and thus appear to be a natural candidates for a data release [11,4], 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 The Differential Privacy Mechanism for Contingency Tables

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 $\{0,1\}^k$. For vectors $\alpha, \beta \in R^{2^k}$, we will denote the L_1 norm as $\|\alpha\|_1 = \sum_i |\alpha_i|$ and the standard inner product as $\langle \alpha, \beta \rangle = \sum_i \alpha_i \beta_i$. Let $\{f^\alpha, \alpha \in \{0,1\}^k\}$ be the Fourier basis for R^{2^k}, whose α element is the vector $f^\alpha = \{f^\alpha_\beta, \beta \in \{0,1\}^k\}$, where

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

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

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

where for $\alpha, \beta \in \{0,1\}^k$, $\alpha \preceq \beta$ signifies that every non-zero coordinate of α is also a non-zero coordinate of β. 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 [12, 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 A from a binary table x, which we reproduce in Table 1. They showed that the algorithm possesses the following properties.

Theorem 1. *The privacy mechanism of Table 1 satisfies differential privacy and, for each* $\delta \in (0,1)$*, with probability at least* $(1-\delta)$*,*

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

uniformly over all $\alpha \in A$*.*

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 a margins that can be realized by an integer-valued table (namely w').

Remarks

1. The result is independent of the sample size, and the accuracy guarantees depend only on the model complexity $|B|$ and the differential privacy parameter ϵ.

Table 1. The differentially private mechanism for binary contingency tables

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

$$\begin{gathered} \min b \\ \text{subject to:} \\ w_\alpha \geq 0, \quad \forall \alpha \\ \phi_\beta - \textstyle\sum_\alpha w_\alpha f_\alpha^\beta \leq b, \quad \forall \beta \in B \\ \phi_\beta - \textstyle\sum_\alpha w_\alpha f_\alpha^\beta \geq -b, \quad \forall \beta \in B. \end{gathered}$$

6. Round w to w', where, for each $\alpha \in \{0,1\}^k$, w'_α is the nearest integer to w_α.
7. Return the A-margins of w'.

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 B-margins given by $\{\phi^{\beta}, \beta \in B\}$.
3. The linear program has typically many (in fact infinite) solutions.
4. The proof of Theorem 7 in [1] implicitly assume that $b = 0$, which, as we mentioned, does not hold in general.

6 Empirical Evaluation of the Differential Privacy Mechanism

We now analyze the statistical properties of the privacy preserving mechanism of [1] on the three real-life datasets. We study empirically whether the algorithm in Table 1 for producing differentially private results,is also statistically robust, 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. We work with three well-analyzed examples, the full data for which we provide in the appendix:

1. The data in Table 4 is a sparse 6-dimensional binary contingency table that was obtained from the cross-classification of six dichotomous categorical variables, labeled with the letters A-F, recording the parental alleles corresponding to six

Table 2. Table dimension, sample size, chosen model and likelihood ratio statistic (2) for the three tables analyzed

Table	Dimension	Sample Size	Model	LR
Edwards	$k = 6$	$n = 70$	[AD][AB][BE][CE][CF]	22.96
Czech	$k = 6$	$n = 1841$	[BF][ADE][ABCE]	48.18
Rochdale	$k = 8$	$n = 665$	[ACE][ACG][ADG][BDH] [BF][BE][CEF][CFG]	238.18

Table 3. Variance of the additive noise and bounds for different values of ϵ

	ϵ		
	0.01	1	10
Edwards	Lap(300) $38400 \log(12/\delta) + 12$	Lap(3) $384 \log(12/\delta) + 12$	Lap(0.3) $38.4 \log(12/\delta) + 12$
Czech	Lap(550) $70400 \log(22/\delta) + 22$	Lap(5.5) $704 \log(22/\delta) + 22$	Lap(0.55) $70.4 \log(22/\delta) + 22$
Rochdale	Lap(362.5) $185600 \log(29/\delta) + 29$	Lap(3.625) $1856 \log(29/\delta) + 29$	Lap(0.3625) $185.6 \log(29/\delta) + 29$

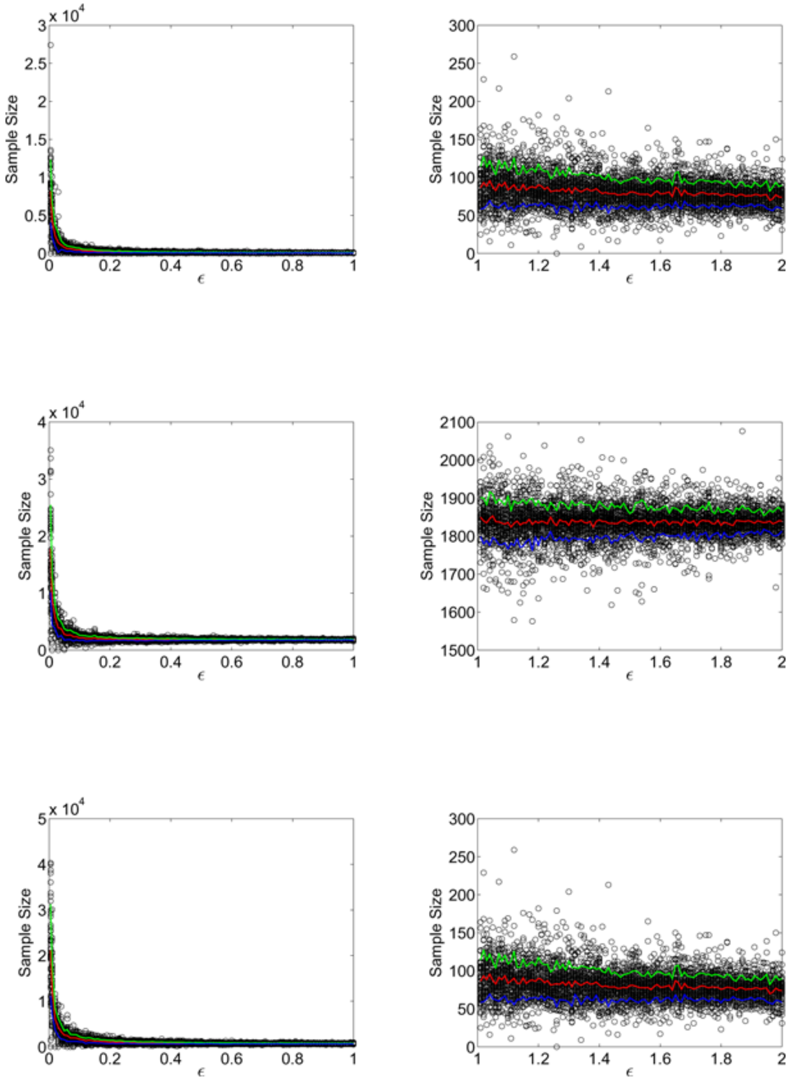


Fig. 1. Sample sizes for the Fungus table (top row), Czech autoworker table (middle) and Rochdale table (bottom). To improve readability, for each table,we split the plot in two parts, for $\epsilon < 1$ (left) and $\epsilon \geq 1$ (right). The three lines represent the mean plus or minus one standard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

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 [8]. Based on the model selection analysis described in [9], 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.

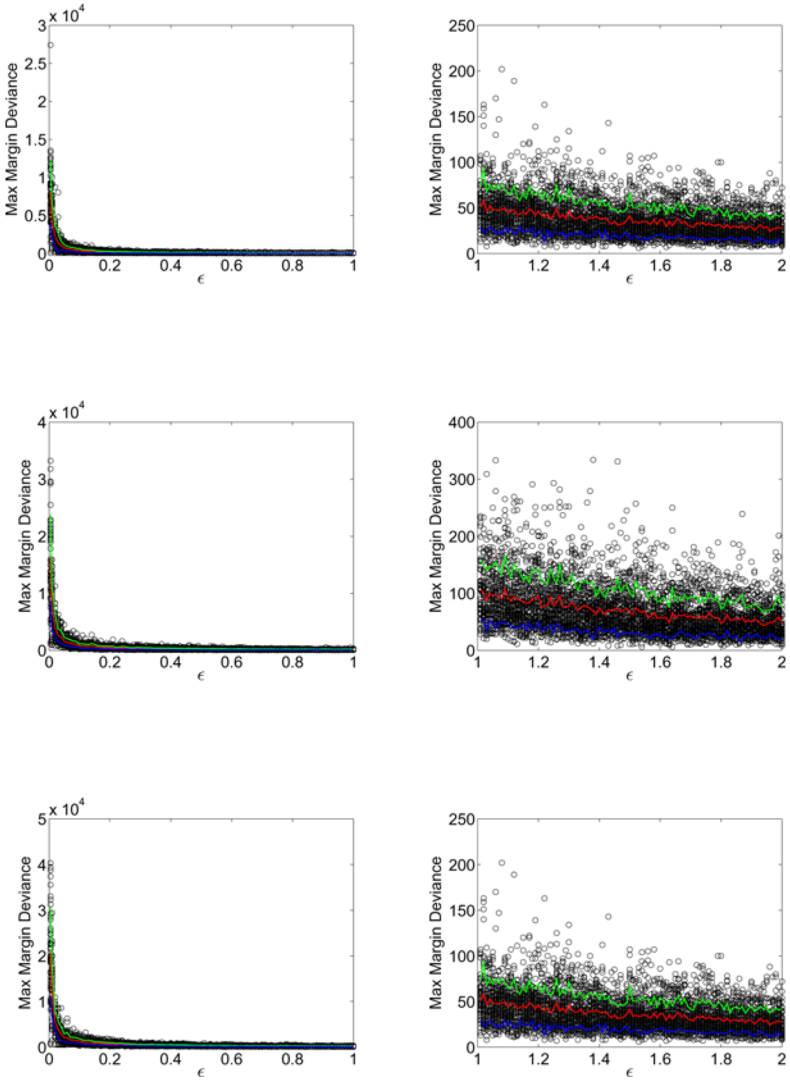


Fig. 2. Maximal L_1 difference between the true and perturbed margins for the Fungus table (top row), Czech autoworker table (middle) and Rochdale table (bottom). To improve readability, for each table, we split the plot in two parts, for $\epsilon < 1$ (left) and $\epsilon \geq 1$ (right). The three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

2. The data in Table 5 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 [10]. 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

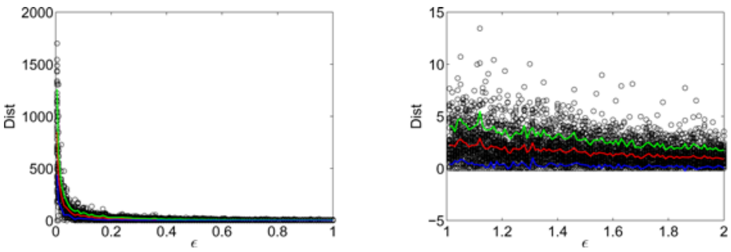


Fig. 3. Optimal values of b for the linear programming part of the algorithm of Table 1 as a function of ϵ for the fungus table

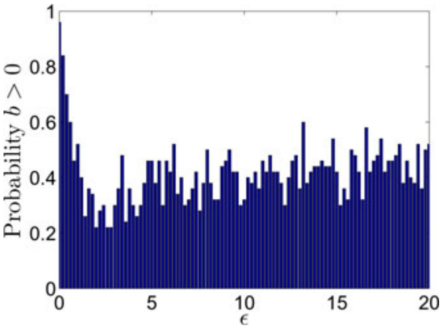


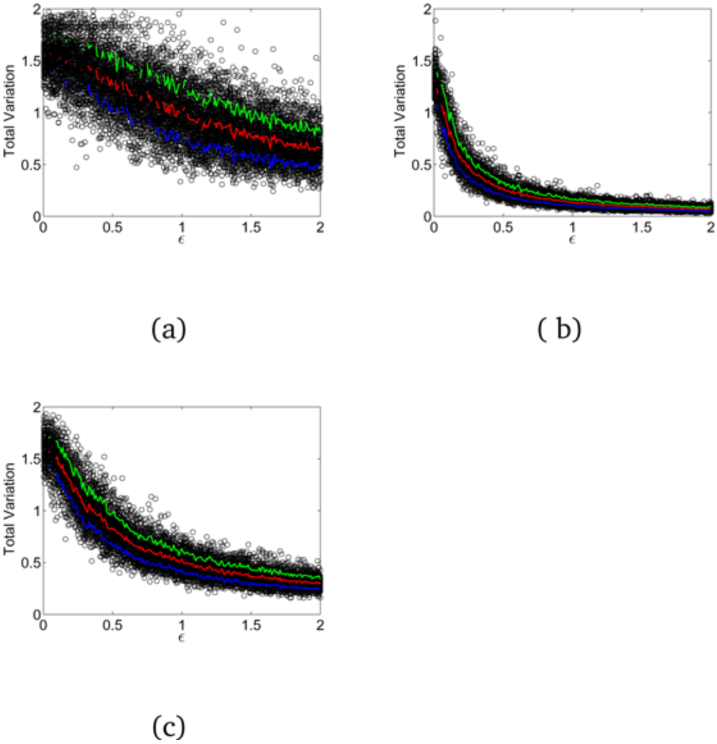
Fig. 4. Fraction of times the optimal value of b in the linear programming part of the algorithm of Table 1 was larger than 0 as a function of ϵ for the fungus table

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.

3. The data in Table 6 involve 8 binary variables (Yes/No) relating women's economic activity and husband's unemployment from a survey of households in Rochdale [14, 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.

For a grid of values of ϵ from 0.005 to 2, we perturbed each of the three tables 50 times using the algorithm of [1]. We summarize key results in a series of figures:

- Figure 1 shows the sample size of 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



(a)

(b)

(c)

Fig. 5. 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) and Rochdale table (c). The three lines represent the mean plus or minus one standard deviation.

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. Once again, for values of ϵ as large as 5, these discrepancies are of the same order of magnitude as the sample size.
- Figure 3 shows the optimal values of b in the linear programming part of the algorithm of Table 1 for the Edward's fungus data as a function of ϵ.
- Figure 4 shows the proportion of times b is larger than 0, which means that there does not exists a real-valued non-negative tables whose margins match the margins of the perturbed table.
- Figure 5 shows the total variation distance between the MLE of the cell probabilities computed using the original distance with the MLE obtained from the perturbed margins, as a function of ϵ. Total variation distance is at most 2. 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 three tables: Edwards–0.83, Czech–0.86, and Rochdale–1.43.

Space precludes a detailed analysis of the information summarized in these figures but 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 is 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 have little practical use.

7 Conclusions

We have explored the differential privacy approach to margin protection in contingency tables proposed by Barak et al. [1]. First we analyzed the theoretical claims and we discovered clear shortcomings. Second, we applied the methodology in a systematic fashion to three binary tables (Edwards fungus data, the Czech autoworkers data, and the data from Rochdale), 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. Through a simulation study for each of the three examples, we demonstrated what we deem to be serious problems with the methodology as originally proposed.

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 methods of Barak et al. 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 [11], Dobra et al. [4] and Winkler [15].

Acknowledgement

This research was partially supported by Army contract DAAD19-02-1-3-0389 to Cylab at Carnegie Mellon University and in part by NSF grant DMS-0631589 and a grant from the Pennsylvania Department of Health through the Commonwealth Universal Research Enhancement Program to the Department of Statistics, Carnegie Mellon University.

References

1. Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K.: Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In: Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (2007)

2. Bishop, Y.M., Fienberg, S.E., Holland, P.W.: Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge (1975); reprinted: Springer (2007)
3. Christiansen, S.K., Giese, H.: Genetic analysis of obligate barley powdery mildew fungus based on rfpl and virulence loci. Theoretical and Applied Genetics 79, 705–712 (1991)
4. Dobra, A., Fienberg, S.E., Rinaldo, A., Slavkovic, A.B., Zhou, Y.: Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation, and disclosure limitation. In: Putinar, M., Sullivant, S. (eds.) Emerging Applications of Algebraic Geometry. IMA Series in Applied Mathematics, pp. 63–88. Springer, Heidelberg (2008)
5. Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R., Roehrig, S.F.: Disclosure limitation methods and information loss for tabular data. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 135–166. Elsevier, Amsterdam (2001)
6. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)
7. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006)
8. Edwards, D.: Linkage analysis using log-linear models. Comp. Statist. and Data Anal. 13, 281–290 (1992)
9. Edwards, D.: Introduction to Graphical Modelling, 2nd edn. Springer, Heidelberg (2000)
10. Edwards, D., Havranek, T.: Fast procedure for model search in multidimensional contingency tables. Biometrika 72, 339–351 (1985)
11. Fienberg, S.E., Slavkovic, A.B.: A survey of statistical approaches to preserving confi- dentiality of contingency table entries. In: Aggarwal, C., Yu, P.S. (eds.) Privacy Preserving Data Mining: Models and Algorithms, pp. 289–310. Springer, Heidelberg (2008)
12. Lauritzen, S.L.: Graphical Models. Oxford University Press, Oxford (1996)
13. Wasserman, L., Shuheng, Z.: A statistical framework for differential privacy. J. Amer. Statist. Assoc. 105, 375–389 (2010)
14. Whittaker, J.: Graphical Models in Applied Multivariate Statistics. Wiley, Chichester (1990)
15. Winkler, W.: General Discret-data Modeling Methods for Producing Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties. Research Report Series, Statistics 2010-02 (2008)

Appendix

Table 4. Cell counts 2^6 table involving genetic linkage in barley powder mildew fungus. Source: Edwards [8].

			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									

Table 5. Cell counts for Czech autoworker 2^6 table. Source: Edwards and Havranek [10].

			1				2				C
			1		2		1		2		B
			1	2	1	2	1	2	1	2	A
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	
F	E	D									

Table 6. Rochdale table. Source: Whittaker [14].

	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																	