Confidentiality and Differential Privacy in the Dissemination of Frequency Tables

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

Yosef Rinott is Professor, The Federmann Center for the Study of Rationality, The Hebrew University, Jerusalem 91904 Israel. e-mail: yosef.rinott@mail.huji.ac.il.

Christine O'Keefe is Senior Principal Research Scientist, CSIRO, GPO Box 1700, Canberra ACT 2601, 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.

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 (ABS) 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, focusing on a version of the ABS TableBuilder as a concrete example of a data release mechanism, and examining its properties. 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, to the relationship between confidentiality protection and utility, and to confidentiality in general.

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

*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 public and private institutions collect data from individuals on economic, health, social and other variables. This paper focuses on frequency tables, which are the most common form of releasing data for use by researchers and the public. The data custodians are obliged to keep such information strictly confidential, not sharing or releasing any data in an identifiable form. The potential for breaches of confidentiality is real. See, for example, Sweeney (1997); Narayanan and Shmatikov (2008); Homer et al. (2008); Gymrek et al. (2013). 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. For further discussion and a general framework for evaluating utility see Karr et al. (2006). 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 (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 measures of disclosure risk that have been used in practice and studied in the SDC literature cited above and references therein, all of them are based on contestable assumptions about an intruder's prior knowledge

of the data and type of confidentiality attacks which they might employ. With the evolution of approaches to data dissemination and the recognition that protecting confidentiality of respondents is becoming increasingly more difficult in the era of data deluge, data custodians need to look for stricter definitions of disclosure risk and a more systematic and quantifiable approach to protecting confidentiality.

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 much attention in the computer science literature, see for example the recent monograph by Dwork and Roth (2014) and its references. The idea was introduced in a mathematically rigorous framework designed to give a well-defined quantification of the confidentiality protection guarantee. By employing a 'worst-case' approach and avoiding strong assumptions about which variables are sensitive to disclosure, intruders' prior knowledge, and attack scenarios, differential privacy has the potential for wide application. This worst-case approach may be deemed overprotective of confidentiality; for example, even sufficient statistics which are usually preserved in SDC approaches need to be perturbed. However, under differential privacy the worst-case approach 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 generally adopted 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 differential privacy under a realistic and popular dissemination scenario and, on the way, to provide a practical introduction to differential privacy 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 in order to keep this paper almost self-contained and therefore will not present the theory in full generality. We shall also present numerical work to assess the trade-off between confidentiality protection, measured via differential privacy parameters and utility, measured in different ways, but taking account of the kinds of analyses undertaken.

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 where individuals may be identified (*identity disclosure*) 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 (*attribute disclosure*)

(Shlomo, 2007). The occurrence of counts of one 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) or *inferential disclosure*. Traditionally, this latter type of disclosure risk was dealt with by manual control of tables that were released.

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. Synthetic data (Little, 1993; Rubin, 1993) methods could also be used (Drechsler, 2012; 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 include table redesign, cell suppression, rounding, or addition of noise directly on the cell counts of the frequency table. 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). Shlomo and Young (2008) developed a method of post-randomisation directly on cell counts based on a probability transition matrix which is related to the differential privacy approach presented in this paper. Perturbing the entire original data is often called *input perturbation* in the differential privacy literature, whereas perturbing responses to queries is called *output perturbation*.

Recently, there has been a growing demand for flexible on-line table generating servers (Thompson, Broadfoot and Elazar; Shlomo, Antal and Elliot, 2015). Typically such systems provide a menu-driven interface for producing confidentiality-protected user-defined frequency tables of counts. These on-line solutions of table generation increase the risk of inferential disclosure since tables can be manipulated and differenced and hence only a few statistical agencies have developed such systems. The server first assesses whether a table can be released based on a set of ad-hoc rules, such as thresholds on the population size and number of small cells, and then implements a confidentiality protection routine to each non-zero cell of the table prior to its release. With the increased disclosure risks, such confidentiality protection typically involves a perturbative method, such as rounding or additive noise to the cell count which leads us to consider the differential privacy framework.

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, (however, see Section 6 and references therein, showing the risk involved in doing this) 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 simply 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 official agencies, in the differential privacy framework the full description of the mechanism $\mathcal{M}$ can be made available along with $\mathcal{M}(D)$. The advantage of this practice is that knowledge of the mechanism allows the user to take the perturbation distribution into account in their analysis for data independent algorithms like those examined here, thereby avoiding potentially misleading conclusions that might arise from ignoring the perturbations.

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. The interactive setting is flexible and may require smaller perturbations, making the released data more useful. On the other hand it requires monitoring of all queries from all users for the whole time the data in question is in use, a task that may be too burdensome for most official agencies. 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 in parts as responses to queries that can be answered as functions of $\mathcal{M}(D)$. If only parts of the data are requested then it may be possible and efficient for the agency to perturb only relevant parts of the data.

In this paper we consider only the non-interactive setting, which is closer to the model table generating systems of interest to us. Therefore, we assume that the whole data set is perturbed, and then the whole or requested parts are released.

If the frequency table data D is treated simply as a set of frequency counts of 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). The extension of this methodology to handle the case where D also includes table margins, consisting of sums of cell counts, and more generally, cells that pertain to overlapping groups, will be considered in Section 7, along with relevant literature, such as Barak et al. (2007).

The rest of the paper is structured as follows. Section 2 presents some features of perturbations for a table generating server, which bear 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 (where two cells *overlap* if there is at least one individual appearing in both) and marginal counts in frequency tables and conclude with Section 8.

2. Perturbation of Frequency Tables

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). 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.

We suppose in this paper that the frequency tables contain population counts, from a census, survey, or administrative sources. Differential privacy treats the dissemination of census data and data arising from samples in the same way. However, in the latter case other considerations may arise, in particular due to the fact that when government agencies produce tables of estimated population counts based on sample survey data, 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.

2.1. Some Terminology and Notation

In this section we introduce some terminology and notation. First, we remark on our use of the terms confidentiality and privacy. This paper deals with the confidentiality of data held by a national statistical agency or other data custodian, as described in the SDC 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 the differential privacy 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. 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 marginal 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. A typical example is a situation where an agency holds a 10-way table, say, but will release only 3-way marginals, and the cells of these marginals (unperturbed) will comprise the list to be perturbed and released. In this case the list $\mathbf{a}$ will consist of all $K = \binom{10}{3} = 120$ three-way tables formed as marginals of the 10-way table. The set $\mathcal{A}$ of possible or potential lists $\mathbf{a} = (a_1, \ldots, a_K)$ is called the *universe* and may include lists with different values of K. We shall suppose that all elements of lists in $\mathcal{A}$ are non-negative integers. The universe is determined by the agency's decision on which parts of the data are to be released. If the agency knows the whole population from which the table to be released is drawn, and the way the data were collected, then the nature of the universe is clear. If not, then the agency has to rely on known ranges of the attributes and possible cell sizes, and perhaps some other information, when considering the universe. The universe plays a major role in providing privacy to microdata, and the case of histograms or tables is much simpler.

We consider a mechanism $\mathcal{M}(\cdot)$ on a universe $\mathcal{A}$ 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 $p(\mathbf{a}, \mathbf{b})$, the conditional probability that the list $\mathbf{a}$ is perturbed to $\mathbf{b}$. In general we shall assume that different cells are perturbed independently and by the same conditional distribution $p(a_k, b_k)$ for $k = 1, \ldots, K$, and then $p(\mathbf{a}, \mathbf{b}) = \prod_k p(a_k, b_k)$.

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 **a** above. In principle all perturbations of **a** could be applied in advance, and the whole perturbed list could be released, however for efficiency's sake perturbations may be applied when users submit queries, using a lookup table whose random values are drawn in advance. There is no monitoring of queries, and from the differential privacy point of view which holds in this paper, this is a non-interactive setting. 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. This guarantees that two cells based on the same group of individuals will be perturbed by the same seed to the same value, and in particular that if a cell is requested by two different users, they will receive the same perturbed output.

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 extreme, but it can be made to seem more realistic, as can be seen in examples in Willenborg and de Waal (2001) such as Table 6.3 on page 148.

The worst-case approach of differential privacy avoids having to consider different kinds of scenarios and how realistic they are. 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. We will demonstrate the breach of differential privacy later, after defining it formally in Section 4.1. 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 here, since we assumed different responses to the two queries). This is one indication why the universe $\mathcal{A}$ must be taken into account, and not just 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 sparseness, 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 avoiding small cells, such as counts of one, and setting a criterion to minimize risk for given utility. The remaining five properties address utility, 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.

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):

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

- 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

We review the basic definitions of differential privacy associated with releasing data sets consisting of lists of counts. 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 $\mathcal{M}$. Roughly speaking, differential privacy requires that the distribution of $\mathcal{M}(\mathbf{a})$ remains almost unchanged when any single individual is removed from the list $\mathbf{a}$. This guarantees that a user of $\mathcal{M}(\mathbf{a})$ cannot infer the presence of any particular individual in the data set, and therefore nothing can be learnt about any individual. We denote the range of the perturbation of $\mathbf{a} \in \mathcal{A}$ by $\mathcal{B}(\mathbf{a})$, that is, $\mathcal{B}(\mathbf{a}) = \{\mathbf{b} : \mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b}) > 0\}$. Then $\mathcal{B}(\mathbf{a}) \subseteq \mathcal{B}$, the range of $\mathcal{M}$, and when $\mathcal{B}(\mathbf{a})$ does not depend on $\mathbf{a}$, we have $\mathcal{B}(\mathbf{a}) = \mathcal{B}$. In this paper $\mathcal{A} = \mathcal{B}$ is assumed, that is, the perturbed list of frequencies has the same structure as the original one. 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.

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. The use of likelihood ratios to measure how much can be learnt about individuals after their data are perturbed by a randomized mechanism may be viewed as a generalization of the method of randomized response proposed by Warner (1965) to protect the privacy of a respondent's answer in a survey. The likelihood ratio could alternatively be viewed as a posterior odds ratio, or Bayes factor, from a Bayesian perspective (Berger, 1985).

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 $\mathcal{A}$, and all subsets $S \subseteq Range(\mathcal{M}) = \mathcal{B}$, 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 $\mathbf{a}'$

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

For small ε this guarantees that the distribution of the released data is not affected by the data of a single participant in the data set, and therefore he can feel safe that his participation and his particular profile is not reflected in the released data. In the words of Dwork (2006): "A mechanism satisfying this definition addresses concerns that any participant might have about the leakage of her personal information: even if the participant removed her data

from the data set, no outputs (and thus consequences of outputs) would become significantly more or less likely."

If there is a very large or small value of the ratio $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})/\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b})$ for given $\mathbf{a} \sim \mathbf{a}'$ and some observed $\mathbf{b}$, then a typical, albeit extreme scenario for a confidentiality breach is the following: suppose an intruder knows the whole original unperturbed list except for one targeted individual. Suppose the intruder wants to know on the basis of $\mathcal{M}(\mathbf{a})$ whether the targeted individual is in the given list $\mathbf{a}$ and if so, in which cell. Denoting the known (to the intruder) list without the target by $\mathbf{a}^*$, the intruder computes the distribution of all $\mathcal{M}(\mathbf{a}'_k)$ where $\mathbf{a}'_k$ is the list $\mathbf{a}^*$ with the targeted individual added to cell k in the list. Note that the intruder is computing the distribution of the output of $\mathcal{M}$ on lists that are known to him. Under DP(ε) with a small ε, all these distributions will be approximately the same, making inference on whether and where the target is in D difficult. Otherwise, an output $\mathcal{M}(\mathbf{a})$ that is very likely if $\mathbf{a} = \mathbf{a}'_k$ suggests that the targeted individual is in cell k, and his privacy is violated.

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

We briefly discuss the example of Bob from the second paragraph of Section 2.2. We consider a mechanism with non-degenerate independent perturbations for cells representing different sets of individuals, and satisfying the principle of 'same-participants-same-perturbation'. Consider the list $\mathbf{a}$ containing cells that contain Bob, and let a_1 be the frequency of workers whose salary exceeds s, and a_2 the frequency of workers whose salary exceeds s, and who have been working for more than one day, so it differs from a_1. Let $\mathbf{b} = \mathcal{M}(\mathbf{a})$ where the coordinates of $\mathbf{b}$ correspond to the perturbed coordinates of $\mathbf{a}$. Since the groups pertaining to a_1 and a_2 are different, the latter frequencies are perturbed independently, and as we obviously consider non-degenerate perturbations, it is easy to see that $\mathbb{P}(b_1 \neq b_2) > 0$. Let $\mathbf{a}'$ represent the same list apart from Bob, and set $\mathbf{b}' = \mathcal{M}(\mathbf{a}')$. Then the corresponding cells satisfy $a'_1 = a'_2$ and the 'same-participants-same-perturbation' principle implies $\mathbb{P}(b'_1 \neq b'_2) = 0$. For the set $S = \{\mathbf{b} : b_1 \neq b_2\}$ we see that (4.1) does not hold, and differential privacy does not hold for any ε.

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 DP(ε, δ) (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 $\mathbf{a}'$ in $\mathcal{A}$. The pa-

rameter 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. Clearly $\mathrm{DP}(\varepsilon, 0) = \mathrm{DP}(\varepsilon)$. 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 $\mathbf{a} \sim \mathbf{a}' \in \mathcal{A}$, where

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

and $0/0$ is defined to be 0. Closely related definitions can be found in Gotz et al. (2012); Machanavajjhala et al. (2008). We give the proof of the following Lemma for completeness, since it will be used later in the paper.

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

Proof. Suppose $\mathcal{M}$ satisfies (ε, δ)-probabilistic differential privacy, and let C denote the complement of G in $\mathcal{B}(\mathbf{a})$. 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. □

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 mechanism, even with $\delta > 1/N$, would not really enable this. 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. Extending this reasoning to the need to avoid releasing the records of a small number of database participants, typically it is desirable to have the value of δ smaller than the inverse of any polynomial in the size of the database (Dwork and Roth, 2014, p18). See Steinke and Ullman (2016) for consideration of some utility implications of reducing the value of δ. Another implication of (4.4) is that with probability δ, the whole data set may be released unperturbed. This can be considered a drawback in the definition (4.4) of $\mathrm{DP}(\varepsilon, \delta)$, suggesting

that δ should be small, and again, the DP(ε, δ) mechanisms described in this paper never release the whole unperturbed data set.

We refer to ε and δ as the DP parameters. The choice of their values should take into account a balance between confidentiality and utility, and perhaps the sensitivity of the information in data; for example, certain health variables may be much more sensitive to disclosure than, say, height and weight, and sensitive variables may require more protection, reflected by smaller DP parameters. Values like $\varepsilon = 0.1$ and $\delta = 0$ guarantee a likelihood ratio of $e^{0.1} = 1.1$ making it very hard to tell whether a particular individual is in the data set. However, it seems that in practice, larger values of the parameters will be required if we are to preserve the data utility. In some settings, the data custodian may consider that it is sufficient for the mechanism to ensure that no adversary could have more than limited evidence that a target individual's data is in the dataset. Evett et al. (2000) propose verbal summaries of ranges of values of a likelihood ratio, in particular interpreting values between 1 and 10 as 'limited evidence'. A threshold of 10 for the likelihood ratio, implying a value of ε of $ln(10) = 2.3$, would therefore ensure that such an objective is met, that is that that no adversary could have more than limited evidence that a target individual's data is in the dataset.

Machanavajjhala et al. (2008) consider data on commuting patterns of the population of the United States. In their experiments they use $\delta = 0.00001$ and $\varepsilon > 4$, which seem rather large. The Netflix dataset is considered by McSherry and Mironov (2009), where for application of the Laplace mechanism ε is chosen to be of the order of 1, and δ is zero. In all cases, the selection of ε (or δ) is a policy decision, not a statistical decision. However, policy makers are not experienced in choosing DP parameters in practical contexts, which points to the need for additional research. Our view regarding DP parameters is that even in cases where they are not small enough to guarantee privacy at a desirable level due to a compromise with utility, they are still useful in comparing different perturbation schemes and selecting an efficient one.

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 universe $\mathcal{A}$, 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.

Since the presence of an individual in a data set is unlikely to be inferred under small DP parameters, 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 private 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 DP 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 DP parameters to compare different perturbation mechanisms operating on the same file, thus avoiding this additional issue.

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 perturbation mechanisms which we shall apply 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 higher loss.

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 $\mathcal{B}$ as before, does not depend on $\mathbf{a}$, that is, every $\mathbf{b} \in \mathcal{B}$ 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

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

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

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

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

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

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, and we focus on such perturbations in order to keep the discussion within the framework of the ABS TableBuilder. We provide some references on dependent perturbations in Section 7. For example, implementation of the method proposed in Li et al. (2015) requires additional work from the releasing agency and/or the data user, which may be prohibitive in practice. Moreover, in Section 6 we discuss data analysis that takes the perturbations into account, assuming their distribution is known. Such an analysis, which is usually nontrivial, is facilitated by the assumption of independent perturbations and may become too complex otherwise. However, independent perturbations may have a negative effect on utility. For 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 will generally 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 the proof is given. We mention again that in Theorem 4.1 we assume that the range of $\mathcal{M}(\mathbf{a})$, denoted by $\mathcal{B}$, does not depend on $\mathbf{a}$. This result shows that under any such exponential mechanism 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 $\mathbf{a} \in \mathcal{A}$ and perturbed lists $\mathbf{b} \in \mathcal{B}$. Then $\mathcal{M}$ is* DP(ε).

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

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

Using $|u(\mathbf{a}, \mathbf{b}) - u(\mathbf{a}', \mathbf{b})| \leq \Delta u$, 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. □

Recalling that d denotes the maximum number of cells in which two neighbours, $\mathbf{a}$ and $\mathbf{a}'$ can differ, consider, for example, the case that all cells pertain to disjoint sets of individuals, as in a standard frequency table, and therefore $d = 1$. As shown in Section 5 we have for u_1 and u_2, as defined above, $\Delta u_1 = \Delta u_1 = 1$ and with perturbations truncated by m we have $\Delta u_2 = 2m + 1$. Therefore, apart from the assumption that $d = 1$, the exponential mechanism under these u_i does not depend on the structure of the data list, such as the cell sizes and the number of cells. This holds for any d with a suitable adjustment of Δu_i.

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 setting $|a_k - b_k| \leq m$ for some m, for all k. In this case, the range of $\mathcal{M}(\mathbf{a})$, denoted by $\mathcal{B}(\mathbf{a})$, 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. With an additional assumption on the utility u in Theorem 4.2, which holds for the examples considered in this paper, the exponent is not divided by 2 ($\eta = \varepsilon$ rather than $\varepsilon/2$ as in Theorem 4.1) implying 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 $\mathcal{B}(\mathbf{a}) = \mathcal{B}$.

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 $\mathbf{a} \in \mathcal{A}$ and perturbed lists $\mathbf{b}$ such that $|a_k - b_k| \leq m \leq \infty$ 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), with $\delta = 0$ when $m = \infty$.*

Proof. Let $\mathbf{a} \sim \mathbf{a}'$ be neighbouring lists and let $\mathbf{b} \in \mathit{Range}(\mathcal{M})$. Clearly, we can assume $\mathbf{b} \in \mathcal{B}(\mathbf{a})$ 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

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

where the second equality follows from the fact that the two sums in the denominators cancel since $\sum_{b:|b-a|\le m} e^{cg(b-a)} = \sum_{z=-m}^{m} e^{cg(z)}$ does not depend on a, and the last inequality follows from $|u(\mathbf{a},\mathbf{b}) - u(\mathbf{a}',\mathbf{b})| \le \Delta u$. 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 so that $p(a,b)=0$ for $|b-a|>m$. Here and in all our applications we assume also that $p(a,b)>0$ for $|b-a|\le m$. 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 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 DP(ε,δ), with

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

where the $p(m) = p(a,a+m)$, the probability that the perturbation takes its maximal value m, which for the symmetric utilities we consider equals the probability of $-m$. In fact, in the above case, if $\mathbf{a},\mathbf{a}'$ differ as above in cell j, and $b_j = a_j+1+m$, then

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

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 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 by 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 our main examples, u_1 and u_2 under the exponential mechanism. 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. 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 consider replacing 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 count a since counts obviously satisfy $a \geq 0$. 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. This would occur if the marginal cell value is perturbed separately from the internal cell values (see Section 7). Such post-processing after a DP perturbation would not affect the differential privacy property of the table.

4.5. Zero Cells

Structural zeros are cells representing combinations of attributes that are known to be impossible and have an expected value of zero, for example, in Table 1, children under the age of 14 are not in the labour force. There is no need to publish them since their value of zero is known a priori, and hence there is no need to perturb them 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. If zeros are not perturbed 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. 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 that zeros are not perturbed. It is straightforward to verify that DP(ε) cannot be satisfied if $p(0, 1) = 0$ and $p(1, 1) > 0$. On the other hand, if we relax to DP(ε, δ) then we need a condition such as $p(1, 1) \leq \delta$ which seems very undesirable for small δ. Thus differential privacy and unbiasedness are contradictory, unless release of negative values is allowed.

4.6. Summary of implications of the structural constraints discussed

The implications of the three different types of structural constraints considered in this section are summarized as follows.

First, it may be desirable to truncate the cell perturbations, as in Section 4.3. However, the increase in utility provided by the truncation is achieved at the cost of relaxing the confidentiality protection standard from DP(ε) to DP(ε, δ), where $\delta > 0$ depends on the truncation bound m and the utility function u.

As described in Section 4.5, structural zeros need not be published and hence do not need to be perturbed. We have demonstrated that non-structural zeros may be informative to intruders and therefore must be perturbed.

Finally, consider the treatment of cells that become negative after perturbation, as in Section 4.4. Such negative values should be released since they may be informative, and replacing them by zeros will introduce bias. Users should be advised to consider replacing them by zeros at a suitable stage of their analysis.

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. The three special cases are easy to explain and implement in practice.

For an alternative optimal perturbation mechanism that may perform better than the Laplace and Gaussian mechanisms, but is more complex, see Geng and Viswanath (2016).

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 appearing in (4.7) is attained in the case of Δu_1 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

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

where the normalizing constant is $C = \sum_{k=-\infty}^{\infty} e^{-\varepsilon k} = 1 + 2e^{-\varepsilon}/(1 - e^{-\varepsilon})$. Theorem 4.2 implies DP(ε).

Clearly one can view this perturbation as adding to each cell count a an independent random variable X satisfying $\mathbb{P}(X = x) = \frac{1}{C} e^{-\varepsilon|x|}$ for all integers x, so the perturbed cell is $a + X$. More generally, it is easy to see that any of the perturbations based on the exponential mechanism and additive utilities such as the above u_i, $i = 1, 2, 3$, results in adding to the data counts noise which is statistically independent of the data, and its distribution does not depend on the data and their distribution, as pointed out at the end of Section 4.2. In the case of u_1 the added noise has the Laplace distribution, and in the case of u_2, the normal, both discretised.

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

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

where $C_m = \sum_{k=-m}^{m} e^{-\varepsilon|k|} = 1 + 2(e^{-\varepsilon} - e^{-(m+1)\varepsilon})/(1 - e^{-\varepsilon})$. In this case, it follows from (4.10) that $\delta = e^{-\varepsilon m}/C_m$ and by Theorem 4.2 we obtain DP(ε, δ). Again, negative perturbed values can be replaced by zero, maintaining the same level of differential privacy. For $\varepsilon = 1$ and $m = 10$ we obtain $\delta = 0.00002$ and for $\varepsilon = 0.5$ and $m = 10$, $\delta = 0.0016$. 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 to be a very efficient choice, better than the normal perturbations of the next section, in the sense of providing higher utility for a given DP level, as indicated also by our simulations and those of Liu (2017).

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 $|a_k - b_k| \le m$ 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

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

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 $\varepsilon = 1$ ($\varepsilon = 0.5$) and $m = 10$ we have $\delta = 0.001$ ($\delta = 0.008$).

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. See, e.g., Cover and Thomas (2006, Chapter 12) for a discussion of maximum entropy distributions. The fact that Laplace perturbations seem to perform better than Normal, suggests that the ABS TableBuiler 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

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

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. 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

Since small cells are considered to be particularly risky, we first compare the utility of the Laplace and Normal perturbations for a given DP level, when negative values are replaced by zero (and therefore the resulting perturbation depends on the original value). In Table 2 we calculate the probability of obtaining a perturbed value in an interval range of $\pm$ 0 to $\pm$ 4 of the original value, when the original values are 0 to 5 and over, $\varepsilon = 1.5$ and $\varepsilon = 0.5$. 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

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

truncated at $m = 10$. The choice of the above parameters for the purpose of this introductory article is somewhat arbitrary, as our goal is to demonstrate how perturbation mechanisms can be compared and not to provide a comprehensive study. However, we chose values which demonstrate well the privacy utility balance and may be considered reasonable choices.

From Table 2, it is clear that the Laplace perturbations are smaller (in probability) and thus have higher utility under differential privacy with the given ε and δ. These results are consistent with those in Liu (2017). 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.

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 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 $\text{Mult}(N, p_{ij})$. We carried out 100 perturbations on each table and

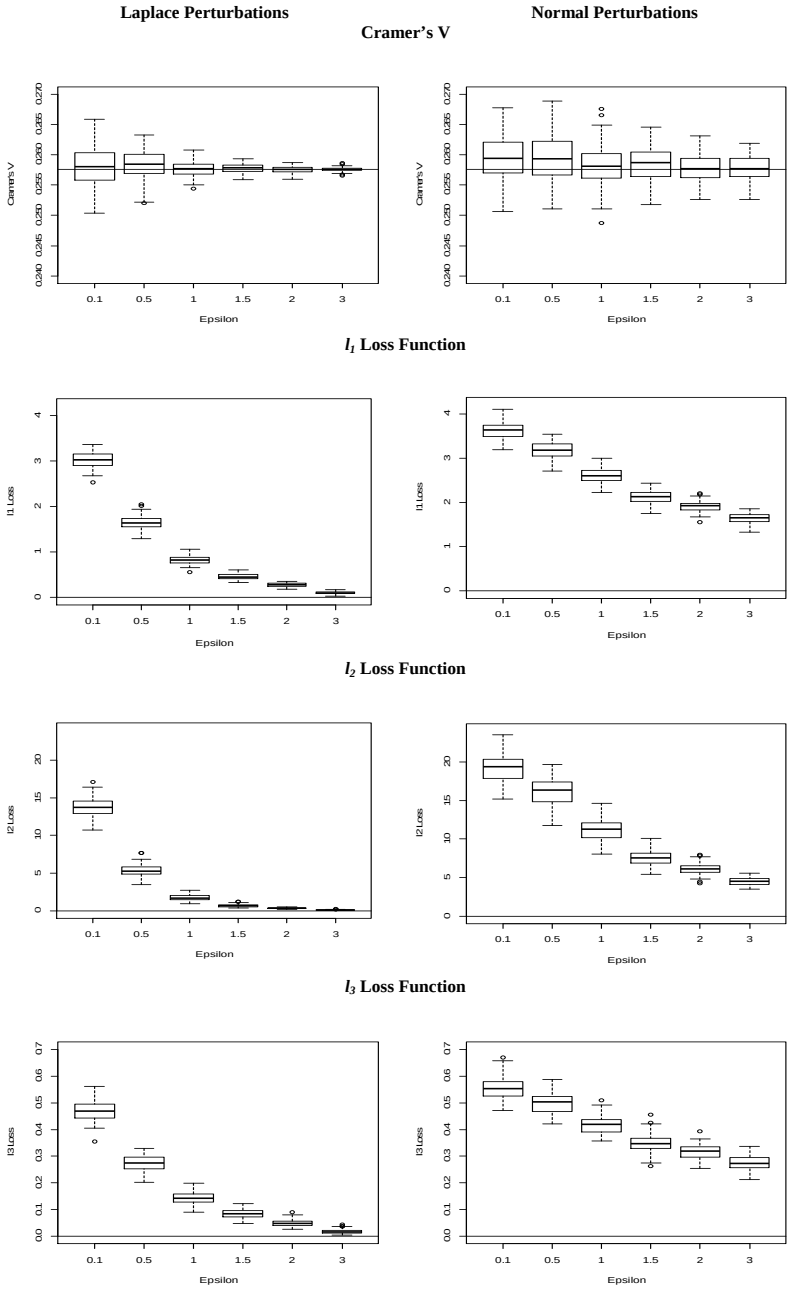

FIG 1. *Values of Cramer's V and three loss functions over 100 perturbation repetitions for each ε for Table 1*

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 tables. 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 by now. 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 perturbation distribution into account

Poisson and multinomial distributions of counts and log-linear models are standard in the analysis of frequency tables. See, e.g., Bishop, Fienberg and Holland (1975, Section 3.2) for a classical reference. Under such models, with additive perturbations having a known distribution, it is natural for the data user to test hypotheses using the model and the perturbation distribution. The example of testing independence in Section 6.1 shows how this can be done with a valid significance level and a power that varies with the DP parameters. On the other hand, it is shown here and in Figures 2 and 3 that applying a standard ('naïve') Chi-square test to the data as if it were not perturbed may lead to a very wrong level of significance, and hence to wrong conclusions, even when the sample sizes are such that standard asymptotic theory (with no perturbations) applies. Unfortunately, since agencies release perturbed tables which have an appearance similar to that of the original table (and this is why the agency may avoid releasing negative cells, for example), it is tempting to ignore the perturbations and analyze the released data 'naïvely'. A further example of a goodness-of-fit test for a single binary attribute is given in Section 6.2.

Uhler, Slavković and Fienberg (2013) and Fienberg, Rinaldo and Yang (2010) have shown that perturbations can lead to unreliable conclusions in the analysis of tables if their presence is ignored, and proposed methods to overcome this problem. Methods of improving the performance of tests of independence in two-way tables under such perturbation have been proposed by Wang, Lee and Kifer (2017). Karwa, Kifer and Slavković (2015) considered working with the true likelihood as we shall do, but they consider that in most cases the likelihood is intractable and that approximate computational methods are needed. Karwa et al. (2016) develop a likelihood-based approach to inference for a particular model of an undirected graph. Charest (2010) suggests a Bayesian approach that accounts for the effects of additive noise on inferences in the context of differential privacy. This area seems to be open to further research.

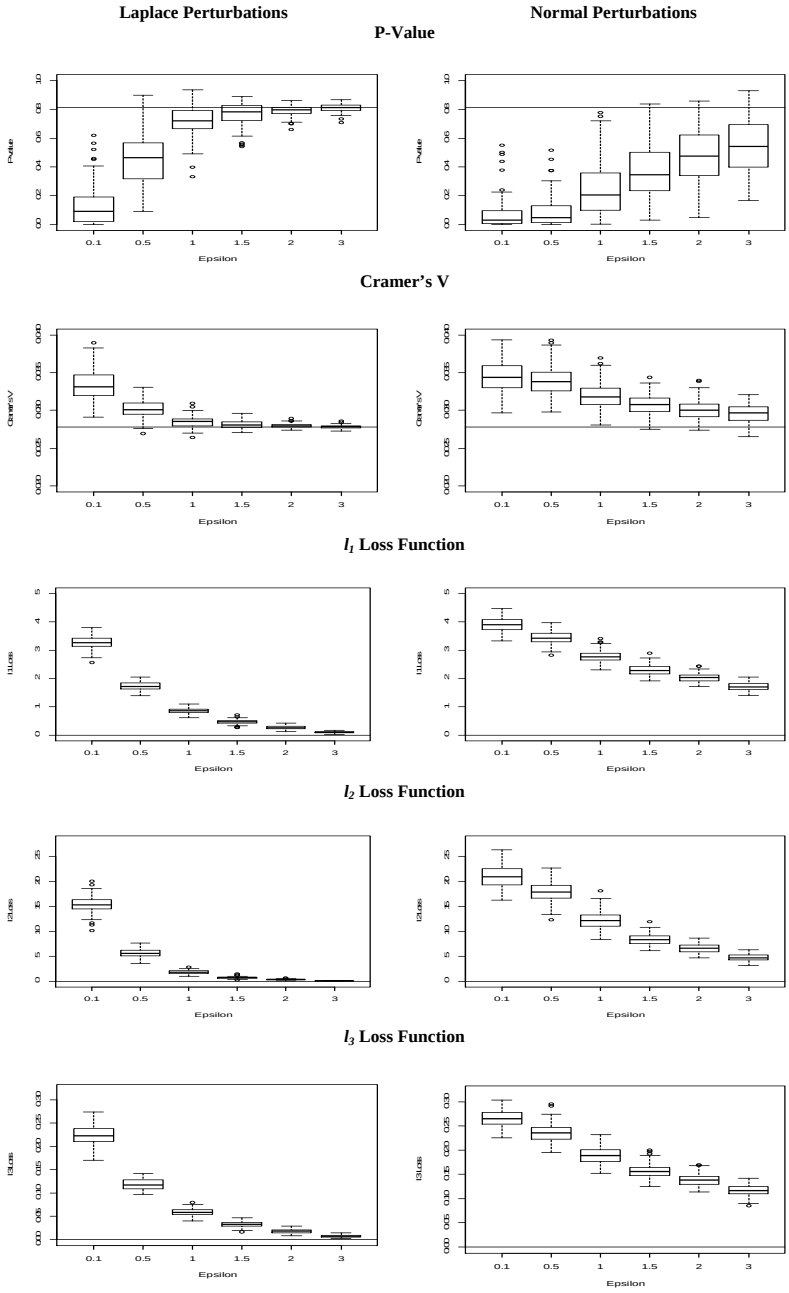

FIG 2. *p-values, Cramer's V, and three loss functions over 100 perturbation repetitions for each ε for the small independent table (average cell size=100)*

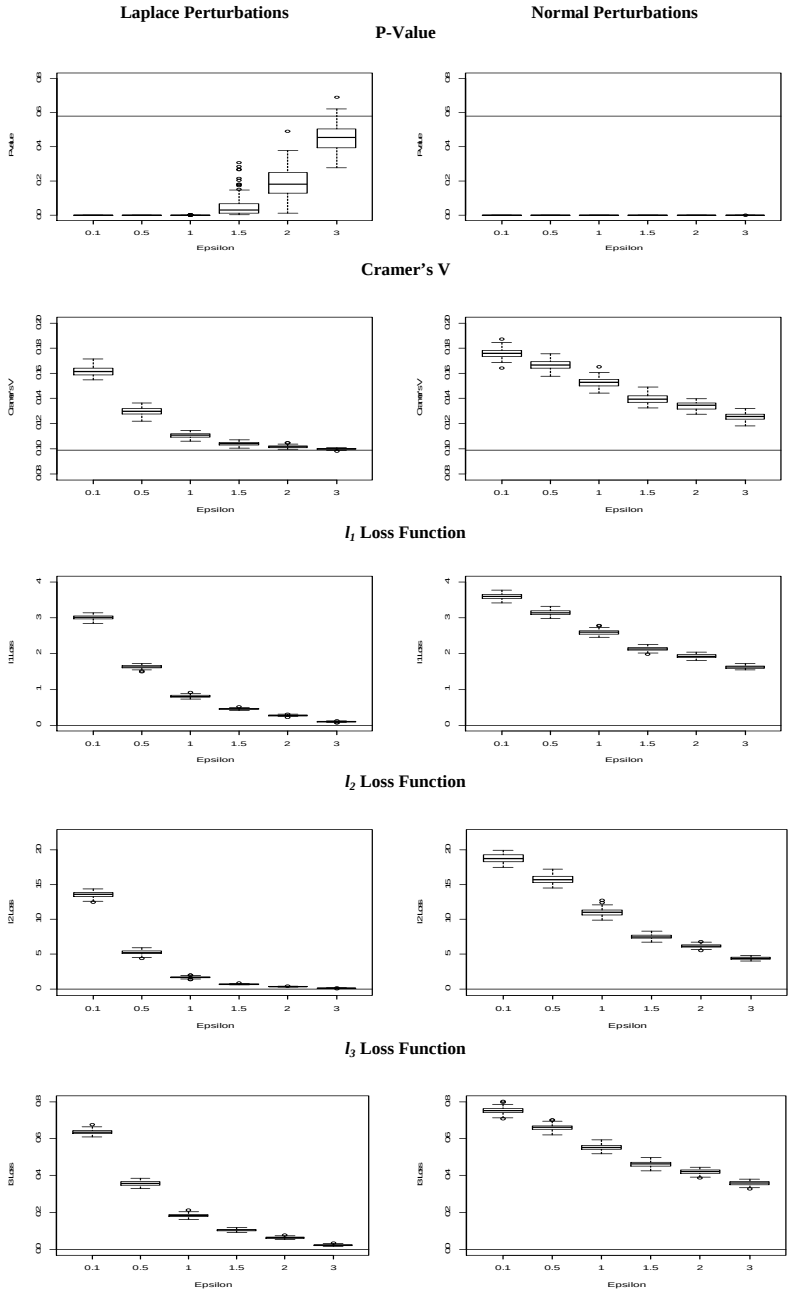

FIG 8. *p-values, Cramer's V, and three loss functionsover 100 perturbation repetitions for each ε for the large independent table (average cell size=10)*

6.1. Testing for independence

Consider an $r \times c$ frequency table $\mathbf{a}$ that is to be released with truncated Laplace perturbation L_{ij} applied independently to each cell, where $-m \le L_{ij} \le m$. Assume that the table consists of independent Poisson(μ_{ij}) entries, a standard model in frequency table analysis. Specifically, we assume that the data $\mathbf{a}$ consists of a table of independent Poisson(μ_{ij}) entries. With additive Laplace noise (see comment in the paragraph following (5.1)) the released table is $\mathbf{X} = \mathbf{a} + \mathbf{L}$ where $\mathbf{L}$ is a $r \times c$ matrix of truncated independent Laplace variables. Independence of the attributes amounts to the hypothesis $H_0 : \log \mu_{ij} = \eta + \alpha_i + \beta_j$.

The standard Chi-square or the asymptotically equivalent likelihood ratio test remain correct asymptotically when applied 'naïvely' to the table with perturbations, when the counts of the original tables increase to infinity (and as pointed out in Sections 4.2 and 5.1 above, the noise distribution remains the same). However, as pointed out in Wang, Lee and Kifer (2017) and references therein "the p-values produced by this method are extremely biased and will often lead to false conclusions." In particular, using a standard test and ignoring the noise in finite samples may lead to tests with a much higher significance level than could be claimed by asymptotic theory. The simulations below show that this is indeed the case, and here we discuss a very natural approach: to compute the likelihood ratio test taking the perturbation distribution into account. Our simulations show that in this case the asymptotic distribution leads to a correct significance level, and that the tests have a reasonable power for certain DP parameters, close to that of the standard tests on the original data under reasonable conditions (depending of course on the data, and the noise parameters).

Our goal is to test the above H_0, taking the perturbation distribution into account. The likelihood ratio test statistic is the ratio of two maximized likelihoods

$$\max_{\mu_{ij}} L_{\mathbf{X}}(\{\mu_{ij}\}) \Big/ \max_{\log \mu_{ij} = \eta + \alpha_i + \beta_j} L_{\mathbf{X}}(\{\mu_{ij}\}) \tag{6.1}$$

where apart from a constant the likelihood $L_{\mathbf{X}}(\{\mu_{ij}\})$ is defined by

$$L_{\mathbf{X}}(\{\mu_{ij}\}) \;=\; \prod_{ij} \sum_{\ell_{ij}=-m}^{\min\{x_{ij},m\}} P_{\mu_{ij}}(x_{ij} - \ell_{ij}) e^{-\epsilon|\ell_{ij}|}, \tag{6.2}$$

and $P_\mu(x) = e^{-\mu}\dfrac{\mu^x}{x!}$ is the Poisson probability that arises from the model on the data; also, x_{ij} are the entries of the released perturbed table, and $e^{-\epsilon|\ell_{ij}|}$ the (unnormalized) Laplace probabilities. In the numerator of the likelihood ratio statistic (6.1), the max is over all μ_{ij}, and in the denominator we need to maximize the function of (6.2) over the parameters η, α_i and β_j, where we set

$$\log \mu_{ij} = \eta + \alpha_i + \beta_j \quad \text{and} \quad \alpha_1 = \beta_1 = 0.$$

For our example, we generate 10×10 tables, so we have 19 parameters to estimate. The maximization was done numerically using SAS procedure NLP.

TABLE 3
Simulation results for testing independence.

Table Type		Para-meters	% p-value ≤0.05	Mean (S.E.) Test Statistic	Mean (S.E.) p-value	Para-meters	% p-value ≤0.05	Mean (S.E.) Test Statistic	Mean (S.E.) p-value
Independent	Original		5.0	81.6 (0.395)	0.487 (0.009)		6.0	81.7 (0.423)	0.487 (0.009)
Attributes	Naive	ε=0.1	86.7	124.5 (0.616)	0.027 (0.002)	ε=0.1	53.3	105.1 (0.524)	0.123 (0.006)
	LR test	m=10	3.0	78.6 (0.388)	0.555 (0.009)	m=7	4.0	80.1 (0.400)	0.521 (0.009)
Dependent	Original	δ=0.0283	79.3	118.8 (0.587)	0.044 (0.003)	δ=0.0470	87.3	124.8 (0.620)	0.027 (0.003)
Attributes	Naive		99.6	162.1 (0.792)	0.001 (0.000)		98.3	149.2 (0.742)	0.004 (0.001)
	LR test		51.0	103.6 (0.526)	0.140 (0.006)		73.3	114.5 (0.583)	0.066 (0.004)
Independent	Original		5.8	81.7 (0.414)	0.485 (0.009)		4.7	81.4 (0.395)	0.485 (0.009)
Attributes	Naive	ε=0.5	25.4	92.9 (0.476)	0.274 (0.008)	ε=0.5	18.7	90.8 (0.435)	0.299 (0.008)
	LR test	m=10	6.9	82.5 (0.419)	0.467 (0.009)	m=7	5.3	82.0 (0.391)	0.473 (0.009)
Dependent	Original	δ=0.0017	82.1	118.3 (0.551)	0.041 (0.003)	δ=0.0076	81.3	119.6 (0.591)	0.040 (0.003)
Attributes	Naive		91.3	129.3 (0.601)	0.017 (0.002)		91.0	128.6 (0.627)	0.018 (0.002)
	LR test		76.3	114.8 (0.532)	0.054 (0.004)		76.9	116.2 (0.567)	0.051 (0.003)

We first generated 1000 10×10 tables under H_0 with α_i and β_j drawn each time from the Uniform$(-0.5, 0.5)$ distribution, and $\eta = 4$, and then added to each table independent Laplace noise with $\varepsilon = 0.1$ or 0.5 with truncation between $-m$ and m for $m = 10$ and 7. The values of δ are obtained from formulas of Section 5.1 leading to DP(ε, δ). The average cell size was about 50, which is not a small sample.

We also generated 1000 10×10 tables where the attributes are dependent, using the Poisson model with $\mu_{ij} = \eta + \alpha_i + \beta_j + 0.7\gamma_{ij}$ where also $\gamma_{ij} \sim$ Uniform$(-0.5, 0.5)$. The results for both independent and dependent attributes are presented in Table 3, where 'Original' refers to applying a standard likelihood ratio test to the unperturbed table, 'Naive' stands for applying the same likelihood ratio test to the perturbed data and ignoring the perturbations, and 'LR test' is the likelihood ratio test of (6.1) that takes the perturbation distribution into account.

Out of 1000 repetitions for each set of parameter values, the table provides the percentage of test statistics whose p-value according to the (asymptotic) Chi-square distribution with 81 degrees of freedom is below 0.05. For example, for independent attributes, $\varepsilon = 0.1$ and $m = 10$ the Original gave exactly 5% below critical value so here the asymptotic significance level was attained perfectly by the simulations. The naive test showed almost 87% below 0.05, meaning that its level of significance is about 0.87, which is extremely high, rendering this test very unreliable for the given sample size. The LR test of (6.1) showed a level of significance of 3%, suggesting that the approach that takes the perturbation distribution into account is reliable with the present sample size. The power of the test on the unperturbed data under the dependence model we chose was 79% whereas the LR test had a smaller power of 51% showing that the perturbations reduce the power. When changing ε to 0.5, the significance level of the naive test was 0.25 which is still unreliable. The LR test had a significance level of 0.058, and the power was 0.76, very close to that of the original unperturbed data of 0.82.

Clearly, more extensive simulations and theoretical study is required, however from these and related simulations not presented here we conclude that when the sample sizes are such that the standard asymptotic theory applies for unperturbed data, it also applied to the proposed LR test of (6.1) in determin-

ing the correct significance level. However, the naive test that applies standard theory and ignores perturbations is useless unless the sample size is very large. Simulations show, for example, that with $\varepsilon = 0.1, m = 10$, and an average count of about 400 per cell, the naive test still has a significance level of about 0.12, rather than the asymptotic value of 0.05, while the LR test achieved a level of about 0.05.

6.2. Testing Goodness-of-Fit for a Binomial distribution

Consider a list consisting of a single cell, with $\mathbf{a} = a_1$, $a_1 \sim \text{Binomial}(N, p)$ and N known. If a_1 is the number of individuals having a certain property, then $\Delta u_1 = 1$. The perturbed data released is $X = a_1 + L$, where L is a Laplace perturbation truncated by m as in (5.2). The likelihood of an observation X is a function of p:

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

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 below (5.2) show that $\delta = 0.02$ so we have $\text{DP}(0.5, 0.02)$. The plot on the left of Figure 4 shows that for testing $H_0 : p = 0.5$ the statistic values are mostly small, and when testing $H_0 : p = 0.7$, 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

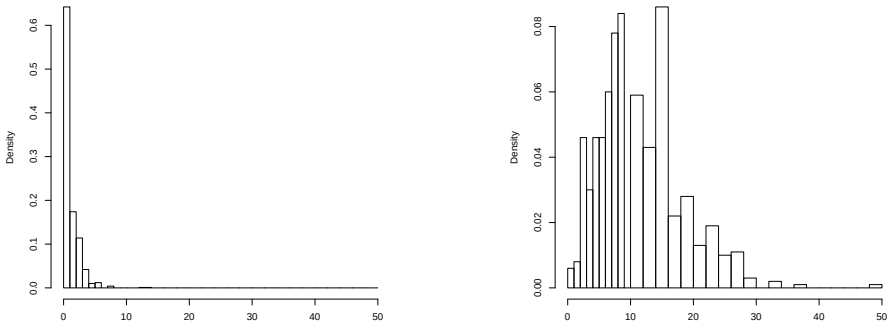

FIG 4. *Histogram of 500 2 log(likelihood ratio) tests when* $N = 80$, $p = 0.5$, $\varepsilon = 0.5$, *and* $H_0 : p = 0.5$ *(left), and* $H_0 : p = 0.7$ *(right) is tested*

be seen from Table 3. Thus, allowing a larger perturbation range, that is ± 10 rather 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 resembles a Chi-square distribution with 1 degree of freedom, 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, for example, if the list includes margins as well as interior cells in a multi-way frequency table, or when the list contains several tables drawn from the same population or overlapping populations. 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 marginal cells 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 marginal tables 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.

An attractive property in the release of non-overlapping cell counts using the perturbations schemes of this paper is that if $DP(\varepsilon)$ holds for each cell, when

perturbed independently, then $DP(\varepsilon)$ also holds for the whole table, irrespective of the number of cells in the table. To see this, note that when cells are non-overlapping, removing or adding an individual affect only one cell and therefore d and Δu_i are determined by a single cell. This is no longer the case for overlapping cells, when achieving $DP(\varepsilon)$ for the whole table will, generally, require greater noise for each entry in the table, with the amount of required noise increasing with d. We now focus on the case where the list $\mathbf{a}$ includes both interior cells and some margins so that $d \geq 2$, and Laplace perturbation is applied to the whole list. We have $\Delta u_1 = d$ and 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 large d 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.

In this section we shall consider various ways of 'spending' a privacy budget of ε if $DP(\varepsilon)$ is to hold for the whole table. Given our focus on the scenario of an official agency implementing an online flexible table generator, we shall only consider straightforward and practical approaches of applying Laplace perturbation independently to the table entries, with possibly varying levels of noise applied to different parts of the table. There is further literature on algorithms which do enable reduced levels of noise to be applied for a given privacy budget by perturbing the interior cells and margins in dependent ways, using the fact that the margins are linear combinations of the interior cells. Several such proposed algorithms are examples of a matrix mechanism (Li et al., 2015). Barak et al. (2007) is an early example, where the perturbation is applied to a transformation of the list using a Fourier basis. Hay et al. (2016) find that the multiplicative weights exponential mechanism of Hardt, Ligett and McSherry (2012) out-performed a number of instances of the matrix mechanism, although this algorithm produces synthetic rather than perturbed tables. Gaboardi et al. (2016) propose a related Dual Query approach for practical applications with high dimensional tables. We shall not pursue such alternative options here, however, due to our focus on flexible table generators. A further potential concern that we shall consider in the perturbation of overlapping cells is that the released table may be inconsistent in the sense that the perturbed margins do not coincide with the relevant sums of the perturbed interior cells, though they will generally be close. In further literature on algorithms which perturb overlapping cells in dependent ways, it is found that the objectives of consistency and reduced levels of noise need not conflict and can be achieved jointly (Barak et al., 2007; Hay et al., 2010).

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 only interior cells are released, then $d = 1$, and if each cell is perturbed by Laplace noise with a given ε, see (5.1), we have DP(ε) and the perturbations have a SD close to $\sqrt{2}/\varepsilon$. In this case the standard deviation of the sum of the perturbations in

a $k-$dimensional marginal table will be proportional to $\sqrt{2C^{t-k}}/\varepsilon$. Consider a 4-way table with $C = 10$, for example. If only interior cells are perturbed then $d = 1$ and the perturbation SD in each cell is $\sqrt{2}/\varepsilon$. Suppose now that the agency releases all 2-dimensional marginal tables. If they are obtained by summing perturbed interior cells, the SD of the perturbation for each cell of a 2-dimensional marginal is proportional to $\sqrt{2C^2}/\varepsilon = \sqrt{2 \cdot 10^2}/\varepsilon \approx 14/\varepsilon$. If only 2-dimensional marginal are perturbed then it is easy to see that $d = 6$ and the SD of each cell in these marginals is $\sqrt{2}d/\varepsilon = \sqrt{2}6/\varepsilon = 8.5/\varepsilon$. If all cells and marginals are perturbed and released then $d = 2^4 - 1$ and then the perturbation SD in each released cell, including cells of 2-dimensional marginals is $\sqrt{2}d/\varepsilon = \sqrt{2}(2^4 - 1)/\varepsilon \approx 21/\varepsilon$, so for such marginals the scheme that perturbs only interior cells is preferable to perturbing all cells in the sense of having a smaller SD, and the smallest SD is achieved by perturbing only 2-dimensional marginals. 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, and release $(\mathcal{M}_1, \dots, \mathcal{M}_k)(\mathbf{a}) := (\mathcal{M}_1(\mathbf{a}), \dots, \mathcal{M}_k(\mathbf{a}))$ which 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$(\varepsilon_i, \delta_i)$ 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 $\mathcal{B}_i$ for $i = 1, 2$ and $S = S_1 \times S_2 \subseteq \mathcal{B} := \mathcal{B}_1 \times \mathcal{B}_2$ 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

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

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. Other more sophisticated composition results can be found in Dwork, Rothblum and Vadhan (2010); Dwork and Rothblum (2016); Abadi et al. (2016); Kairouz, Oh and Viswanath (2017).

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:

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

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 naturally obtain again DP(ε).

However, 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). It should be noted that large high-dimensional tables, which arise in certain surveys, will often be very sparse, and it does not seem useful to perturb every cell. In fact, a common practice of agencies in this situation is to merge cells and to reduce the dimension and hence sparseness, and then to perturb the resulting

list. Obviously, this incurs loss of information. The development of practical methods for the confidentiality-protected release of such tables seems to be a worthwhile direction for research.

The above discussion indicates that the data releasing agency has a great amount of flexibility in deciding 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 practical perturbation schemes that resemble ones currently used by some official agencies when releasing frequency tables, with the goal of assessing how random perturbations, along with other common practices of these agencies, protect confidentiality in terms of the differential privacy standard. We have seen how this approach can highlight specific issues, such as the effect of truncation, not perturbing zeros, or 'same participants-same perturbation' schemes. We focused on a few alternative perturbation mechanisms and the Laplace perturbation seems to have a clear advantages in terms of the utility of the resulting tables for a given level of confidentiality protection. The extent to which the perturbations damage the value of tables for analysis will depend on user needs and it is hard to draw any general conclusions. Our numerical work in Section 5 compared the properties of a small number of algorithms that we believe would be likely candidates for practical implementation by an official agency. See Hay et al. (2016) for a framework for undertaking a comprehensive evaluation of differentially private algorithms and for the findings of such an evaluation of a broader range of algorithms for answering 1- and 2- dimensional range queries over 27 datasets.

Maximum entropy perturbation subject to variance constraints is one existing criterion for selecting perturbations in statistical disclosure control, but the implied approximately normal perturbations did not perform well in our assessment. We found that insisting on releasing only non-negative perturbed frequencies may result in loss of utility, without a well defined gain in confidentiality protection. Other desiderata that have been proposed for perturbation, for example that perturbed frequencies be unbiased for the true frequencies and that perturbations be truncated by a specified bound, may be contradictory, and compromises of these criteria may be desirable.

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 the desirability of making the perturbation mechanism and its parameters 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 may be

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 simple examples. In practice, the computational challenges may be severe for the kinds of tables released by national statistical agencies, but this is an area for further research. We also noted that testing independence on perturbed data using 'naïve' test statistics that ignore the perturbations will be wrong for reasonable sample sizes, even if asymptotically justified.

Another area needing further research relates to tables based on sample data rather than on population counts. 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 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 official 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 over-protection. The differential privacy literature proposes interactive query submission and monitoring for all users online, 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 potentially be automated. Further research on interactive dissemination by official agencies and its implications seems to be needed.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* 308–318. ACM.

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 solu-

tion to contingency table release. In *Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)* 273–282.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*, 2nd ed. ed. *Springer Series in Statistics*. Springer, New York.

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.

Charest, A.-S. (2010). How can we analyse differentially-private synthetic datasets? *Journal of Privacy and Confidentiality* **2** 21-33.

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. (2012). New data dissemination approaches in old Europe–synthetic datasets for a German establishment survey. *Journal of Applied Statistics* **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. (2006). Differential Privacy. In *ICALP 2006* (M. Bugliesi, B. Preneel, V. Sassone and I. Wegener, eds.). *Lecture Notes in Computer Science* **4052** 1-12. Springer, Heidelberg.

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

Dwork, C., Rothblum, G. N. and Vadhan, S. (2010). Boosting and differential privacy. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on* 51–60. IEEE.

Dwork, C. and Rothblum, G. N. (2016). Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*.

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.

Evett, I., Jackson, G., Lambert, J. and McCrossan, S. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice* **40** 233–239.

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.

Gaboardi, M., Arias, E. J. G., Hsu, J., Roth, A. and Wu, Z. S. (2016). Dual Query: practical private query release for high dimensional data. *Journal of Privacy and Confidentiality* **7** 53-77.

Geng, Q. and Viswanath, P. (2016). The optimal noise-adding mechanism in differential privacy. *IEEE Transactions on Information Theory* **62** 925-951.

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 logsa comparative study of privacy guarantees. *IEEE Transactions on Knowledge and Data Engineering* **24** 520–532.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. and Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science* **339** 321–324.

Hardt, M., Ligett, K. and McSherry, F. (2012). A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems* 2339–2347.

Hay, M., Rastogi, V., Miklau, G. and Suciu, D. (2010). Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment* **3** 1021–1032.

Hay, M., Machanavajjhala, A., Miklau, G., Chen, Y. and Zhang, D. (2016). Principled evaluation of differentially private algorithms using DPBench. In *Proceedings of the 2016 International Conference on Management of Data* 139-154. ACM.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F. and Craig, D. W. (2008). Resolving individuals contributing trace amounts of

DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics* **4** e1000167.

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.

Kairouz, P., Oh, S. and Viswanath, P. (2017). The composition theorem for differential privacy. *IEEE Transactions on Information Theory* **63** 4037–4049.

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*.

Karwa, V., Slavković, A. et al. (2016). Inference using noisy degrees: Differentially private β-model and synthetic graphs. *The Annals of Statistics* **44** 87–112.

Li, C., Miklau, G., Hay, M., McGregor, A. and Rastogi, V. (2015). The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB Journal* **24** 757–781.

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

Liu, F. (2017). Generalized gaussian mechanism for differential privacy. *arXiv preprint arXiv:1602.06028v5*.

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 Mironov, I. (2009). Differentially private recommender systems: building privacy into the net. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* 627–636. ACM.

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.

Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large datasets. In *Proc IEEE Security & Privacy Conference* 111–125.

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.

Shlomo, N. and Young, C. (2008). Invariant post-tabular protection of census frequency counts. In *International Conference on Privacy in Statistical Databases* 77–89. Springer.

Steinke, T. and Ullman, J. (2016). Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality* **7** 3-22.

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *J Law Med Ethics* **25** 98 – 110.

Thompson, G., Broadfoot, S. and Elazar, D. 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 / Revue Internationale de Statistique* **70** 269-288.

Wang, Y., Lee, J. and Kifer, D. (2017). Revisiting Differentially Private Hypothesis Tests for Categorical Data. *arXiv preprint arXiv:1511.03376v4*.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **60** 63–69.

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.