Confidentiality and Differential Privacy in the Dissemination of Frequency Tables

Yosef Rinott, Christine M. O'Keefe, Natalie Shlomo and Chris Skinner

Abstract. For decades, national statistical agencies and other data custodians have been publishing frequency tables based on census, survey, and administrative data. In order to protect the confidentiality of individuals represented in the data, tables based on original data are modified before release. Recently, in response to user demand for more flexible and responsive table publication services, frequency table publication schemes have been augmented with on-line table generating servers such as the US Census Bureau FactFinder and the Australian Bureau of Statistics TableBuilder. These systems allow users to build their own custom tables, and make use of automated perturbation routines to protect confidentiality. Motivated by the growing popularity of table generating servers, in this paper we study confidentiality protection for perturbed frequency tables, including the trade-off with analytical utility. Confidentiality protection is assessed in terms of the differential privacy standard, and this paper can be used as a practical introduction to differential privacy, to calculations related to its application, and to the relationship between confidentiality protection and utility.

Key words and phrases: Differential Privacy, Statistical Disclosure Control, Contingency Tables, Utility.

Yosef Rinott is Professor, The Federmann Center for the Study of Rationality, The Hebrew University, Jerusalem 91904 Israel and LUISS, Rome. (e-mail: yosef.rinott@mail.huji.ac.il). Christine O'Keefe is Senior Principal Research Scientist, CSIRO, GPO Box 664, Canberra ACT 2602, Australia. (e-mail: Christine.O'Keefe@csiro.au). Natalie Shlomo is Professor of Social Statistics, University of Manchester, Manchester M13 9PL, United Kingdom. (e-mail: natalie.shlomo@manchester.ac.uk). Chris Skinner is Professor of Statistics, London School of Economics and Political Science, London WC2A 2AE, United Kingdom. (e-mail: c.j.skinner@lse.ac.uk).

[*]The work of all authors was partially supported by the Simons Foundation. All authors thank the Isaac Newton Institute for Mathematical Sciences, University of Cambridge, for support and hospitality during the programme Data Linkage and Anonymisation which was supported by EPSRC grant no EP/K032208/1. We also thank those participants in the programme from the differential privacy computer science community, from whom we learnt a great deal.

1. INTRODUCTION

Sharing data for statistical purposes is increasingly important. National statistical agencies and other custodians collecting data from individuals are obliged to keep such information strictly confidential to the agency, including not sharing or releasing such data in an identifiable form. Therefore, a key constraint on data sharing is the need to protect the confidentiality of the individuals or other entities to which the data refer. A canonical confidentiality protection problem can be formulated as follows. For given data, denoted D, how can we determine a (possibly stochastic) transformation $\mathcal{M}(\cdot)$, called a *perturbation mechanism* (or simply *mechanism*), such that if $\mathcal{M}(D)$ is disseminated then confidentiality will be protected and also the value of D for statistical analysis, called *utility*, will be preserved in $\mathcal{M}(D)$?

A key issue in the development of solutions to this problem is how to define confidentiality and utility. The basic idea of utility should be more familiar territory to statisticians. If the data are being disseminated for statistical purposes, for example for estimation of various parameters, then the reduction in utility arising from releasing $\mathcal{M}(D)$ rather than D might be measured in terms of increases in the bias and variance of the resulting estimators. The question of how to measure confidentiality has historically been a more specialised topic in statistics and has been considered mainly within the field of *statistical disclosure control* (SDC), which has developed in association with a long tradition of data dissemination practice by government statistical offices (see Duncan, Elliot and Salazar-Gonzàlez, 2011; Hundepool et al., 2012; Willenborg and de Waal, 2001).

To protect the confidentiality of individuals in a data set D, *de-identification*, that is, removing identifiers such as names, addresses, and identification numbers from D before its release, is standard. However, this may not prevent a knowledgeable intruder from obtaining information about individuals in D, see (O'Keefe and Chipperfield, 2013). Here is a simple example: let D represent a t-way frequency table with counts of individuals having certain combinations of t attributes in a certain population, or a sample from the population. Suppose an intruder knows that there is an individual in the population with a given combination of r of the attributes for some $r < t$, and that this individual is the only one with this combination. If this individual is in D, and D is released, the intruder can locate the individual on the basis of the r known attributes, and then learn all other $t - r$ attributes.

Although there are established measures of disclosure risk used widely in practice and studied in the SDC literature cited above and in references therein, there is considerable interest in alternative ways of measuring confidentiality for a number of reasons. Such reasons include that existing methods may be based upon contestable assumptions about an intruders' prior knowledge of the data and type of confidentiality attacks which they might employ, and that the continuing evolution of approaches to data dissemination requires flexible approaches to confidentiality that can be applied in systematic ways.

In this paper we focus on *differential privacy* (Dwork et al., 2006) as a way of defining confidentiality, measuring confidentiality protection, and comparing perturbation mechanisms. Differential privacy has recently been attracting a lot of attention in the computer science literature, see for example the recent book, Dwork and Roth (2014), and its references. The idea has been introduced in a mathematically rigorous framework with the potential for wide application and, by employing a 'worst case' approach, avoids strong assumptions about which variables are sensitive to disclosure, and intruders' prior knowledge and attack scenarios, leading to a well-defined quantification of the confidentiality protection guarantee. This worst case approach may be deemed overprotective of confidentiality, however this is intentional as it is designed to protect against a potentially sophisticated adversary who may take advantage of a rare weakness of the release mechanism. Only time will tell whether differential privacy as a risk measure, or some of its relaxations, will be applied by official agencies. In any case, we find it very illuminating as a framework of thinking about SDC.

Our goal in this paper is to explore and describe the application of the notion of differential privacy under a realistic and popular dissemination scenario and, on the way, to provide a practical introduction to this notion for statisticians. We shall focus on the dissemination of frequency tables in a government statistical setting, where the underlying data D are cross-classified tables of frequencies. Further, in order to keep our discussion realistic, where possible we shall model our system requirements and objectives (but not our perturbation mechanism) on the existing Australian Bureau of Statistics (ABS) TableBuilder system (Chipperfield, Gow and Loong, 2016). We shall derive the results from the theory of differential privacy that are useful to us in the most direct ways, not trying to present the theory in full generality, but trying to keep this paper almost self contained.

Since increased confidentiality protection is generally traded off against reduced utility, it is vital to know how alternative confidentiality protection methods affect what utility can be achieved. The guarantee in differential privacy is defined in terms of one or two parameters, allowing different perturbation schemes to be compared by fixing these parameters, and comparing the utility of the perturbed data. The analytic impact of perturbation will depend heavily on the kinds of analyses undertaken and these may be hard to anticipate at the time the protection takes place, but see Karr et al. (2006) for a general framework for evaluating utility. Even given a definition of utility, one perturbation scheme may be preferable to another for some values of the parameters, and vice versa for other values. We shall use differential privacy parameters to compare perturbation mechanisms, but when we extend the comparison to also include utility, then it will clearly depend on both the parameters and on how utility is measured, and therefore it is not straightforward.

In order to help to put our work in its historical context, we now give a brief review of disclosure risk assessment and confidentiality protection methods for frequency tables, see Duncan et al. (2001); Hundepool et al. (2012); Shlomo (2007). Disclosure risk assessment typically focuses on small cell counts and on the possibility that information on one classifying variable can be learnt about an individual for whom values of other classifying variables are known. This is usually called *attribute disclosure* (Shlomo, 2007), in contrast to *identity disclosure* in which information in the data is associated with an individual. The occurrence of counts of 1 in the table may be treated as a potential problem of identity disclosure in itself but can also magnify the threat of attribute disclosure if a second table is available cross-classifying these variables with a further variable, leading to what may be called *residual disclosure* (Fellegi, 1972).

There are two main classes of confidentiality protection methods for frequency tables, namely, *pre-tabular* methods that modify microdata before aggregation into a table, and *post-tabular* methods that modify a table directly. Any method for protecting confidentiality in microdata can be used as a pre-tabular confidentiality protection method, including: rounding, suppression of variables or variable values, variable recoding, sampling, data swapping, perturbation, and post-randomisation methods similar to randomised response. Synthetic data (Little, 1993; Rubin, 1993) methods could also be used, see Drechsler (2011); Drechsler and Reiter (2011). In this approach, the original process that generated the microdata is modelled, and synthetic microdata are generated from this model with a view to preserving the statistical properties of the implied table. Post-tabular methods are generally conducted in two steps. The first step is to identify whether the release of the data in any table cell could lead to a disclosure, for example, if the cell contains a very small count. The second step is to reduce the disclosure risk associated with the identified cells, with a method such as table redesign, cell suppression, rounding, or addition of noise. Table redesign typically refers to the combining of categories of classifying variables but it also includes releasing only marginal and conditional tables corresponding to subsets of the cross-classifying variables (Fienberg and Slavković, 2008).

Recently, there has been a growing demand for flexible on-line table generating servers (Thompson, Broadfoot and Elazar, 2013; Shlomo, Antal and Elliot, 2015). Typically such systems provide a menu-driven interface for producing confidentiality-protected user-defined frequency tables of counts or quantiles. Instead of using the two-step confidentiality protection routine mentioned above, such systems may add a random perturbation amount to each non-zero cell of the table, not just to a subset of the cells.

In the differential privacy framework, a mechanism $\mathcal{M}(\cdot)$ operating on datasets is required to be stochastic, and it is this stochasticity that provides the confidentiality protection, as we shall explain. From the utility perspective, a common assumption is that statistical analysis will generally be conducted on $\mathcal{M}(D)$ as if it were D itself, and so utility is often measured in terms of some kind of discrepancy measure between D and $\mathcal{M}(D)$ (Wasserman and Zhou, 2010). Such measures include the information-theoretic Hellinger's distance, and the more intuitive average absolute difference per cell (Gomatam and Karr, 2003; Shlomo, 2007).

It is a property of differential privacy that the confidentiality protection guarantee does not rely on hiding the parameters of the perturbation. This fact is reminiscent of Kerckhoffs' principle in cryptography, that *a cryptosystem should be secure even if everything about the system, except the key, is public knowledge* (Auguste, 1883) and Shannon's maxim in information theory, that *one ought to design systems under the assumption that the enemy will immediately gain full familiarity with them* (Shannon, 1949). As a consequence, in contrast to common practice in some government agencies, in the differential privacy framework the full description of the mechanism $\mathcal{M}$ can be made available along with $\mathcal{M}(D)$. This would include all details on the distributions of perturbations, but would, of course, exclude their actual randomly drawn values, which would not be revealed. The advantage of this practice is that knowledge of the mechanism allows the user to take the perturbation into account in their analysis, thereby avoiding potentially misleading conclusions that might arise from ignoring the perturbation.

Methods for correcting for perturbation have been considered for microdata on both continuous and categorical variables (Fuller, 1993; van den Hout and van der Heijden, 2002) but do not appear to have been considered for the dissemination of frequency tables. A basic general idea is that the likelihood for a parametric model for D may be naturally extended, in principle, to the likelihood for $\mathcal{M}(D)$ and so valid likelihood-based inference could be conducted (Karwa, Kifer and Slavković, 2015). This idea will be illustrated in Section 6.

The differential privacy literature distinguishes between what are called interactive and non-interactive data dissemination settings. In the *interactive* setting, the data custodian agency provides a system interface, typically on-line, through which users may pose a series of queries say $f_1, f_2, \ldots$ about a dataset D and receive a series of confidentiality-protected responses $\mathcal{M}_1(f_1(D))$, $\mathcal{M}_2(f_2(D)), \ldots$. The system monitors the queries, and decides based on the outputs already released, whether to stop dissemination altogether, whether to answer the particular query, and if so then the amount of perturbation to be applied. In the *non-interactive* setting, for a dataset D, the whole data set is perturbed off-line to produce a confidentiality-protected dataset $\mathcal{M}(D)$. The protected dataset can be released as a whole, or functions of it are provided as responses to queries that can be answered with $\mathcal{M}(D)$. If only parts of the data are requested then it may be possible and efficient for the agency to perturb only those parts. In this paper, we consider only the non-interactive setting, which is closer to the model table generating systems of interest to us. If the frequency table data D is treated simply as a set of frequency counts in disjoint cells then this is analogous to a histogram with disjoint bins and is a core field of application of differential privacy methodology (Dwork et al., 2006; Dwork and Roth, 2014; Wasserman and Zhou, 2010). Barak et al. (2007) extended this core methodology to handle the case where D also includes table margins, consisting of sums of cell counts, and where perturbed margins are released which are arithmetically consistent with the perturbed cell counts. Fienberg, Rinaldo and Yang (2010) explore this approach in the context of a number of examples. They express doubt about the suitability of this methodology for the type of large sparse tables often produced by statistics agencies.

The rest of the paper is structured as follows. Section 2 presents some features of perturbations for a table generating server, which bear some resemblance to those recommended by the ABS TableBuilder system, with an example table presented in Section 3. Section 4 introduces some aspects of differential privacy theory for the dissemination of frequency tables. In Section 5 we define and compare different perturbation mechanisms and present some results illustrating the trade-off between disclosure risk and data utility on the example table from Section 3 and other simulated tables. In Section 6 we demonstrate how to carry out correct statistical inference when the perturbation mechanism is known to the analyst. In Section 7 we address the issue of overlapping cells and marginal counts in frequency tables and conclude with Section 8.

2. PERTURBATION OF FREQUENCY TABLES

As a starting point for this exploration of perturbation mechanisms and differential privacy, we have chosen to focus on the problem of dissemination of frequency tables. We suppose in this paper that such tables contain population counts, from a census or administrative sources. Government agencies also produce tables of estimated population counts based on sample survey data, where an estimated cell count is typically the sum of survey weights across the sample units in the cell. There are somewhat different considerations in the potential application of differential privacy ideas to such survey-based tables and we shall only return to comment on this possible extension in the final section of the paper.

Frequency tables are important data products in government statistical settings, and recently various dissemination schemes in addition to the publication of pre-specified collections of confidentiality-protected tables have appeared. One flexible on-line table generating system is the ABS TableBuilder (Chipperfield, Gow and Loong, 2016; Fraser and Wooton, 2005; Thompson, Broadfoot and Elazar, 2013). This system has attracted interest from other agencies in the context of the protection of census outputs (Andersson, Jansson and Kraft, 2015; Jansson, 2012; Longhurst et al., 2007). While we refer to the requirements and objectives of the TableBuilder system to motivate our assumptions, we do not attempt to replicate its properties exactly nor do we seek to replicate its confidentiality protection methods.

2.1 Some Terminology and Notation

In this section we introduce some terminology and notation. First, we make a remark about our use of the terms confidentiality and privacy. This paper deals with the confidentiality of data held by a national statistical agency or other custodian, as in the statistical disclosure control literature, and we use the term confidentiality in that context. In the computer science literature the term differential privacy is used to mean a particular way of defining a standard of confidentiality protection, and the term privacy is used in association with that. To be consistent with that literature, we will use the term privacy in the context of the differential privacy theory.

Consider a data set in the form of a frequency table or a set of tables, where each cell is defined by values of a given fixed set of attributes. We assume that the given data set belongs to a universe of possible data sets that could have been realised, denoted by $\mathcal{U}$. An agency may decide to release the data or parts of it. The collection of all frequencies that could be released is arranged in a *list* $\mathbf{a} = (a_1, \ldots, a_K)$ consisting of K cells in some order, where a_k denote the frequency in cell k, that is, the number of individuals taking the attribute values corresponding to the cell, for $k = 1, \ldots, K$. The list $\mathbf{a}$ will be released after undergoing a perturbation in order to preserve confidentiality. If, for example, the data consists of a 10-way table, the list may include all interior cells, and also some marginal tables, or only some marginal tables. Marginal tables are computed by aggregating interior cells, and we shall see later why both marginals and interior cells may be included in the list. It is thus possible that different cells in a list might refer to overlapping subsets of individuals, that is, some individuals may appear in more than one cell, and that different cell frequencies might correspond to the same set of individuals. A typical situation is that an agency holds a 10-way table, say, but will release only perturbed versions of 3-way marginals, and the cells of these marginals (unperturbed) will comprise the list.

The set A of possible lists $\mathbf{a} = (a_1, \ldots, a_K)$ is called the *list-space*. We shall suppose that all elements of lists in A are non-negative integers. The list-space is determined by the agency's decision on which parts of the data are to be released, thus determining the structure of $\mathbf{a}$, and on the universe $\mathcal{U}$ of potential data sets that could have been realised, that is, A is the set of lists with a given structure that could arise from all data sets in $\mathcal{U}$.

We consider a mechanism $\mathcal{M}(\cdot)$ on a list-space A that replaces the list $\mathbf{a} = (a_1, \ldots, a_K)$ by the perturbed list to be published $\mathcal{M}(\mathbf{a}) = \mathbf{b} = (b_1, \ldots, b_K)$ containing perturbed frequencies b_k. In this paper we consider mechanisms that are random functions. The mechanism can be represented by a conditional probability distribution, denoted $p(a, b)$, where for cell $k = 1, \ldots, K$, the perturbed cell frequency b_k takes a value b with probability $p(a, b)$, when $a_k = a$. Thus $p(a, b)$ defines the conditional probability distribution of b_k given a_k and we assume that the same distribution applies to all cells k with a given value a of a_k. In general we shall assume that different cells are perturbed independently.

2.2 Some Properties of the ABS TableBuilder

The ABS TableBuilder, which we use as a model for table generating servers, has been evolving and its description varies in different papers. Chipperfield, Gow and Loong (2016) describe a list as above, and in principle all perturbations could be applied in advance, however for efficiency's sake they are applied when users submit queries, using a lookup table whose random values are drawn in advance. According to Fraser and Wooton (2005) different cells are perturbed independently, unless the cell counts are associated with the same underlying set of individuals. If two cell counts do in fact correspond to the same group of individuals, then the ABS TableBuilder requires that the perturbed value is also the same. In this method, this 'same-participants-same-perturbation' property is implemented in a straightforward manner by attaching a random key drawn from some continuous distribution to each individual in the population underlying the data, and a cell's key being the sum of the keys of its members. This cell key is used as a seed for the random perturbation mechanism and two cells based on the same group of individuals will be perturbed by the same seed to the same value.

The 'same-participants-same-perturbation' property is aimed at preventing repeated queries on the same group with independent perturbations, which can be averaged to reduce the noise and thus leak information. However, as we shall see, the 'same-participants-same-perturbation' property will have to be abandoned if differential privacy is adopted. We explain it here informally by demonstrating a scenario of confidentiality breach that results from this principle. As often happens, the scenario below may seem contrived, but it can be made to seem more realistic easily. Suppose our data D is about a given group, say workers in a factory, and an intruder wishes to obtain information about the salary of a particular person, say Bob, the only worker hired today. Suppose the following two queries are allowed: 1, the frequency of workers whose salary exceeds s, and 2, the frequency of workers whose salary exceeds s, and who have been working for more than one day. Suppose the responses (with perturbation) to the two queries are different. Under the 'same-participants-same-perturbation' principle Bob's salary must exceed s, and thus new information was obtained due to Bob's participation in D. Once differential privacy will be defined, it is an easy task to translate this scenario to a breach of differential privacy. Note that in the above scenario we obtained the information only because the two groups defined by 1 and 2 above could have been the same (which was not the case in the above realisation). This shows that the universe $\mathcal{U}$ must be taken into account, and not just identical groups in the realised data or list.

This breach can be avoided if two queries with different descriptions as shown in 1 and 2 above are perturbed independently, and the principle is modified to 'same-participants and description-same-perturbation'. A similar scenario appears in Chipperfield, Gow and Loong (2016), leading them to the above modification of the principle. However this modification opens the possibility of submitting queries for the same group in different ways, and averaging to cancel the perturbation noise. It may perhaps be possible to circumvent the whole problem, and in particular such an averaging attack, by setting rules on the structure of the list $\mathbf{a}$ and queries' formulations which prevent the possibility of referring to the same group in different ways. An example of such a rule is a restriction on the structure with respect to sparsity, e.g., the number of zeros (and sometimes also ones and twos) that may cause a margin to equal an internal cell.

Some additional properties of a protection method for a frequency table dissemination server that are similar to those of the ABS TableBuilder are set out below. The first three properties address disclosure risk concerns, via either removing concerns with small cells, such as counts of one, and setting a criterion to minimize risk for given utility. The remaining five properties address utility concerns, via being broadly concerned with either preserving important features of the original table or reducing differences between the original and perturbed tables.

1. The perturbation does not produce values below a specified threshold, that is $p(a_k, b_k) = 0$ if $b_k \leq c$ for a specified value $c > 0$, for any value of a_k.
2. The distribution of b_k given a_k has maximal entropy subject to constraints on the range and variance of the perturbation.
3. Sparse tables according to given thresholds are not published.
4. The perturbed frequencies are non-negative integers, that is, $b_k \geq 0$.
5. Structural zeros, that is, counts of attribute combinations that are impossible to observe in the population, are not perturbed.
6. The perturbations are unbiased, that is, the expected value of b_k given a_k equals a_k.
7. The variance of b_k given a_k is constrained not to exceed a given value.
8. The distribution of b_k given a_k is *truncated* by imposing a bound on $|b_k - a_k|$, the absolute difference between the perturbed and original values.

We remark that these properties are not all consistent, for example, properties 4 and 6 are generally contradictory. As discussed later, some of these properties, such as 1, 2 and 4 above, may not be advantageous under the differential privacy framework. They may well be justifiable if other risk measures are considered.

Table 1
Typical user-specified sub-table of a larger frequency table (interior cells only) for NUTS2 Region = 1 and Country of birth = rest of Europe. The variables of interest are Age in banded 5-year groups from 15 to 74, and Occupation classified as one of A,...,K.

Age group	Occupation										
	A	B	C	D	E	F	G	H	I	J	K
15-19	2	2	8	7	31	0	7	2	20	0	80
20-24	55	68	110	54	134	0	23	13	138	2	129
25-29	115	147	132	78	83	0	19	15	45	0	18
30-34	191	129	127	89	68	0	18	8	33	4	10
35-39	153	113	119	74	49	1	34	15	44	4	9
40-44	102	70	78	70	43	1	20	21	24	3	8
45-49	94	65	55	72	47	2	29	16	36	4	14
50-54	92	81	75	80	65	1	43	17	36	1	8
55-59	74	51	56	64	72	2	49	21	67	2	13
60-64	63	41	40	70	53	3	22	22	56	4	59
65-69	12	5	7	3	12	0	6	4	8	2	287
70-74	4	4	1	5	4	0	2	1	4	0	307

3. EXAMPLE OF FREQUENCY TABLE

In order to provide a realistic example, we selected the following variables used in data from the 2001 census in the United Kingdom (UK):

- NUTS2 Region - 11 regions
- Gender - 2 categories
- Age in banded 5 year age groups - 21 categories
- Current Employment Status - 5 categories
- Occupation - 12 categories
- Educational attainment - 9 categories
- Country of birth - 5 categories

Here the NUTS (Nomenclature of Territorial Units for Statistics) is a hierarchical system for dividing up the economic territory of the European Union and NUTS2 comprises basic regions for the application of regional policies, defined for the purpose of socio-economic analyses. We generated a 7-way frequency table by multiplying each of the UK 2001 census proportions by $N = 1,500,000$, to obtain a table that mimics a real population of size N.

In Table 1 we present a realistic example of a sub-table of the 7-way frequency table that might be requested by a user. The sub-table is defined by fixing NUTS2 Region = 1 and Country of birth = rest of Europe, and requesting a 2-way frequency table of counts for occupation and age groups from 15 to 74.

Table 1 has some small cells, that normally have high associated disclosure risks. We will use this table (in addition to some simulated tables) later , in order to illustrate the implementation of our confidentiality protection approach.

4. DIFFERENTIAL PRIVACY FOR FREQUENCY TABLES

4.1 Basic Ideas and Definitions

As indicated in Section 1, privacy loss occurs when an intruder can learn from the perturbed list $\mathcal{M}(\mathbf{a})$ about an individual contributing to the original list $\mathbf{a}$. We consider a randomized mechanism $\mathcal{M}(\mathbf{a})$ that produces a random value $\mathbf{b}$, the perturbed value of $\mathbf{a}$, with probability $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})$ depending only on the mechanism $\mathcal{M}$. We denote the range of the perturbation of $\mathbf{a} \in A$ by $B(\mathbf{a})$, that is, $B(\mathbf{a}) = \{\mathbf{b} : \mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) > 0\}$. Then $B(\mathbf{a}) \subseteq B$, the range of $\mathcal{M}$, and when $B(\mathbf{a})$ does not depend on $\mathbf{a}$, we have $B(\mathbf{a}) = B$. Sometimes $A = B$ is assumed. For lists $\mathbf{a}$, $\mathbf{a}'$, we write $\mathbf{a} \sim \mathbf{a}'$ and refer to $\mathbf{a}$ and $\mathbf{a}'$ as *neighbours*, if $\mathbf{a}'$ can be obtained from $\mathbf{a}$ by adding or removing exactly one individual.

As explained below, we may measure how much can be learnt about individuals by the likelihood ratios $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})/\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b})$ for $\mathbf{a} \sim \mathbf{a}'$. It is the ratio of the intruder's likelihoods for observed $\mathbf{b}$ under $\mathbf{a}$ or $\mathbf{a}'$ considered as parameters. (Recall that we assume that the nature of $\mathcal{M}(\cdot)$ is released together with $\mathcal{M}(\mathbf{a})$.) The likelihood ratio could alternatively be viewed as a posterior odds ratio, or Bayes factor, from a Bayesian perspective.

Placing a bound on this likelihood ratio motivates the definition of *ε-differential privacy*, which we denote by DP(ε). We specialise the definition to lists as follows.

DEFINITION 1. *(Dwork et al., 2006) A mechanism $\mathcal{M}$ satisfies* ε-differential privacy *if for all neighbouring lists* $\mathbf{a}, \mathbf{a}'$ *in A, and all subsets $S \subseteq Range(\mathcal{M}) = B$, we have:*

$$\mathbb{P}(\mathcal{M}(\mathbf{a}) \in S) \leq e^{\varepsilon}\mathbb{P}(\mathcal{M}(\mathbf{a}') \in S). \tag{4.1}$$

Since in our setting $Range(\mathcal{M})$ is discrete, we can use the simpler condition that $\mathcal{M}$ satisfies *ε-differential privacy* if for all neighbouring lists $\mathbf{a}, \mathbf{a}'$, and all lists $\mathbf{b}$ we have:

$$\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) \leq e^{\varepsilon}\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b}). \tag{4.2}$$

As the neighbourhood relation is symmetric, we can equivalently say that the mechanism $\mathcal{M}$ satisfies ε-differential privacy if for all perturbed lists $\mathbf{b}$ and neighbouring $\mathbf{a}$ and $\mathbf{a}'$

$$e^{-\varepsilon} \leq \mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})\big/\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b}) \leq e^{\varepsilon}. \tag{4.3}$$

If there is a very large or small value of the ratio $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})\big/\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b})$ for given $\mathbf{a} \sim \mathbf{a}'$ and some observed $\mathbf{b}$, then a typical scenario for a confidentiality breach is the following: suppose an intruder knows the whole original unperturbed list apart from the cell where one targeted individual belongs. Suppose the intruder wants to know whether the targeted individual is in the data set D, and if so, in which cell. Denoting the list without the target by $\mathbf{a}'$, say, the intruder computes the ratio for all $\mathbf{a}$ where the target is added into one cell. Under DP(ε) with a small ε all these ratios will be close to 1, and inference on whether the target is in D is impossible. But otherwise, a large likelihood ratio will suggests the inference that the targeted individual is in D and in which cell, according to the $\mathbf{a}$ that yielded that high likelihood ratio. One can describe such a scenario in a way that does not require the intruder to know too much. However, some prior information is needed.

Note that 'all S' in the DP(ε) definition refers to all possible subsets S of B. Thus, the definition does not only refer to the realised list $\mathbf{a}$ and its neighbours and the outcome $\mathbf{b}$ observed by the intruder but rather to all potential lists in A, and all possible outcomes of the perturbation. In this sense DP(ε) can be viewed as a 'worst case' requirement, and the definition refers to the mechanism and not to the perturbed data, and is applicable at the stage of designing the mechanism before the perturbation has taken place.

As we shall discuss, a key challenge with the differential privacy requirement is the possible effect on utility. We introduce two relaxations of differential privacy that seek to reduce confidentiality protection in a controlled way, in order to gain utility. Both of these relaxations will be used later in the paper.

The most widely known relaxation of the definition of differential privacy for $\mathcal{M}$, which may result in enhanced utility, is (ε, δ)-*differential privacy*, or $\mathrm{DP}(\varepsilon, \delta)$ (Dwork and Roth, 2014, Definition 2.4), under which

$$\mathbb{P}(\mathcal{M}(\mathbf{a}) \in S) \leq e^{\varepsilon}\mathbb{P}(\mathcal{M}(\mathbf{a}') \in S) + \delta \tag{4.4}$$

for all subsets S of the range of $\mathcal{M}$ and neighbouring $\mathbf{a}$ and $\mathbf{a}'$. The parameter δ adds flexibility by allowing the randomly perturbed list to have a probability of δ of having an undesirable likelihood ratio with associated higher disclosure risk.

An alternative relaxation of $\mathrm{DP}(\varepsilon)$ requires the likelihood ratio to be bounded by e^{ε}, as in (4.1), across a set of possible outcomes with probability at least $1-\delta$. As a definition, (ε, δ)-*probabilistic differential privacy* is satisfied if $\mathbb{P}\big(\mathcal{M}(\mathbf{a}) \in G(\mathbf{a}, \mathbf{a}')\big) > 1 - \delta$ for all $\mathbf{a} \sim \mathbf{a}' \in A$, where

$$G = G(\mathbf{a}, \mathbf{a}') = \{\mathbf{b} \in B(\mathbf{a}) : \mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})/\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b}) \leq e^{\varepsilon}\}, \tag{4.5}$$

and 0/0=0. Closely related definitions can be found in Gotz et al. (2012); Machanavajjhala et al. (2008).

Lemma 1. *(Gotz et al., 2012) If a mechanism $\mathcal{M}$ satisfies (ε, δ)-probabilistic differential privacy then it also satisfies $DP(\varepsilon, \delta)$.*

Proof. Suppose $\mathcal{M}$ satisfies (ε, δ)-probabilistic differential privacy, and let C denote the complement of G in $B(\mathbf{a})$. For a subset S of the range of $\mathcal{M}$ and for neighbouring lists $\mathbf{a} \sim \mathbf{a}'$, we have:

$$\begin{aligned}
\mathbb{P}(\mathcal{M}(\mathbf{a}) \in S) &= \sum_{\mathbf{b} \in S \cap G} \mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) + \sum_{\mathbf{b} \in S \cap C} \mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) \\
&\leq \sum_{\mathbf{b} \in S \cap G} e^{\varepsilon}\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b}) + \sum_{\mathbf{b} \in S \cap C} \mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) \\
&\leq e^{\varepsilon}\mathbb{P}(\mathcal{M}(\mathbf{a}') \in S) + \delta,
\end{aligned}$$

where the first inequality follows from the definition of the set G and the second from the definition of (ε, δ)-probabilistic differential privacy. □

Recall that two lists $\mathbf{a}$ and $\mathbf{a}'$ in A are neighbours if $\mathbf{a}'$ can be obtained from $\mathbf{a}$ by adding or removing a single individual. Given a list-space A, let d denote the maximum number of cells in which two neighbours, $\mathbf{a}$ and $\mathbf{a}'$ can differ. If each individual appears only in a single cell, then $d = 1$, as one cell frequency decreases by one when an individual is removed from the cell, and increases by one if an individual is added to the cell. The number d will play a role in utility computation, see Section 7, with a larger d leading to smaller utility. Other than in Section 7 we assume throughout that $d = 1$, which occurs, for example, if the data to be released consist of the interior cells of a standard frequency table.

Given two original, unperturbed lists that differ in a single individual, the differential privacy condition involves quantification of the difference between the distributions of the corresponding released perturbed lists. For small ε and δ, $\mathrm{DP}(\varepsilon, \delta)$ means that an individual's participation in the original data set is likely to have a small influence on the released data. An individual considering participation in a census, for example, is assured that by doing so the risk of a confidentiality breach rises only in a limited way: to be precise, only with (small) probability δ, the presence of the individual in a given cell may be inferred by a likelihood ratio that exceeds e^{ε}. Moreover, since the very presence of an individual in a data set is unlikely to be inferred, participation in any past or future data set is unlikely to increase the individual's risk. In other words, the data environment in which the perturbed data set is released is irrelevant to the confidentiality guarantees under differentially privacy release with small parameters. On the other hand, if an intruder can learn certain attributes of an individual with high probability, he can later try to use these attributes to find the individual in other data sets and obtain further information about them. In this case the environment may matter, and if individuals in the data set appear in other data sets, past or future, the risk may increase. If the differential privacy parameters of different perturbation schemes are not small, and they are used for comparing confidentiality protection in different data sets, one has to take the environments into account, and compare only files which have similar environments. In this paper we focus on using the parameters of differential privacy to compare different perturbation mechanisms operating on the same file, thus avoiding this additional issue.

In the differential privacy literature it is stated that δ should be smaller than $1/N$ where N is the total number of individuals in the protected data (Dwork and Roth, 2014). The reason is that if $\delta = 1/N$ then a mechanism that chooses one individual at random and just releases her data without any perturbation, would satisfy $\mathrm{DP}(\varepsilon, \delta)$ for any ε. Releasing the data of a single individual is indeed inappropriate, but a realistic perturbation algorithm, even with $\delta > 1/N$, would not really do that. Indeed, $\delta > 1/N$ means that the probability that the likelihood ratio of (4.3) will be outside the defined desirable interval is larger than $1/N$. If this happens then testing whether the data set in question is $\mathbf{a}$ or a neighbouring $\mathbf{a}'$ may have a higher power than we would like, but that does not necessarily amount to releasing the unperturbed data of some individual.

4.2 Utility/loss Functions and the Exponential Mechanism

As mentioned above, differential privacy is defined as a property of a mechanism. Various candidates for differentially private mechanisms $\mathcal{M}(\cdot)$ have been proposed in the literature, see for example Dwork and Roth (2014). We shall consider some alternative choices that might be suitable for implementation in table-generating servers, specifically those that are cases of the general 'exponential mechanism' (McSherry and Talwar, 2007). Informally, the exponential mechanism is defined with respect to some utility function u which assigns a utility score to possible perturbed values so that the mechanism is more likely to produce values with higher utility scores (see Dwork and Roth, 2014).

The exponential mechanism includes the class of perturbation mechanisms which we shall apply in different versions in the remainder of this paper. The approach starts by specifying a utility function $u(\mathbf{a}, \mathbf{b})$, measuring the utility of the perturbed list $\mathbf{b}$ given the original list $\mathbf{a}$. Following (Dwork and Roth, 2014), we shall generally consider additive utility functions of the form $u(\mathbf{a}, \mathbf{b}) = \sum_{k=1}^{K} v(a_k, b_k)$. As we shall see, this additive form enables us to specify a mechanism which ensures that the K cells in the list are perturbed independently. Statisticians are familiar with loss functions, so we start with examples of those, and then transform them to utilities by a sign change. The loss functions we shall use are:

$$\begin{aligned} \ell_1 &= \ell_1(\mathbf{a}, \mathbf{b}) = \sum_{k=1}^{K} |a_k - b_k| \\ \ell_2 &= \ell_2(\mathbf{a}, \mathbf{b}) = \sum_{k=1}^{K} (a_k - b_k)^2, \\ \ell_3 &= \ell_3(\mathbf{a}, \mathbf{b}) = \sum_{k=1}^{K} |\sqrt{a_k} - \sqrt{b_k}|. \end{aligned}$$

The utility functions considered in this paper are $u_i = -\ell_i$ for $i = 1, 2, 3$.

As loss functions, ℓ_1 and ℓ_2 are natural and standard. The loss ℓ_3 is reminiscent of Hellinger distance. It has the intuitively appealing property that the loss varies with the size of the perturbed cell: for example, the same loss of 2 is incurred by perturbing 0 to 4, 100 to 144 and 10000 to 10404. This is in contrast to ℓ_1 for which the perturbation from 10000 to 10404 has a much higher loss than perturbing 100 to 144 or 0 to 4. Although as a loss function ℓ_3 seems very reasonable, and we use it to demonstrate some points, we shall see that it does not turn out to be very useful in practice when using the exponential mechanism for protecting frequency tables. Note that the Hellinger distance, $(\sum_{k=1}^{K} (\sqrt{a_k} - \sqrt{b_k})^2)^{1/2}$, proposed as a loss function in Shlomo (2007) is not of an additive form.

To describe the exponential mechanism, consider mechanisms where the range of $\mathbf{b}$, denoted by B as before, does not depend on $\mathbf{a}$, that is, every $\mathbf{b} \in B$ satisfies $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) > 0$ for all $\mathbf{a}$. This assumption will be modified later. The *exponential mechanism* is defined by

$$\text{given } \mathbf{a} \text{ choose } \mathbf{b} \in B \text{ with probability proportional to } e^{\eta u(\mathbf{a}, \mathbf{b})/\Delta u}, \tag{4.6}$$

where η is a specified value, depending on the differential privacy parameter ε, and the scale factor Δu is

$$\Delta u = \max_{\mathbf{b} \in B} \max_{\mathbf{a} \sim \mathbf{a}' \in A} |u(\mathbf{a}, \mathbf{b}) - u(\mathbf{a}', \mathbf{b})|. \tag{4.7}$$

It is easy to see that this mechanism attaches higher probability to perturbed lists which have higher utility. For any additive utility function of the form $u(\mathbf{a}, \mathbf{b}) = \sum_{k=1}^{K} v(a_k, b_k)$, the K cells in the list are perturbed independently and the probability that list $\mathbf{a}$ is perturbed to $\mathbf{b}$ is

$$P(\mathbf{a}, \mathbf{b}) = \prod_{k=1}^{K} p(a_k, b_k) \propto \prod_{k=1}^{K} e^{\eta v(a_k, b_k)/\Delta u},$$

where $p(a_k, b_k)$ is the probability of a cell of size a_k being perturbed to b_k. Independent perturbations are simple to apply and to analyse, however in theory they may lead to an undesirable result. For example, suppose one cell represents a subset of the set of individuals contributing to another cell. Under independent perturbations, the perturbed count of the cell with the subset may be larger than the perturbed count of the original cell. As another example, if one cell in the list to be perturbed consists of a marginal count, that is, the sum of some other cells, then this additive relationship need not hold after independent perturbations have been applied.

A key property of the exponential mechanism is that DP(ε) holds for a suitable η depending on ε in a simple way. The following result is Theorem 3.10 in Dwork and Roth (2014), where a proof is given. In fact, the result is a special case of our Theorem 4.2 below. We mention again that in Theorem 4.1 we assume that the range of $\mathcal{M}(\mathbf{a})$, denoted by B, does not depend on $\mathbf{a}$. This result shows that we obtain DP(ε) by choosing $\eta = \varepsilon/2$.

Theorem 4.1. *Let u be a utility function and $\mathcal{M}$ a perturbation mechanism such that $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})$ is proportional to $e^{\varepsilon u(\mathbf{a},\mathbf{b})/2\Delta u}$ for all possible lists $\mathbf{a} \in A$ and perturbed lists $\mathbf{b} \in B$. Then $\mathcal{M}$ is* DP(ε).

Proof. For $\mathbf{a}, \mathbf{a}' \in A$ and $\mathbf{b} \in B$ we have

$$\begin{aligned}\frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b})} &= \left\{\frac{e^{\varepsilon u(\mathbf{a},\mathbf{b})/2\Delta u}}{\sum_{\mathbf{b}\in B} e^{\varepsilon u(\mathbf{a},\mathbf{b})/2\Delta u}}\right\} \Big/ \left\{\frac{e^{\varepsilon u(\mathbf{a}',\mathbf{b})/2\Delta u}}{\sum_{\mathbf{b}\in B} e^{\varepsilon u(\mathbf{a}',\mathbf{b})/2\Delta u}}\right\} \\ &= \left\{\frac{e^{\varepsilon u(\mathbf{a},\mathbf{b})/2\Delta u}}{e^{\varepsilon u(\mathbf{a}',\mathbf{b})/2\Delta u}}\right\}\left\{\frac{\sum_{\mathbf{b}\in B} e^{\varepsilon u(\mathbf{a}',\mathbf{b})/2\Delta u}}{\sum_{\mathbf{b}\in B} e^{\varepsilon u(\mathbf{a},\mathbf{b})/2\Delta u}}\right\} \le e^{\varepsilon}.\end{aligned}$$

Using $|u(\mathbf{a},\mathbf{b}) - u(\mathbf{a}',\mathbf{b})| \le \Delta u$, it is easy to see that each of the two terms in the latter product is bounded by $e^{\varepsilon/2}$, and the result follows. □

4.3 Truncated Cell Perturbations

Recall from Section 2.2 that it can be desirable in terms of increased utility to truncate cell perturbations by bounding the perturbation according to $|a_k - b_k| \le m$ for some m, for all k. In particular, the range of $\mathcal{M}(\mathbf{a})$, denoted by $B(\mathbf{a})$, will depend on $\mathbf{a}$. Theorem 4.2, a variant of Theorem 4.1, demonstrates that the increased utility provided by the truncation is achieved at the cost of relaxing DP(ε) to DP(ε, δ) with $\delta > 0$ depending on the truncation bound m and the utility function u. Note that in contrast to Theorem 4.1, in Theorem 4.2 the exponent is not divided by 2 ($\eta = \varepsilon$ rather than $\varepsilon/2$ above), which implies a smaller spread of the perturbation in addition to the truncation by m. Consistent with these adjustments, the definition (4.7) is replaced by

$$\Delta u = \Delta u(\mathbf{a}) = \max_{\mathbf{b}\in B(\mathbf{a}')} \max_{\mathbf{a}\sim\mathbf{a}'\in A} |u(\mathbf{a},\mathbf{b}) - u(\mathbf{a}',\mathbf{b})|. \tag{4.8}$$

Note that (4.7) is a special case of (4.8) where for all $\mathbf{a}$ we have $B(\mathbf{a}) = B$.

Theorem 4.2. *Let u be a utility function of the form $u(\mathbf{a},\mathbf{b}) = g(\mathbf{a} - \mathbf{b})$ for some g, and $\mathcal{M}$ a perturbation mechanism such that $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})$ is proportional to $e^{\varepsilon u(\mathbf{a},\mathbf{b})/\Delta u}$ for all possible lists $\mathbf{a} \in A$ and perturbed lists $\mathbf{b}$ such that $|a_k - b_k| \le m$ for all k, and otherwise $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) = 0$, and Δu is given in* (4.8)*. Assume also that for all $\mathbf{a} \sim \mathbf{a}'$, $\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b}) = 0$ implies $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) < \delta$. Then $\mathcal{M}$ is DP(ε, δ).*

Proof. Let $\mathbf{a} \sim \mathbf{a}'$ be neighbouring lists and let $\mathbf{b} \in \mathit{Range}(\mathcal{M})$. Clearly, we can assume $\mathbf{b} \in B(\mathbf{a})$ as otherwise $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) = 0$ and (4.9) holds trivially. If $\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b}) = 0$ then $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) < \delta$ so that $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) \leq e^{\varepsilon}\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b}) + \delta$ as required. If $\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b}) > 0$ then

$$\begin{aligned} \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b})} &= \left\{ \frac{e^{\varepsilon u(\mathbf{a},\mathbf{b})/\Delta u}}{\sum_{\mathbf{b}} e^{\varepsilon u(\mathbf{a},\mathbf{b})/\Delta u}} \right\} \Big/ \left\{ \frac{e^{\varepsilon u(\mathbf{a}',\mathbf{b})/\Delta u}}{\sum_{\mathbf{b}} e^{\varepsilon u(\mathbf{a}',\mathbf{b})/\Delta u}} \right\} \\ &= \frac{e^{\varepsilon u(\mathbf{a},\mathbf{b})/\Delta u}}{e^{\varepsilon u(\mathbf{a}',\mathbf{b})/\Delta u}} \leq e^{\varepsilon}, \end{aligned} \tag{4.9}$$

where the second equality follows from the fact that the two sums in the denominators cancel since $\sum_{b:|b-a|\leq m} e^{cg(b-a)} = \sum_{z=-m}^{m} e^{cg(z)}$ does not depend on a, and the last inequality follows from $|u(\mathbf{a}, \mathbf{b}) - u(\mathbf{a}', \mathbf{b})| \leq \Delta u$. Thus $\mathcal{M}(\mathbf{a}) = \mathbf{b} \in G(\mathbf{a}, \mathbf{a}')$, where $G(\mathbf{a}, \mathbf{a}')$ is defined in (4.5). It follows that $\mathbb{P}(\mathcal{M}(\mathbf{a}) \in G(\mathbf{a}, \mathbf{a}')) > 1 - \delta$, and the result follows from Lemma 1. □

We now demonstrate the calculation of the value δ when applying Theorem 4.2. Suppose we wish to impose a bound m on $|b - a|$, the difference between the perturbed and original value. In other words, we assume $p(a, b) = 0$ for $|b - a| > m$. Here and in all our applications we assume also that $p(a, b) > 0$ for $|b - a| \leq m$. For neighbouring $\mathbf{a}, \mathbf{a}'$, $\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b}) = 0$ and $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) > 0$ occurs only when the value in a particular cell, say j, of $\mathbf{a}$ is $a + 1$ and that of $\mathbf{a}'$ is a, and all other cells of $\mathbf{a}, \mathbf{a}'$ are equal. We have $p(a + 1, a + 1 + m) > 0$ and $p(a, a + 1 + m) = 0$ and therefore, if cell j of $\mathbf{b}$ has the value $a + 1 + m$ then $\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b}) = 0$. With a similar argument for $p(a, a - m)$, we claim that the exponential mechanism of Theorem 4.2 is $\mathrm{DP}(\varepsilon, \delta)$, with

$$\delta = \max\{\max_a p(a + 1, a + 1 + m), \max_a p(a, a - m)\}. \tag{4.10}$$

In fact, in the above case, if $\mathbf{a}, \mathbf{a}'$ differ as above in cell j, and $b_j = a_j + m + 1$, then

$$P(\mathbf{a}, \mathbf{b}) \leq \delta \prod_{k \neq j} p(a_k, b_k) \leq \delta. \tag{4.11}$$

Thus for any such $\mathbf{b}$ we have $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) < \delta$ as required in the theorem. Note that there may be a considerable slack in the second inequality of (4.11), implying that the δ parameter in the differential privacy could be much better, that is, smaller than stated.

4.4 Post-Processing and Negative Perturbed Values

In general, agencies will be reluctant to disseminate perturbed tables with negative frequencies. However, as our brief discussion below shows, this policy should be reconsidered if differential privacy is to be adopted. Our proofs of DP allow negative values, but as we shall see, the same DP level continues to hold if all negative values are replaced by zeros. We show below that negative values may be useful and informative in various ways and that information may be lost by replacing negative values with zero.

If the perturbations are unbounded, as in Theorem 4.1, then $\mathcal{M}(\mathbf{a})$ may have negative cells for any $\mathbf{a}$ depending on the utility u. This is the case for u_1 and

u_2. If the perturbations are truncated by m as in Theorem 4.2, then cells with $a < m$ may be perturbed to a negative b. Negative values are required to achieve unbiasedness of the perturbed data. In the exponential mechanism with u_1 or u_2, which are our main examples, the perturbing distribution is symmetric and negative values of the perturbed data may occur. Unbiasedness is clearly desirable on its own, and when computing marginals as sums of perturbed interior cells, unbiasedness implies that the perturbation would cancel rather than accumulate. Therefore, it seems reasonable to allow release of negative values, and advise users to replace them by zeros at a suitable stage of their analysis, e.g., after computing marginals or merged cells from interior cells.

However, if publishing data with negative perturbed frequencies is not acceptable for some reason, the data releasing agency can just report all negative values as zeros. This will effectively replace the perturbed value b by a value closer to the original a since $a \geq 0$. More generally, if for some reason an agency wishes the released entries of the list to satisfy some constraints such as $b \geq c$ for some c, it can replace all smaller values by c. Such *post-processing* preserves differential privacy, see Proposition 2.1 in Dwork and Roth (2014). To see this in the current context, let $\mathcal{M}(\cdot)$ be a DP(ε, δ) mechanism and let f be any function not depending on the unperturbed data, such as the function that maps negative values to zero. Then $f(\mathcal{M}(\cdot))$ is DP(ε, δ), since

$$\begin{aligned}\mathbb{P}(f(\mathcal{M}(\mathbf{a})) \in S) &= \mathbb{P}(\mathcal{M}(\mathbf{a}) \in f^{-1}(S)) \leq e^{\varepsilon}\mathbb{P}(\mathcal{M}(\mathbf{a}') \in f^{-1}(S)) + \delta \\ &= e^{\varepsilon}\mathbb{P}(f(\mathcal{M}(\mathbf{a}')) \in S) + \delta.\end{aligned}$$

Another common post-processing step performed on perturbed tables is the application of an algorithm to ensure that each marginal cell value equals the sum of the corresponding cell values. Such post-processing after a DP perturbation would not affect the differential privacy property of the table.

4.5 Zero Cells

Structural zeros, that is, cells representing combinations of attributes that are known to be impossible, need not be published since their value, zero, is known a priori. Therefore, there is no need to publish a structural zero, and no need to perturb it if published. We shall simply assume that our lists do not contain structural zeros.

In the case of non-structural zeros, there may be an impression that such zero cells do not constitute a disclosure risk, since an empty cell cannot reveal information about anyone. However, consider the following scenario: suppose the intruder wishes to know the health status of a targeted individual, who lives in a certain area and is in a known age group. Suppose the intruder knows that excluding the targeted individual, there is no individual having the given disease in this area and age group. If non-structural zeros are not perturbed, and if the targeted individual does not have the disease then the corresponding cell would be empty in the released data. Observing a zero in this cell, the intruder can conclude that the targeted individual does not have the disease. This is reflected in differential privacy as follows. Consider only the cell in question, as if this is the whole list. Then $\mathbb{P}(\mathcal{M}(0) = 1) = 0$ while $\mathbb{P}(\mathcal{M}(1) = 1) > 0$. Taking $S = \{1\}$ in (4.4) we can have $\mathbb{P}(\mathcal{M}(1) = 1) \leq e^{\varepsilon}\mathbb{P}(\mathcal{M}(0) = 1) + \delta$ only with $\delta = \mathbb{P}(\mathcal{M}(1) = 1)$, and in general there is no reason for this value to be small. Note that neighbouring lists can differ in the above way in a given cell.

Therefore we conclude that non-structural zeros should be perturbed. However, constraining the perturbed values to be non-negative can introduce statistical bias. Unless $p(0,0) = 1$, there is a positive bias, and $p(0,0) = 1$ implies $p(0,1) = 0$. It is straightforward to verify that $\mathrm{DP}(\varepsilon)$ cannot be satisfied if $p(0,1) = 0$ and $p(1,1) > 0$. On the other hand, if we relax to $\mathrm{DP}(\varepsilon,\delta)$ then we need a condition such as $p(1,1) \le \delta$ which seems very undesirable for small δ. Thus differential privacy and unbiasedness are contradictory, unless release of negative values is allowed.

5. EXAMPLES OF EXPONENTIAL PERTURBATION MECHANISMS

In this section, we study in more detail three special cases of the general exponential mechanism introduced in Section 4.2. We discuss the nature of these mechanisms, compare their differential privacy properties and illustrate numerically the utility consequences of the different choices of differential privacy parameters.

5.1 Laplace Perturbations

Corresponding to ℓ_1 in Section 4.2, we have the utility function $u_1 = u_1(\mathbf{a},\mathbf{b}) = -\sum_{k=1}^{K} |a_k - b_k|$. We first consider perturbation without truncation. To construct an exponential mechanism as in Equation (4.6), we need to determine Δu_1. Assume for now that each individual appears in the list only in one cell and therefore when one individual is removed or added relative to the list, only one cell count changes by 1. This assumption will be removed later. In terms of d defined above as the maximal number of cells in which two neighbours, $\mathbf{a}$ and $\mathbf{a}'$ can differ, we have $d = 1$. It follows readily that $\Delta u_1 = 1$. We remark that the maximum value of Δu as in (4.7) is attained for all $\mathbf{a}$, $\mathbf{a}'$, $\mathbf{b}$ so here the worst case is typical. This is one explanation why the exponential mechanism constructed from u_1 is very efficient for frequency tables.

Under this choice of utility function, the exponential mechanism becomes a discretised Laplace perturbation distribution, or a symmetric geometric distribution having probability $p(a,b)$ of perturbing a cell count a to b given by the proportionality relation

$$p(a,b) = \frac{1}{C} e^{-\varepsilon|b-a|}, \quad a = 0,1,\ldots,\ b = 0,\pm 1,\pm 2,\ldots. \tag{5.1}$$

where the normalizing constant is $C = \sum_{k=-\infty}^{\infty} e^{-\varepsilon k} = 1 + 2e^{-\varepsilon}/(1-e^{-\varepsilon})$. Theorem 4.1 implies $\mathrm{DP}(\varepsilon)$. Clearly one can view this perturbation as adding to a cell count a a random variable X satisfying $\mathbb{P}(X = x) = \frac{1}{C} e^{-\varepsilon|x|}$ for all integers x. As always, one can replace negative perturbed frequencies by zero.

We can impose truncation of the type $|a_k - b_k| \le m$ as above to improve utility, and the conditions of Theorem 4.2 hold. In this case we have

$$p(a,b) = \frac{1}{C_m} e^{-\varepsilon|b-a|} \quad \text{for } b \text{ satisfying } -m \le |b-a| \le m, \tag{5.2}$$

where $C_m = \sum_{k=-m}^{m} e^{-\varepsilon|k|} = 1 + 2(e^{-\varepsilon} - e^{-(m+1)\varepsilon})/(1-e^{-\varepsilon})$, obtained by using the geometric series formula. In this case, it follows from the formula of Section 4.5, that $\delta = e^{-\varepsilon m}/C_m$ and by Theorem 4.2 we obtain $\mathrm{DP}(\varepsilon,\delta)$. Again, negative perturbed values can be replaced by zero, maintaining the same level of differential privacy. For $\varepsilon = 1$ and $m = 10$ we obtain $\delta = 0.00002$. It is readily seen that δ decreases in m for each ε, so in terms of the differential privacy parameters the larger m the better.

A strong universal optimality property of the discrete Laplace (two-sided geometric) perturbation for the case of perturbing a single cell appears in Ghosh, Roughgarden and Sundararajan (2012). They show that without truncation, the Laplace perturbation of a single cell is optimal relative to a wide class of loss functions that includes the ones we consider, provided some post processing of the kind we do, e.g., replacing negative outputs by zero, is performed. More specifically, they show that Laplace with DP(ε) minimizes $E_b[\sum_{\mathbf{a}} \ell(a,b)] = \sum_{\mathbf{a}} \sum_b \mathbb{P}(\mathcal{M}(a = b)\ell(a,b)$ among all DP(ε) mechanisms having the same range, provided $\ell(a,b)$ is non-negative and non-decreasing in $|a-b|$ for all a, the frequency in the single cell. This was followed by Brenner and Nissim (2010) where it is shown such universality does not extend beyond a single cell, and therefore does not apply for tables as in this paper. Still, the Laplace perturbation seems like a very efficient choice.

5.2 Normal Perturbations

As a further example of the exponential mechanism, consider the utility function u_2. We show below that without truncation we have $\Delta u_2 = \infty$. Therefore, in order to determine a finite Δu_2 , we truncate the perturbations by m so that $|a_k - b_k| \leq m$ for all k. This forces us to consider DP(ε, δ) with $\delta > 0$.

Making the same assumption as in the previous section that $d = 1$, we have $\Delta u_2 = 2m + 1$, since in cells that differ between neighbouring lists we have $(a + 1 - b)^2 - (a - b)^2 = 2(a - b) + 1$ and likewise if +1 is replaced by -1. Clearly Δu_2 can be finite only if m is finite. The probability $p(a,b)$ is now given by the proportionality relation

$$(5.3) \qquad p(a,b) = \frac{1}{D_m} e^{-\varepsilon(b-a)^2/(2m+1)}, \quad \text{for } b \text{ satisfying } |b-a| \leq m$$

where $D_m = \sum_{k=-m}^{m} e^{-\varepsilon k^2/(2m+1)}$. This is a discretised and truncated normal normal distribution. Theorem 4.2 guarantees DP(ε, δ) with $\delta = e^{-\varepsilon m^2/(2m+1)}/D_m$. For $\varepsilon = 1$ and $m = 10$ we have $\delta = 0.001$.

5.3 Maximum Entropy Perturbation

One of the desiderata of frequency table dissemination mechanisms noted in Section 2.2 is that the distribution of the perturbations has maximum entropy, subject to the range and first two moments (see Andersson, Jansson and Kraft, 2015; Marley and Leaver, 2011). This may be intuitively appealing, and if one takes the variance of the perturbation as being indicative of its confidentiality protection performance, then maximum entropy subject to variance makes sense, although we are not aware of a formal statement regarding its advantage.

The normal distribution is well known to have maximum entropy subject to a given variance and range on the real line. Numerical calculations show that a discretised version as used above has approximately maximum entropy. An exact calculation of the discrete maximum entropy perturbation distribution subject to variance and range constraint requires a calculation using Lagrange multipliers.

The Laplace distribution has a similar characterization, if the range and expectations are prescribed. In fact, the principle of maximum entropy in statistics goes back to Laplace. Again the discrete version inherits an approximate maximum entropy property. As the tables above and results given later on testing independence suggest, Laplace perturbations perform better than Normal, suggesting that the principle of maximum entropy subject to variance should be reconsidered.

5.4 Hellinger-type Perturbations

Turning to the utility function $u_3 = u_3(\mathbf{a}, \mathbf{b}) = -\sum_{k=1}^{K} |\sqrt{a_k} - \sqrt{b_k}|$, easy calculations show that $\Delta u_3 = 1$, assuming again that $d = 1$. However, in this case the maximum in (4.7) is attained in the extreme case of small $\mathbf{a}$, $\mathbf{a}'$ due to the concavity of $\sqrt{x}$, so here the worst case is not typical unless all cells are very small. In other words, for large cells, the value of Δu in the exponential mechanism is too large, making the inequalities in the proof of Theorem 4.1 crude, and therefore leading to over-perturbation and loss of utility. For the exponential mechanism with u_3 we have

$$(5.4) \qquad p(a, b) \propto e^{-\varepsilon|\sqrt{b}-\sqrt{a}|/2}, \quad a, b = 0, 1, \ldots.$$

and Theorem 4.1 implies DP(ε).

Although the loss function ℓ_3 that corresponds to u_3 has very attractive properties, the worst case aspect explained above implies that as a perturbation mechanism the scheme defined in (5.4) performs very poorly in terms of data utility. We believe that it is a somewhat interesting lesson that a loss function that appears so natural leads to a poor mechanism.

5.5 Comparisons of Perturbation Mechanisms

In Table 2 we calculate the probability of obtaining a perturbed value in an interval range of ± 0 to ± 4 of the original value, when the original values are 0 to 5 and over, $\varepsilon = 1.5$ and $\varepsilon = 0.5$ for both the Laplace and Normal perturbations, and negative values are replaced by zero. In order to compare the two perturbation mechanisms we fix the value of δ for each ε. For $\varepsilon = 1.5$ and $\delta = 0.00002$, Laplace perturbations are truncated at $m = 7$ and Normal perturbations are truncated at $m = 12$. For $\varepsilon = 0.5$ and $\delta = 0.008$, Laplace perturbations are truncated at $m = 7$ and Normal perturbations are truncated at $m = 10$.

From Table 2, it is clear that the Laplace perturbations have higher utility under differential privacy with given ε and δ. All perturbed values are within ± 3 for $\varepsilon = 1.5$ and $\delta = 0.00002$ and over 92% of the perturbed values are within ± 4 for $\varepsilon = 0.5$ and $\delta = 0.008$. The corresponding probabilities for the normal perturbations are between 6% and 25% lower. Note that replacing all negative perturbed values by zero impacts on the perturbation ranges when a zero is included in the interval.

A similar calculation for Hellinger-type perturbations shows that they are considerably worse than the other perturbation mechanisms, and the probabilities are very small compared to those in Table 2. Therefore, we will not include the Hellinger-type perturbations in further analyses.

TABLE 2

Probability of range for Laplace and Normal perturbations with negative values replaced by zero

	Range for $\varepsilon = 1.5$ and $\delta = 0.00002$					Range for $\varepsilon = 0.5$ and $\delta = 0.008$				
Original Value	± 0	± 1	± 2	± 3	± 4	± 0	±1	± 2	± 3	± 4
	Laplace $m = 7$					Laplace $m = 7$				
0	0.82	0.96	0.99	1.00	1.00	0.63	0.78	0.87	0.93	0.96
1	0.64	0.96	0.99	1.00	1.00	0.25	0.78	0.87	0.93	0.96
2	0.64	0.92	0.99	1.00	1.00	0.25	0.55	0.87	0.93	0.96
3	0.64	0.92	0.98	1.00	1.00	0.25	0.55	0.74	0.93	0.96
4	0.64	0.92	0.98	1.00	1.00	0.25	0.55	0.74	0.85	0.96
≥5	0.64	0.92	0.98	1.00	1.00	0.25	0.55	0.74	0.85	0.92
	Normal $m = 12$					Normal $m = 10$				
0	0.57	0.70	0.81	0.89	0.94	0.54	0.63	0.71	0.78	0.84
1	0.14	0.70	0.81	0.89	0.94	0.09	0.63	0.71	0.78	0.84
2	0.14	0.40	0.81	0.89	0.94	0.09	0.26	0.71	0.78	0.84
3	0.14	0.40	0.62	0.89	0.94	0.09	0.26	0.42	0.78	0.84
4	0.14	0.40	0.62	0.78	0.94	0.09	0.26	0.42	0.57	0.84
≥5	0.14	0.40	0.62	0.78	0.88	0.09	0.26	0.42	0.57	0.69

5.6 Risk-Utility Analysis

5.6.1 Utility of the Laplace and Normal Perturbations We begin by presenting some expressions for the expected loss under these mechanisms. Beginning with Laplace perturbation and setting $\alpha = e^{-\varepsilon}$ we have

$$E(|b-a|) = \sum_{k=-m}^{m} |m|e^{-\varepsilon k} = 2\alpha(m\alpha^{(m+1)} - (m+1)\alpha^{m} + 1)/C_m(\alpha-1)^2,$$

where C_m is defined in (5.2). Letting $m \to \infty$ we obtain for the untruncated case, $E(|b-a|) = e^{-\varepsilon}/C(e^{-\varepsilon}-1)^2$ with $C = 1 + 2e^{-\varepsilon}/(1-e^{-\varepsilon})$. If we replace negative outputs by zero, the loss improves.

Turning to normal perturbations, we have

$$E(|b-a|^2) = \sum_{k=-m}^{m} |m|^2 e^{-\varepsilon k^2/(2m+1)}/D_m,$$

where D_m is defined after (5.3). Again, if we replace negative outputs by zero, this utility improves.

5.6.2 Risk-Utility Plots In this section, we shall present risk-utility plots for the real Table 1 and for additional two-way tables that were generated assuming independence of the two attributes, in order to assess the impact of the perturbation mechanisms on statistical inference. Risk is measured in terms of the value of ε, from $\varepsilon = 0.1$ to $\varepsilon = 3.0$, for both the Laplace and Normal perturbations. The truncation of m is fixed at $m = 7$ for the Laplace perturbations and allowed to vary for the Normal perturbations to ensure the same value of δ for each ε. For $\varepsilon = 0.1, 0.5, 1.0, 1.5, 2.0, 3.0$ the corresponding values of m for the Normal perturbations are 8, 10, 12, 12, 13, 14, respectively. Utility is measured using the loss functions ℓ_1, ℓ_2, and ℓ_3 defined in Section 4.2 as well as by the accuracy of the Cramer's V statistic and the associated p-value for the Chi-square test for independence.

Figure 1 presents results of applying perturbations to Table 1. For each ε, the table was perturbed 100 times in order to produce the box plots. The real table is

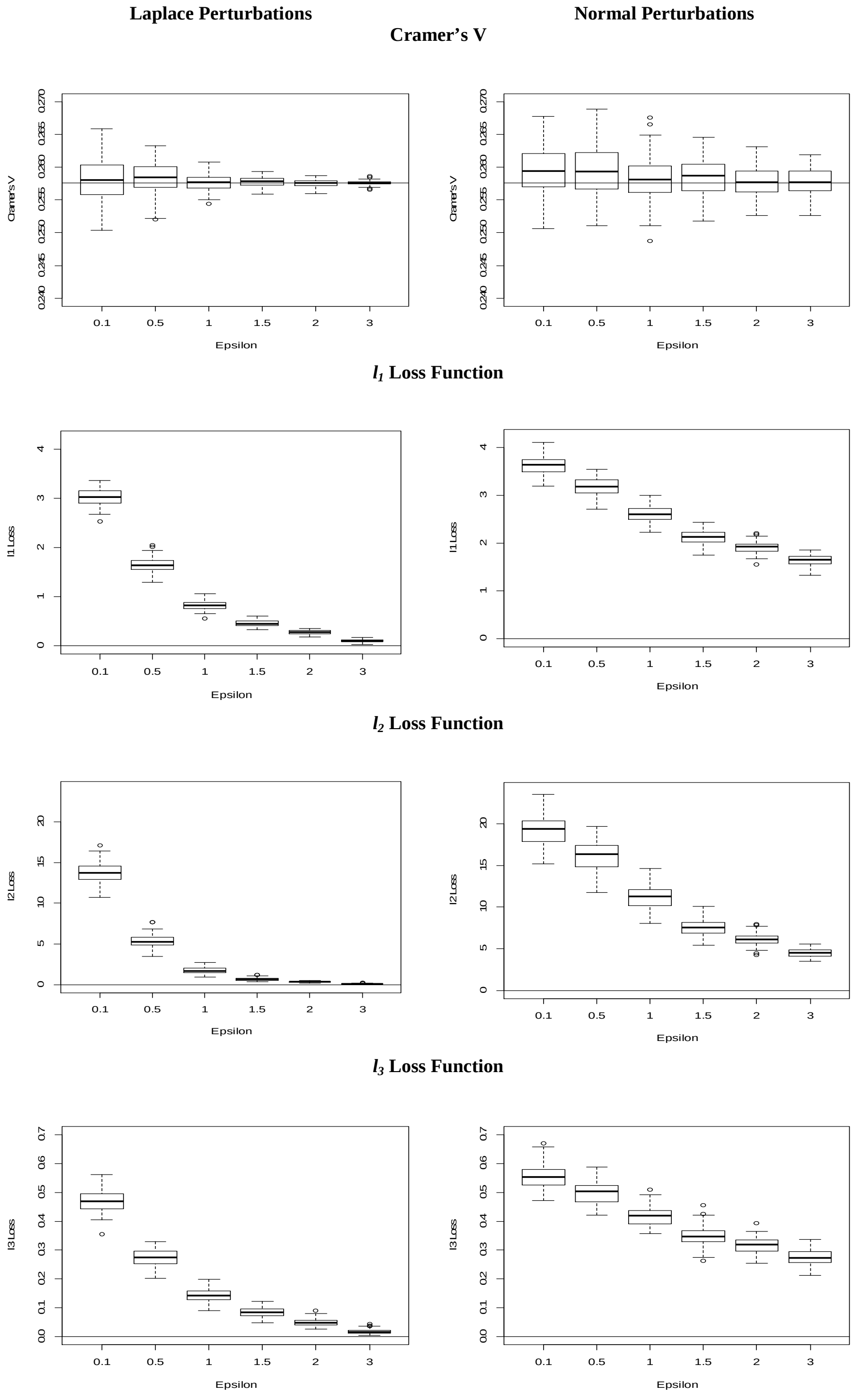

FIG 1. *Values of Cramer's V and three loss functions over 100 perturbation repetitions for each ε for Table 1*

highly dependent and hence p-values (not shown) for testing independence were close to zero for the original table and all perturbations and the inference did not change. The true value of Cramer's V is represented by the horizontal line and we can see that under both perturbation mechanisms, the inter-quartile range of the statistic is less than 0.005. The three loss functions are also included in Figure 1 where the smaller the value, the higher the utility. It is clear that utility improves as ε increases. In all cases, the Laplace perturbations show higher utility and in fact out-performs the Normal perturbations even for the ℓ_2 loss function which defines the exponential mechanism for Normal perturbations.

In order to assess the impact of the perturbations on statistical inference when testing for independence on the perturbed data as if they were true data, we generated two tables having two independent attributes, both with a population size of $N = 10,000$, a large table with 1,000 cells (average cell size of 10) and a small table with 100 cells (average cell size of 100). The marginal probabilities of the tables were generated by the Dirichlet distribution. From the marginal probabilities, we define the internal probabilities under the assumption of independence $p_{ij} = p_{i.}p_{.j}$. Finally, we generated the counts in the table by random draws from $\mathrm{Mult}(N, p_{ij})$. We carried out 100 perturbations on each table and under each ε for the Laplace and Normal perturbations using the same settings of m as described above to ensure equal δ.

Figures 2 and 3 show the risk-utility plots for the two table. The horizontal lines for the p-value and Cramer's V statistic show the true values obtained from the original tables. We see that utility improves as ε increases and the Laplace perturbations out-perform the Normal perturbations as expected. Under both perturbation mechanisms we rarely change the inference from independence to dependence for the small table (with large counts) but this is not the case for the large table (with small counts). For the latter table under the Normal perturbations, we are unable to obtain correct inference for any of the ε whilst under the Laplace perturbations we would need ε over 2.0 in order not to reject independence. For the Cramer's V statistic the Normal perturbations in the large table show greater discrepancies than the small table, and compared to the Laplace perturbations. The three loss functions are also shown in the figures for comparison.

6. DATA ANALYSIS TAKING THE PERTURBATIONS INTO ACCOUNT

In Section 5 we compared the properties of the alternative mechanisms and generally found that Laplace perturbations led to higher utility for a given value of ε. Nevertheless, in absolute terms, the distortion of analyses arising from even Laplace perturbations was non-negligible for values of ε of, say, 0.5 or 1.0. Thus, for such values of ε, we see in Figures 2 and 3 clear evidence of bias in the estimation of the Cramer's V parameter and evidence that p-values for testing independence are often very different to the p-value in the original table. Fienberg, Rinaldo and Yang (2010) and Wang, Lee and Kifer (2015) have also discussed how such perturbations can lead to unreliable conclusions in the analysis of tables if their presence is ignored. As a result, it is of interest to consider approaches to analysis of the perturbed table which take account of the perturbation mechanism.

As mentioned in the Introduction, the standard assumption in differential pri-

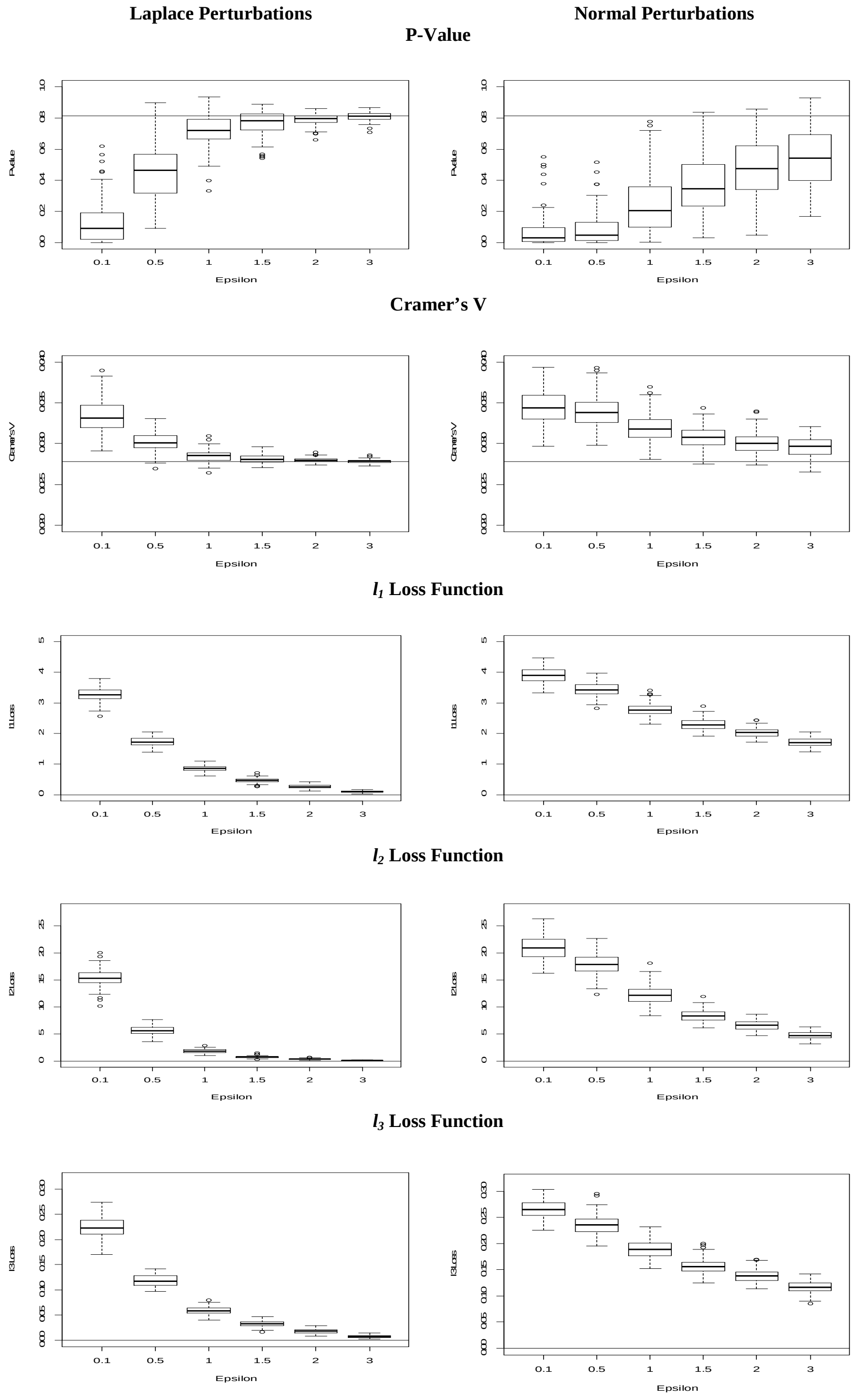

FIG 2. *p-values, Cramer's V, and three loss functions over 100 perturbation repetitions for each ε for the small independent table (average cell size=100)*

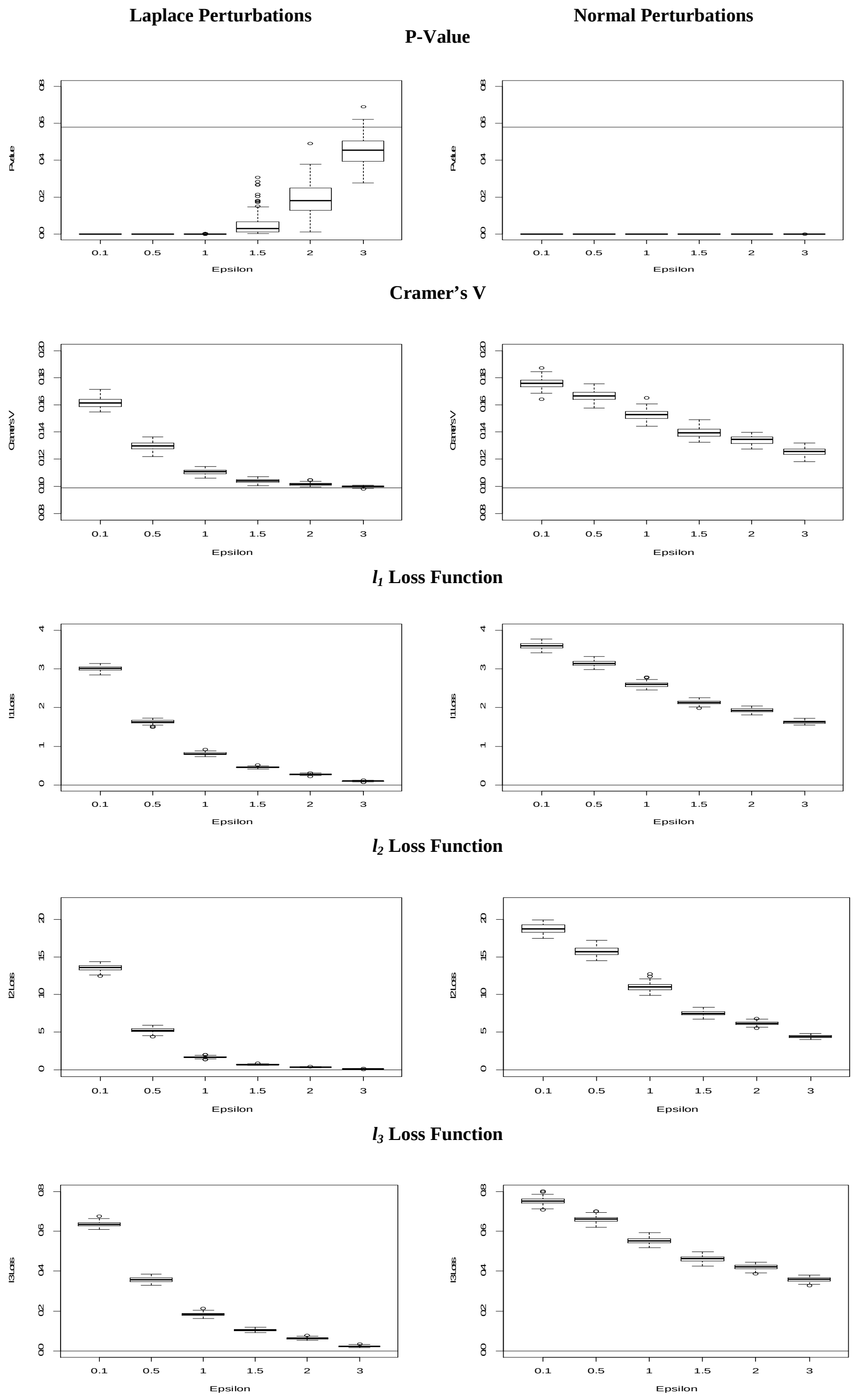

FIG 3. *p-values, Cramer's V, and three loss functionsover 100 perturbation repetitions for each ε for the large independent table (average cell size=10)*

vacy is that the perturbation distribution should be known to data users analysing the perturbed data. Our goal in this section is to show that if the perturbation distribution is known, it can be taken it into account when estimating or testing hypotheses on parameters of models pertaining to the unperturbed data. This is demonstrated here very briefly by a simple toy example. The likelihood-based approach we present generalizes in theory to more complex models and hypotheses, however, with heavier computations. Karwa, Kifer and Slavković (2015) consider that in most cases the likelihood is intractable and that approximate computational methods are needed. Other methods of improving the performance of tests of independence in two-way tables under such perturbation have been proposed by Uhler, Slavković and Fienberg (2013) and Wang, Lee and Kifer (2015) .

Consider a list of two cells, $\mathbf{a} = (a_1, a_2)$, where a_1 is a random variable, and $a_1 \sim$ Binomial(N, p) is assumed, and $a_2 = N - a_1$. If a_1 is the number of individuals having a certain property, then clearly $\Delta u_1 = 1$. The perturbed data to be released is $X = a_1 + L$, where L has the Laplace perturbation defined in (5.1). The likelihood of an observation X is a function of p (we consider N known, as usual):

$$\begin{aligned} L_x(p) &= P(X = x) = P(a_1 = x - L) \\ &= \sum_{\ell=\max\{-m,x-N\}}^{\min\{x,m\}} \binom{N}{x-\ell} p^{x-\ell}(1-p)^{N-x+\ell} \, \frac{e^{-\varepsilon|\ell|}}{\sum_{k=-m}^{m} e^{-\varepsilon|k|}}. \end{aligned}$$

The likelihood ratio statistic for the goodness of fit of the parameter value p_0 given $X = x$ is

$$\max_p L_x(p)/L_x(p_0),$$

and we reject $H_0 : p = p_0$ if the statistic is large.

Figure 4 shows histograms of 500 values of 2 log(likelihood ratio statistic) obtained by simulation when the data comes from $p = 0.5$ and we test $H_0 : p = 0.5$ and $H_0 : p = 0.7$, with $N = 80$ and for the perturbation we have $\varepsilon = 0.5$ and $m = 5$. In this case the formulas around (5.2) show that $\delta = 0.02$ so we have DP$(0.5, 0.02)$. The plot on the left of Figure 4 shows that for $p = 0.5$ the statistic values are mostly small, and when testing $p = 0.7$, the plot on the right shows that most values of the statistic are large, and $H_0 : p = 0.5$ is rejected. For numerical reasons, if twice the likelihood ratio exceeded 50, it was set as 50.

Of the 500 values for testing $H_0 : p = 0.5$, 95% are below the (empirical) critical point of $c = 3.36$. This should be compared with the critical value of 3.84 for the Chi-square with df=1 asymptotic distribution. For testing $H_0 : p = 0.7$, the proportion of statistics out of the simulated 500 that are above $c = 3.36$ is 0.95. Thus the power of our test, at level of significance $\alpha = 0.05$ is 0.95, whereas the power of the same test without the Laplace noise is 0.96. The added noise did not reduce the power by much in the present case. If one uses the asymptotic critical value of 3.84, rather than the empirical 3.36, the empirical power and level of significance change very little, implying that the asymptotic theory of the likelihood ratio statistic applies at this sample size.

For $m = 10$ with other parameters as above we obtain $c = 3.82$, the empirical power for testing $H_0 : p = 0.7$ with $\alpha = 0.05$ is 0.92, and $\delta = 0.00166$ as can be seen from Table 2. Thus, allowing a larger perturbation range, that is ±10 rather

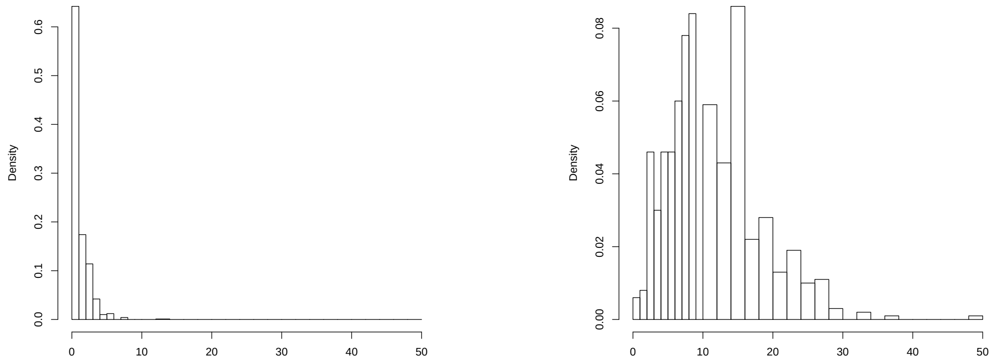

Fig 4. *Histogram of 500 2 log(likelihood ratio) tests when $N = 80$, $p = 0.5$, $\varepsilon = 0.5$, and $H_0 : p = 0.5$ (left), and $H_0 : p = 0.7$ (right) is tested*

than ±5, improves (reduces) δ, at the cost of some reduction in the power of the test.

From the histograms (for $m = 5$) one can obtain the power of the test for any given significance level by choosing a point on the x-axis and looking at the percentage of values below the point in the left histogram (level of significance) and above in the right one (power). A comparison to the case of no noise shows that the loss of power is not very significant, and the left histogram quite resembles a Chi-square$_1$ distribution, to which it converges with N.

7. COMPLEX LISTS WITH OVERLAPPING CELLS

In this section we deal with lists in which an individual may appear in more than one cell. This arises, in particular, if the list includes margins as well as interior cells in a multi-way frequency table. Margins (perturbed) can be computed by summing perturbed interior cells, however, such aggregation results in a standard deviation (SD) that becomes larger with the number of summands. If some marginals are of special interest, the agency can release them with their own perturbation, which may have a smaller SD than that obtained by aggregation.

Overlapping cells affect the number d of cells in which two neighbouring lists can differ. For example, if the list consists of a t-way table and all its marginals except for the total which is almost always known, then it is easy to see that each individual appears in $2^t - 1$ cells, and therefore two neighbouring lists can differ in $d = 2^t - 1$ cells.

Focusing now on the Laplace exponential mechanism, we now have $\Delta u_1 = d$. The exponential mechanism will now perturb according to $p(a, b) \propto e^{-\varepsilon|b-a|/d}$, which is equivalent to replacing ε by ε/d, in order to obtain DP(ε). For d large this results in large perturbations and reduced utility. In fact, the discrete Laplace perturbation distribution of (5.1) with ε replaced by ε/d has SD approximately $\sqrt{2}d/\varepsilon$, which will apply to all released cells. Note that if we perturb interior cells and marginals independently, then the released table will be inconsistent in the sense that perturbed marginals are unlikely to coincide with the relevant sums of the perturbed interior cells, though they will generally be close.

Consider a t-way table where each of its t attributes has C categories, say, and the user computes marginals by summing over interior cells. In this case consistency of interior cells and marginals is obvious and each cell in a $k-$dimensional marginal table is obtained as the sum of C^{t-k} frequencies. If each cell is perturbed with a SD proportional to $\sqrt{2}/\varepsilon$ and only interior cells are released, then $d = 1$ and we obtain DP(ε), and the standard deviation of the sum of the perturbations in a $k-$dimensional marginal table will be proportional to $\sqrt{2C^{t-k}}/\varepsilon$. In a 4-way table with $C = 10$, for example, if only interior cells are perturbed, the SD of the perturbation for each cell of a 2-dimensional marginal is proportional to $\sqrt{2C^2}/\varepsilon = \sqrt{2 \cdot 10^2}/\varepsilon \approx 14/\varepsilon$, and if all marginals are perturbed the SD is approximately to $\sqrt{2}d/\varepsilon = \sqrt{2}(2^4 - 1)/\varepsilon \approx 21/\varepsilon$ so for such marginals the scheme that perturbs only interior cells is preferable in the sense of having a smaller SD. For the interior cells themselves the situation is of course even better because if only interior cells are perturbed then $d = 1$ and the perturbation SD is $\sqrt{2}/\varepsilon$ rather than $\sqrt{2}d/\varepsilon$, which is 15 times larger in the above example. When considering the release of a table, the importance of some marginals relative to others and interior cells should be considered when deciding on the perturbation scheme, and in many situations, perturbing only interior cells, and letting users compute marginals from those perturbed cells, is efficient.

It may also be useful to perturb interior cells and different marginal tables with different values of ε, depending on the importance of these marginals. We can allow smaller perturbation for some marginals and compensate by larger perturbations in others. In this case we consider several mechanisms $\mathcal{M}_i$ for $i = 1, \dots k$ and apply them on the same data. This is known in the differential privacy literature as composition. To assess whether such schemes satisfy differential privacy, the composition Theorem 3.16 in Dwork and Roth (2014) is relevant. We bring a proof in order to keep the paper as self contained as possible.

THEOREM 7.1. *Let $\mathcal{M}_i$ be independent DP(ε_i, δ_i) mechanisms for $i = 1, \dots k$. Then $(\mathcal{M}_1 \dots, \mathcal{M}_k)$ is DP($\sum_{i=1}^k \varepsilon_i, \sum_{i=1}^k \delta_i$).*

Proof It suffices to consider $k = 2$, and then proceed by induction. Let the ranges of $\mathcal{M}_i$ be B_i for $i = 1, 2$ and $S = S_1 \times S_2 \subseteq B := B_1 \times B_2$ and denote $S_1(s_2) = \{s_1 : (s_1, s_2) \in S\}$. Below, the first inequality uses the differential privacy property of $\mathcal{M}_1$ and the second uses $(c + \delta) \wedge 1 \leq c \wedge 1 + \delta$. The third inequality uses the differential privacy property of $\mathcal{M}_2$ and the last one and the first equality are obvious. We have

$$\begin{aligned}
\mathbb{P}((\mathcal{M}_1(\mathbf{a}), \mathcal{M}_2(\mathbf{a})) \in S) &= \sum_{s_2 \in S_2} \mathbb{P}(\mathcal{M}_1(\mathbf{a}) \in S_1(s_2))\mathbb{P}(\mathcal{M}_2(\mathbf{a}) = s_2) \\
&\leq \sum_{s_2 \in S_2} [\{e^{\varepsilon_1}\mathbb{P}(\mathcal{M}_1(\mathbf{a}') \in S_1(s_2)) + \delta_1\} \wedge 1]\mathbb{P}(\mathcal{M}_2(\mathbf{a}) = s_2) \\
&\leq \sum_{s_2 \in S_2} [\{e^{\varepsilon}\mathbb{P}(\mathcal{M}_1(\mathbf{a}') \in S_1(s_2))\} \wedge 1]\mathbb{P}(\mathcal{M}_2(\mathbf{a}) = s_2) + \delta_1 \\
&\leq \sum_{s_2 \in S_2} [\{e^{\varepsilon_1}\mathbb{P}(\mathcal{M}_1(\mathbf{a}') \in S_1(s_2))\} \wedge 1][e^{\varepsilon_2}\mathbb{P}(\mathcal{M}_2(\mathbf{a}') = s_2) + \delta_2] + \delta_1 \\
&\leq e^{\varepsilon_1 + \varepsilon_2}\mathbb{P}((\mathcal{M}_1(\mathbf{a}'), \mathcal{M}_2(\mathbf{a}')) \in S) + \delta_1 + \delta_2. \quad \square
\end{aligned}$$

Theorem 3.20 in Dwork and Roth (2014) provides a more advanced composition result, where instead of obtaining DP($k\varepsilon$) when composing k mechanisms with DP(ε), as in Theorem 7.1, a composition with DP(ε', δ) is obtained with ε' of order $\sqrt{k}\varepsilon$ but with constants depending on δ that make it useful only for rather large values of k. We shall not present or use this result.

As an example consider now a 3-way table $\{X_{ijk}\}$, and suppose we wish to perturb independently all interior cells and marginals. In this case, the list $\mathbf{a}$ consists of 7 tables:

$$\mathbf{a} = \Big(\{X_{ijk}\}, \{\sum_i X_{ijk}\}, \{\sum_j X_{ijk}\}, \{\sum_k X_{ijk}\}, \{\sum_{ij} X_{ijk}\}, \{\sum_{ik} X_{ijk}\}, \{\sum_{jk} X_{ijk}\}\Big).$$

For the whole list $\mathbf{a}$ we have $d = 2^3 - 1 = 7$, and we can apply (5.1) with ε replaced by $\varepsilon/7$ to obtain DP(ε). Alternatively, we can apply Theorem 7.1. Each of the above 7 tables has $d = 1$, and if we apply a Laplace perturbation with $\varepsilon/7$ for each of the 7 tables of the above $\mathbf{a}$, we obtain again DP(ε).

More generally one can release the rth table of $\mathbf{a}$ with DP(ε_r), $r = 1, \ldots, 7$, using the corresponding Laplace perturbation, and by Theorem 7.1, the whole list will be released with DP($\sum_{i=1}^{7} \varepsilon_i$). Suppose we expect users to be more interested in 2-dimensional tables, and less in others. For example, if the attributes are Income, Education, and Ethnicity, then it may be that the releasing agency or the data users consider Ethnicity to be of lesser importance, and the important table might be Income by Education, and the table of interior cells, so that one can see the Income by Education table for each fixed Ethnicity. In this case $\{X_{ijk}\}$ and $\{\sum_k X_{ijk}\}$ could be released with DP($\varepsilon/3$), say, and the other 5 tables with DP($\varepsilon/15$). The latter tables may be quite perturbed, much more than the important ones, and the whole release will satisfy DP(ε).

The above discussion provide tools that can help the data releasing agency decide on the construction of the list and the amount of perturbations of different parts according to the number of categories of the attributes, the expected interest in particular marginals (which are often more relevant than interior cells), and the dimension of the table and the marginals of interest.

8. CONCLUSIONS

In this paper we have considered modern perturbation schemes that resemble ones used by some official agencies when releasing frequency tables, with the goal of assessing how random perturbations protect confidentiality in terms of differential privacy. We have seen how this approach can highlight specific issues, such as the effect of truncation. We have studied some alternative perturbation mechanisms and found that Laplace perturbation has clear advantages in terms of the utility of the resulting tables for a given level of confidentiality protection. Maximum entropy perturbations, subject to variance constraints is one existing criterion for selecting perturbations in disclosure control, but the implied approximately normal perturbations did not perform well in our assessment. We found that insisting on releasing only nonnegative perturbed frequencies may result in loss of utility, without a well defined gain in confidentiality protection.

We have studied the trade off between different values of the two parameters ε and δ governing differential privacy and the utility of the resulting tables, and seen how compromises in the former values can make a considerable difference to the level of utility. We have noted that there are many desiderata that have been proposed for perturbation, for example that perturbed frequencies be non-negative, that they be unbiased for the true frequencies and that perturbations be truncated by a specified bound. We have also seen that some compromises of these criteria may be desirable. To what extent perturbation will damage the value of the tables for analysis will depend on user needs and it is hard to draw any general conclusion. Nevertheless, our examples suggest that Laplace perturbations may guarantee differential privacy with what we consider to be acceptable parameters, and preserve a fair amount of utility.

We have noted the desirability of making the nature and parameters of the perturbation mechanism available to users and the possibility that users could take account of this knowledge when analysing the data. Thus, in principle, given a specified model for the data and a perturbation mechanism, it is feasible to determine a likelihood function for the perturbed data, and make inference on the parameters of the data model. We demonstrated this procedure in a simple example. In practice, the computational challenges are severe for the kinds of tables released by national statistical agencies. But this is an area for further research.

Another area needing further research relates to tables based on sample data rather than on population counts, on which we focused in this paper. The cells in tables based on sample data may contain sample-based estimated counts, consisting of sums of survey weights. In this case, adding or removing a sample unit from the dataset will change the estimated count by the value of the corresponding survey weight. If $d = 1$ and w is the maximal possible weight then $\Delta u_1 = w$, and the earlier differential privacy methodology applies. In this paper we did not pursue this direction, the practicality of which seems to be worthwhile of investigation. Confidentiality considerations for sample-based tables may also take account of the potential confidentiality protection afforded by sampling, when sample membership can be assumed unknown (e.g. Chaudhuri and Mishra, 2006). Further protection may arise from the fact that sampling error considerations often lead government agencies to design tables that do not include cell estimates based on small numbers of sample units.

This paper focused on the non-interactive setting, where the list and all perturbations are prepared in advance to satisfy a given level of DP (although the perturbations can be applied only to the data actually requested). If some cells in the list are never requested, then their contribution to d or ε (and δ) can be seen as overprotection. The differential privacy literature proposes interactive query submission and monitoring for all users on line, responding to queries with a certain level of DP which accumulates as in Theorem 7.1, and allocating a "budget" of a certain ε_j to user j so that the total of all ε's (and δ's) achieves the required DP level. Such monitoring is quite demanding of the agencies, but could hopefully be automated. Further research on interactive dissemination by official agencies and its implications seems to be needed.

REFERENCES

Andersson, K., Jansson, I. and Kraft, K. (2015). Protection of frequency tables - current work at Statistics Sweden. Joint UNECE/Eurostat work session on statistical data confidentiality (Helsinki, Finland, 5-7 October). 20pp.

Auguste, K. (1883). La cryptographie militaire. *Journal des sciences militaires* **9** 538.

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F. and Talwar., K. (2007). 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 (PODS)* 273–282.

Brenner, H. and Nissim, K. (2010). Impossibility of differentially private universally optimal mechanisms. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on* 71–80. IEEE.

Chaudhuri, K. and Mishra, N. (2006). When random sampling preserves privacy. In *Proceedings of the 26th Annual International Conference on Advances in Cryptology: CRYPTO 2006* (C. Dwork, ed.). *LNCS* **4117** 198–213. Springer-Verlag, Berlin.

Chipperfield, J., Gow, D. and Loong, B. (2016). The Australian Bureau of Statistics and releasing frequency tables via a remote server. *Statistical Journal of the IAOS* **32** 53–64.

Drechsler, J. (2011). New data dissemination approaches in old Europe - synthetic datasets for a German establishment survey. *J Appl Stat* **39** 243–265.

Drechsler, J. and Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis* **55** 3232-3243.

Duncan, G. T., Elliot, M. and Salazar-Gonzàlez, J. J. (2011). *Statistical Confidentiality.* Springer, New York.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R. and Roehrig, S. F. (2001). Disclosure limitation methods and information loss for tabular data. *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* 135–166.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* **9** 211–407.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *3rd IACR Theory of Cryptography Conference* 265–284.

Fellegi, I. P. (1972). On the question of statistical confidentiality. *Journal of the American Statistical Association* **67** 7–18.

Fienberg, S. E., Rinaldo, A. and Yang, X. (2010). Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *International Conference on Privacy in Statistical Databases* 187–199. Springer.

Fienberg, S. E. and Slavković, A. B. (2008). A survey of statistical approaches to preserving confidentiality of contingency table entries. In *Privacy-Preserving Data Mining* 291–312. Springer.

Fraser, B. and Wooton, J. (2005). A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing. In *Joint UNECE/Eurostat work session on statistical data confidentiality. Topic (v): Confidentiality aspects of tabular data, frequency tables, etc* **WP. 35** 5pp. United Nations Statistical Commission and Economic Commission for Europe Conference of Europe Statisticians. European Commission Statistical Office of the European Communities (Eurostat), Geneva, Switzerland.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9** 383–383.

Ghosh, A., Roughgarden, T. and Sundararajan, M. (2012). Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing* **41** 1673–1693.

Gomatam, S. and Karr, A. (2003). Distortion measures for categorical data swapping Technical Report, National Institute of Statistical Sciences. www.niss.org/downloadabletechreports.html.

Gotz, M., Machanavajjhala, A., Wang, G., Xiao, X. and Gehrke, J. (2012). Publishing search logs– a comparative study of privacy guarantees. *IEEE Transactions on Knowledge and Data Engineering* **24** 520–532.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and de Wolf, P. P. (2012). *Statistical Disclosure Control. Wiley Series in Survey Methodology.* John Wiley & Sons, United Kingdom.

Jansson, I. (2012). Issues and plans for the disclosure control of the Swedish Census 2011 Technical Report No. 2012-04-02, Statistika centralbyrån.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P. and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* **60** 224–232.

Karwa, V., Kifer, D. and Slavković, A. B. (2015). Private Posterior distributions from

Variational approximations. *arXiv preprint arXiv:1511.07896.*

Little, R. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics* **9** 407–426.

Longhurst, J., Tromans, N., Young, C. and Miller, C. (2007). Statistical disclosure control for the 2011 UK census. In *Joint UNECE/Eurostat conference on Statistical Disclosure Control, Manchester* 17–19.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J. and Vilhuber, L. (2008). Privacy: Theory meets Practice on the Map. In *Proceedings of the IEEE 24th International Conference on Data Engineering ICDE* 277 -286.

Marley, J. K. and Leaver, V. L. (2011). A Method for Confidentialising User-Defined Tables: Statistical Properties and a Risk-Utility Analysis. Proc 58th Congress of the International Statistical Institute, ISI 2011, 21–26 August.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on* 94–103. IEEE.

O'Keefe, C. M. and Chipperfield, J. O. (2013). A Summary of Attack Methods and Protective Measures for Fully Automated Remote Analysis Systems. *International Statistical Review* **81** 426–455.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9** 462–468.

Shannon, C. E. (1949). Communication theory of secrecy systems. *Bell system technical journal* **28** 656–715.

Shlomo, N. (2007). Statistical Disclosure Control Methods for Census Frequency Tables. *International Statistical Review* **75** 199–217.

Shlomo, N., Antal, L. and Elliot, M. (2015). Measuring disclosure risk and data utility for flexible table generators. *Journal of Official Statistics* **31** 305–324.

Thompson, G., Broadfoot, S. and Elazar, D. (2013). Methodology for automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics. Joint UNECE/Eurostat work session on statistical data confidentiality (Ottawa, Canada, 28-30 October 2013). 37pp.

Uhler, C., Slavković, A. and Fienberg, S. E. (2013). Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality* **5** 137-166.

van den Hout, A. and van der Heijden, P. G. M. (2002). Randomized Response, Statistical Disclosure Control and Misclassification: A Review. *International Statistical Review* **70** 269-288.

Wang, Y., Lee, J. and Kifer, D. (2015). Differentially Private Hypothesis Testing, Revisited. *arXiv preprint arXiv:1511.03376.*

Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association* **105** 375–389.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control. Lecture Notes in Statistics* **155**. Springer.