Measuring Disclosure Risk and Data Utility for Flexible Table Generators

Natalie Shlomo, Laszlo Antal and Mark Elliot[1]

[1]University of Manchester, e-mail: natalie.shlomo@manchester.ac.uk, laszlo.antal@postgrad.manchester.ac.uk, and mark.elliot@manchester.ac.uk

Abstract

Statistical agencies are considering making more use of the internet to disseminate census tabular outputs through on-line flexible table generating servers that allow users to define and generate their own tables. The key questions in the development of these servers are what data should be used to generate the tables and what statistical disclosure control (SDC) method should be applied. For flexible table generating, the server has to measure the disclosure risk in the table, apply the SDC method and then reassess the disclosure risk. SDC methods may be applied either to the underlying data used to generate the tables and/or to the final output table generated from original data. Besides disclosure risk, the server should provide measures of information loss comparing the perturbed table to the original table. In this paper, we examine the development of a flexible table generating server and compare different SDC methods. We propose measures for disclosure risk and data utility that are based on Information Theory.

Keywords: Statistical Disclosure Control, Entropy, Hellinger Distance

Acknowledgement: The project is funded by the EU 7th framework infrastructure research grant: 262608, Data Without Boundaries (DwB) and the ONS-ESRC funded PhD studentship (Ref. ES/J500161/1).

1. Introduction

Driven by demand from policy makers and researchers for specialized and tailored census tables, many statistical agencies are considering using flexible table generating servers that allow users to define and generate their own tables. The United States Census Bureau and the Australian Bureau of Statistics have developed such servers for disseminating census tables. Users access the servers via the internet and define their own table from a set of pre-defined variables and categories typically from drop down lists.

The key questions in the development of these servers are what data should be used in the background for producing the tables and what method of statistical disclosure control (SDC) should be applied. The Computer Science literature has contributed much research on the theory of guaranteeing privacy in outputs from query-based systems based on perturbative SDC methods under specific parameterization (Dinur and Nissim, 2003) which can inform these new modes of data dissemination.

For the dissemination of census tables from European member states, Eurostat is developing a table generating server through the European Census Hub Project. Each member state is required to produce a fixed set of pre-defined multi-dimensional tables (hypercubes) containing their country's census counts: 19 hypercubes at the geography level of LAU2 and over 100 hypercubes at the geography level of NUTS2, cross- classified with as many as six other census variables. The hypercubes will then be used as the underlying data behind the flexible table generating server. The platform will allow comparative tables across member states and the combination of census data from multiple member states. The fixed set of hypercubes allow harmonization of census results and have the additional advantage that they

provide some a priori protection against disclosure since no data below the level of the cells of the hypercube can be released.

When selecting the SDC method for a flexible table generating server, there are two approaches: apply SDC to the underlying data so that all tables generated in the server are deemed safe for dissemination (pre-tabular SDC), or produce tables directly from original data and apply the SDC method to the final tabular output (post-tabular SDC). Although sometimes a neater and less resource intensive for data from a single source, the pre-tabular approach is problematic for the dissemination of European Census data for two reasons. Firstly all member states would have to agree on a common SDC method in order to provide consistent hypercubes across member states. For example, if one member state employs a rounding algorithm whilst another member state employs cell suppression, there will be little utility in a table that is generated based on both member states' data. Secondly, when combining data which has been separately disclosure controlled we compound the SDC impact, for example aggregating rounded counts exacerbates the data utility impact and overprotects the data. With the second approach of protecting only the final tabular output, SDC methods are not compounded.

For flexible table generating, the server has to measure the disclosure risk in the original table, apply an SDC method and then reassess the disclosure risk. There are two types of disclosure risks in census tables: identity disclosure where small cell counts may lead to an identification, and attribute disclosure where rows/columns contain empirical (real) zeros and only a small number of cells are non-zero. This leads to the ability to learn attributes about an individual or group of individuals. Differencing tables generated through the server can lead to residual tables that are more susceptible to the above disclosure risks and to the

reconstruction of individual records. After the table is protected, the server should also calculate data utility impact of the disclosure control by comparing the perturbed table to the original table.

In this paper, we compare both pre- and post-tabular SDC methods. The comparison is made through disclosure risk and data utility measures which must be able to be calculated 'on-the-fly' within the table generating server. We propose new disclosure risk and data utility measures based on Information Theory (IT).

Section 2 describes the hypercube that will be used in our simulation study. The SDC methods for the study are described in Section 3 and the development of a table generating server is discussed in Section 4. The disclosure risk and data utility measures are presented in Section 5. The results of the comparison of SDC methods are presented in Section 6 with a discussion in Section 7.

2. Simulation Hypercube

To investigate and compare SDC methods for a table generating server, we simulate a hypercube with an underlying population of 1,500,000 individuals for two NUTS2 regions. The variables defining the hypercube follow the Eurostat specification for one of the hypercubes:

- NUTS2 Region - 2 regions
- Gender – 2 categories
- Banded age groups – 21 categories
- Current Activity Status – 5 categories

- Occupation – 13 categories
- Educational attainment – 9 categories
- Country of citizenship – 5 categories

From the UK Census 2001, we calculated cell proportions from available published tables, multiplied the proportions by the 1,500,000 individuals in the population and calculated all cross-classified proportions of the table through iterative proportional fitting to produce the final synthetic hypercube. The hypercube used in the simulation study had 245,700 cells. The distribution of cell counts is skewed with a large proportion of zero cells as seen in Table 1.

The distributions in the synthetic hypercube were compared to those obtained from real hypercubes produced by member states Italy and Estonia at the NUTS2 region level according to the above specification and similar distributions were obtained.

Table 1: Distribution of Cell Counts in the Synthetic Hypercube

Cell Value	Number of Cells	Percentage of Cells
0	226,939	92.36%
1	4,028	1.64%
2	2,112	0.86%
3-5	2,964	1.21%
6-8	1,664	0.68%
9-10	720	0.29%
11 and over	7,273	2.96%
Total	245,700	100.00%

3. Statistical Disclosure Control Methods

In this section, we describe SDC methods for protecting the hypercubes: record swapping, semi-controlled random rounding and a probabilistic perturbation mechanism. From each of the disclosure controlled hypercubes, we generate an output table and compare the SDC

methods through disclosure risk and data utility measures. The comparison will also include the case where the SDC is applied directly on the output table that is generated from the original hypercube.

3.1 Record Swapping

Record swapping is based on the exchange of values of variable(s) between similar pairs of population units (often households). In order to minimize bias, pairs of population units are typically determined within strata defined by control variables, such as a large geographical area, household size and the age-sex distribution of individuals in the households. In addition, record swapping can be targeted to high-risk population units found in small cells of census tables. In a census context, geographical categories are often swapped. Swapping places of residence attempts to minimize bias on the assumption that place of residence is independent to other target variables (conditional on the control variables). Also, place of residence is itself a highly visible variable so swapping this variable has a double benefit. In addition, if one swaps *between* low levels of geography but *within* higher levels of geography then at the higher aggregations of geography, marginal distributions are preserved. For more information on record swapping, see Dalenius and Reiss, 1982, Fienberg and McIntyre, 2005 and Shlomo, 2007.

For this study, we carried out random record swapping at the individual level. In addition, to keep the study simple, a random sample of 5% of the individuals was selected in each NUTS2 region. The selected individuals were paired randomly with other individuals in different LAU2 geographies within the NUTS2 region, and the LAU2 geographies swapped between them. This produced a total of 10% of the individuals in each NUTS2 region having their LAU2 geography variable swapped.

3.2 Semi-Controlled Random Rounding

The most common post-tabular method of SDC for Census frequency tables is based on unbiased random rounding. The entries of the table x are first converted to residuals of the rounding base b, $res(x)$. Let $Floor(x)$ be the largest multiple k of the base b such that $bk \leq x$ for an entry x. In this case, $res(x) = x - Floor(x)$. For an unbiased rounding procedure, x is rounded up to $Floor(x) + b$ with probability $\frac{res(x)}{b}$ and rounded down to $Floor(x)$ with probability $\left(1 - \frac{res(x)}{b}\right)$. If x is already a multiple of b, it remains unchanged.

In general, each small cell is rounded independently in the table, i.e. a random uniform number u between 0 and 1 is generated for each cell. If $u < \frac{res(x)}{b}$ then the entry is rounded up, otherwise it is rounded down. This ensures an unbiased rounding scheme, i.e. the expectation of the rounding perturbation is zero. However, the realization of this stochastic process on a finite number of cells in a table will not necessarily ensure that the sum of the perturbations will exactly equal zero. To place some control in the random rounding procedure, we use a semi-controlled random rounding algorithm for selecting the entries to round up or down as follows: First the expected number of entries of a given $res(x)$ that are to be rounded up is predetermined (for the entire table or for each row/column of the table). The expected number is rounded to the nearest integer. Based on this expected number, a random sample of entries is selected (without replacement) and rounded up. The other entries are rounded down. This process ensures that the rounded internal cells aggregate to the controlled rounded total.

Due to the large number of perturbations in the table, margins are typically rounded separately from internal cells and therefore tables are not additive. When using semi-

controlled random rounding this alleviates some of the problems of non-additivity since one of the margins and the overall total will be controlled, i.e. the rounded internal cells aggregate to the rounded total. Another problem with random rounding is the consistency of the rounding across same cells that are aggregated in different tables. The consistency can be solved by the use of microdata keys. For each record in the microdata, a random number (i.e., a key) is defined which when combined with other records to form a cell of a table defines the seed for the rounding. Records that are aggregated into same cells will always have the same seed and therefore a consistent rounding (Fraser and Wooton, J, 2005, Shlomo and Young, 2008).

For this study, we carry out full random rounding to base 3 semi-controlled to the two NUTS2 totals in the hypercube. As will be seen, we also apply semi-controlled random rounding to base 3 on a final output table generated from the original hypercube.

3.3 Stochastic Perturbation

A more general method than rounding is stochastic perturbation which involves perturbing the internal cells of the hypercube using a mechanism based on a probability transition matrix (similar to the method that is used in PRAM; see Gouweleeuw, Kooiman, Willenborg, and De Wolf, 1998).

Let $\mathbf{P}$ be a $(L+1) \times (L+1)$ transition matrix containing conditional probabilities: $p_{ij} = P(\text{perturbed cell value is } j | \text{original cell value is } i)$ for cell values from 0 to L (usually a cap is put on the cell values and any cell value above the cap would have the same perturbation probabilities). Let $\mathbf{t}$ be the vector of frequencies of the cell values where the last component would contain the number of cells above cap L and $\mathbf{v}$ the vector of relative frequencies: $\mathbf{v} = \mathbf{t}/K$,where K is the number of cells in the table. In each cell of the table, the cell value i

is changed or not changed according to the prescribed transition probabilities in the matrix **P** and the result of a draw of a random multinomial variate u with parameters p_{ij} $(j = 0, 1, \ldots, L)$. If the j-th value is selected, value i is moved to value j. When $i = j$, no change occurs.

Placing the condition of invariance on the transition matrix **P** (i.e. **tP = t)** means that the marginal distribution of the cell values are approximately preserved under the perturbation. As described in the random rounding procedure, in order to obtain the exact overall total a without replacement strategy for selecting the cell values to change can be carried out. For each particular cell value, we calculate the expected number of cells that need to be changed to a different value according to the probabilities in the transition matrix. The expected number of cells is rounded to the nearest integer. We then randomly select (without replacement) the cells and carry out the change.

To preserve exact additivity in the table, an Iterative Proportional Fitting algorithm can be used to fit the margins of the table after the perturbation according to the original margins. This results in cell values that are not integers. Exact additivity with integer counts can be achieved by controlled rounding to base 1 using for example Tau-Argus (Salazar-Gonzalez, Bycroft, and Staggemeier, 2005). Cell values can also be rounded to their nearest integers resulting in 'close' additivity because of the invariance property of the transition matrix. Finally, the use of microdata keys can ensure consistent perturbation of cells across hypercubes.

For this study, we implement the stochastic perturbation based on an invariant probability matrix with controls in the overall totals of the two NUTS2 regions. We carry out the

perturbation on cells of values in the range 0-10; all cells above a value of 11 were not perturbed. The invariant perturbation matrix used in this study is presented in Table 2.

Table 2: Invariant Perturbation Matrix used to Perturb Hypercube

Cell Value	Perturbed Cell Value										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	0.998	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.080	0.760	0.080	0.047	0.024	0.004	0.002	0.001	0.001	0.000	0.000
2	0.080	0.153	0.686	0.047	0.024	0.005	0.002	0.001	0.001	0.000	0.000
3	0.000	0.148	0.078	0.703	0.027	0.031	0.007	0.002	0.002	0.001	0.001
4	0.000	0.103	0.054	0.037	0.725	0.022	0.024	0.020	0.006	0.005	0.005
5	0.000	0.023	0.014	0.055	0.029	0.783	0.031	0.025	0.023	0.009	0.008
6	0.000	0.013	0.007	0.012	0.032	0.032	0.814	0.029	0.026	0.025	0.010
7	0.000	0.005	0.003	0.005	0.035	0.034	0.037	0.797	0.029	0.027	0.027
8	0.000	0.005	0.003	0.005	0.013	0.039	0.042	0.036	0.798	0.030	0.030
9	0.000	0.005	0.003	0.005	0.013	0.017	0.046	0.039	0.034	0.807	0.032
10	0.000	0.005	0.003	0.005	0.013	0.017	0.021	0.043	0.037	0.034	0.823

4. Table Generating Servers

The design of remote table generating servers typically involves many ad-hoc preliminary SDC rules that can easily be programmed within the system to determine a priori tables that should not be released. These SDC rules may include:

- Limiting the number of dimensions in the tables,
- Ensuring consistent and nested categories of variables to avoid disclosure by differencing,
- Ensuring minimum population thresholds,
- Ensuring that the percentage of small cells is above a minimum threshold,
- Ensuring average cell size above a minimum threshold.

Despite these preliminary rules, the output tables generated in the system may still be disclosive and require the application of SDC methods. As mentioned, the SDC methods can be applied on the underlying data or applied directly to the final output table produced from the original data. We compare these approaches in Section 6.

For the flexible table generating server, we assume the scenario that the number of dimensions for generating a table is limited to three with one additional variable defining the population. For our output table, we define the population as those in the first NUTS2 region and define the table as: banded age group*education*occupation. This table contains 2,457 cells with 854,539 individuals, giving an average cell size of 347.8 individuals. The cell counts of the final output table are shown in Table 3.

Table 3: Distribution of Cell Counts in the Generated Table: Banded Age Group*Education*Occupation for NUTS2=1

Cell Value	Number of Cells	Percentage of Cells
0	1534	62.43%
1	44	1.79%
2	35	1.42%
3	27	1.10%
4	20	0.81%
5 and over	797	32.44%
Total	2457	100.00%

5. Information Theory Based Disclosure Risk and Data Utility Measures

For each output table generated, the server must provide disclosure risk and data utility measures. We propose to use Information Theory (IT) to define these measures since the theory is particularly sensitive to the case of attribute disclosure which is caused by a dominant number of empirical (real) zeros in a row/column or table.

5.1 Theoretical Development for the Measures

Information theory is covered comprehensively in Cover and Thomas (2006). One of the most important formulas is entropy. Entropy is a measure of uncertainty in a random variable. Let X be a discrete random variable having a distribution $= (p_1, p_2, \dots, p_K)$. The entropy is defined as:

$$H(X) = H(P) = -\sum_{i=1}^{K} p_i \cdot \log\, p_i.$$

If $p_i = 0$ for a category i, the respective term in the sum will be considered 0, since $\lim_{x\to 0} x \log x = 0$.

One can see easily that $H(P) \geq 0$, since $-p_i \cdot \log\, p_i \geq 0$. Entropy is equal to 0 if the probability mass is concentrated on one point. Under the uniform distribution U_K=$\left(\frac{1}{K}, \frac{1}{K}, \dots, \frac{1}{K}\right)$, we obtain the maximum entropy: $H(U_K) = \log K$.

The entropy of the frequency vector in a table of size K, $F = (F_1, F_2, \dots, F_K)$ where $\sum_{i=1}^{K} F_i = N$ is:

$$H(P) = H\left(\frac{F}{N}\right) = -\sum_{i=1}^{K} \frac{F_i}{N} \cdot \log \frac{F_i}{N} = \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N} \quad (1)$$

To compare two probability distributions ($P = (p_1, p_2, \dots, p_K)$ and $Q = (q_1, q_2, \dots, q_K)$), we use the f-divergences. The concept of f-divergence is specified in Csiszár (1967) and Csiszár and Shields (2004). To define an f-divergence we need a convex function: $\mathbb{R}^+ \to \mathbb{R}$. We assume that $f(1) = 0$. The divergence between the two distributions determined by f is defined as:

$$D_f(P \parallel Q) = \sum_{i=1}^{K} q_i \cdot f\left(\frac{p_i}{q_i}\right)$$

We also assume that: $0 \cdot f\left(\frac{0}{0}\right) = 0$, $f(0) = \lim_{x\to 0} f(x)$, $0 \cdot f\left(\frac{c}{0}\right) = \lim_{x\to 0} x \cdot f\left(\frac{c}{x}\right)$.

Using only these properties of function f, we are able to derive an inequality that has several applications. Let $u_1, u_2, \dots, u_K$ and $v_1, v_2, \dots, v_K$ positive real numbers with sums $u = \sum_{i=1}^{K} u_i$ and $v = \sum_{i=1}^{K} v_i$ respectively. Then the following inequality holds:

$$\sum_{i=1}^{K} v_i \cdot f\left(\frac{u_i}{v_i}\right) \geq v \cdot f\left(\frac{u}{v}\right).$$

The proof of the above inequality is based on the convexity of f. According to Jensen's inequality $\sum_{i=1}^{K} \frac{v_i}{v} \cdot f\left(\frac{u_i}{v_i}\right) \geq f\left(\sum_{i=1}^{K} \frac{v_i}{v} \cdot \frac{u_i}{v_i}\right) = f\left(\frac{u}{v}\right)$. If f is strictly convex at $\frac{u}{v}$, then equality holds if and only if $u_i = \frac{u}{v} \cdot v_i$ for every i.

It follows immediately that $D_f(P \parallel Q) \geq 0$, since choosing $u_i = p_i$ and $v_i = q_i$ provides the inequality of

$$D_f(P \parallel Q) = \sum_{i=1}^{K} q_i \cdot f\left(\frac{p_i}{q_i}\right) \geq 1 \cdot f(1) = 0.$$

Relative entropy or Kullback-Leibler divergence has a similar formulation to entropy but provides a comparison of two distributions. Relative entropy is an f-divergence with $f(x) = x \cdot \log x$ which is a convex function. So, for P=$(p_1, p_2, \dots, p_K)$ and $Q = (q_1, q_2, \dots, q_K)$ the relative entropy is:

$$D_{x \cdot \log x}(P \parallel Q) = D(P \parallel Q) = \sum_{i=1}^{K} p_i \cdot \log\left(\frac{p_i}{q_i}\right). \quad (2)$$

Here $0 \cdot \log\left(\frac{0}{q_i}\right) = 0$, if $q_i > 0$, and $p_i \cdot \log\left(\frac{p_i}{0}\right) = \infty$, if $p_i > 0$.

Relative entropy is not symmetric and triangle inequality also does not hold. Therefore, $D(P \parallel Q)$ does not meet the criteria of distances. The non-negativity of relative entropy follows from the non-negativity of f-divergences, $D(P \parallel Q) \geq 0$, with equality if and only if $p_i = q_i$ for all i.

With this inequality, $H(P) \leq \log K$ as shown above since:

$$0 \leq D(P \parallel U_K) = \sum_{i=1}^{K} p_i \cdot \log\left(\frac{p_i}{1/K}\right) = \sum_{i=1}^{K} p_i \cdot \log p_i + \sum_{i=1}^{K} p_i \cdot \log K = \log K - H(P).$$

with equality holding if and only if the distribution is uniform.

Note that $D(Q \parallel P)$ is also f-divergence with $f(x) = -\log x$.

To measure the distance between two distributions, the L_p-norm can also be used. For an arbitrary vector $x = (x_1, x_2, \dots, x_K)$ the L_p-norm ($1 \leq p < \infty$) of x is defined as:

$$\|x\|_p = \left(\sum_{i=1}^{K} |x_i|^p\right)^{1/p}.$$

The L_2-norm is the Euclidean-norm. As p converges to infinity, $\|x\|_p$ tends to $\max_i |x_i|$. Therefore $\|x\|_\infty = \max_i |x_i|$ is referred to as the L_∞-norm of x.

The distance of two distributions can be expressed as the L_p-norm of the difference: $\|P - Q\|_p$. If $p = 1$, this distance is equivalent to the f-divergence given by $f(x) = |x - 1|$.

L_p-norm induces a metric on $\mathbb{R}^K$ as proven in Serre, 2010, since

1. $\|x - y\|_p \geq 0$, (non-negativity)
2. $\|x - y\|_p = 0 \iff x = y$,
3. $\|x - y\|_p = \|y - x\|_p$ (symmetry)

and also the triangle inequality,

4. $\|x-y\|_p + \|y-z\|_p \geq \|x-z\|_p$, is fulfilled.

Let $p > 1$ and $f(x) = -\log x$. Selecting $u_i = {p_i}^p$ and $v_i = p_i$ results in the following inequality according to the proven inequality of the f-divergences:

$$\sum_{i=1}^{K} p_i \cdot (-\log {p_i}^{p-1}) \geq -\log \sum_{i=1}^{K} {p_i}^p,$$

or

$$H(P) \geq -\frac{1}{p-1} \cdot \log \|P\|_p^p.$$

If $p = 2$, the inequality simplifies into

$$H(P) \geq -\log \|P\|_2^2.$$

For $P = (p_1, p_2, \dots, p_K)$ and $Q = (q_1, q_2, \dots, q_K)$ we denote $\sqrt{P} = \left(\sqrt{p_1}, \sqrt{p_2}, \dots, \sqrt{p_K}\right)$ and $\sqrt{Q} = (\sqrt{q_1}, \sqrt{q_2}, \dots, \sqrt{q_K})$. These are not (necessarily) probability distributions, however, as vectors, their L_2-norms are 1.

We define the Hellinger distance as the following L_2-norm and preserves the properties of a distance:

$$HD(P,Q) = \frac{1}{\sqrt{2}} \cdot \left\|\sqrt{P} - \sqrt{Q}\right\|_2$$

Obviously, $HD(P,Q) \geq 0$. On the other hand, $HD(P,Q) \leq 1$, since

$HD(P,Q) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2} = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(p_i + q_i - 2\sqrt{p_i \cdot q_i}\right)} =$

$\frac{1}{\sqrt{2}} \cdot \sqrt{2 - 2 \cdot \sum_{i=1}^{K} \sqrt{p_i \cdot q_i}} = \sqrt{1 - \sum_{i=1}^{K} \sqrt{p_i \cdot q_i}} \leq 1.$

Suppose that $f(x) = 1 - \sqrt{x}$. Then $\sum_{i=1}^{K} q_i \cdot f\left(\frac{p_i}{q_i}\right) = \sum_{i=1}^{K} q_i \cdot \left(1 - \sqrt{\frac{p_i}{q_i}}\right) = \sum_{i=1}^{K} \left(q_i - \sqrt{p_i \cdot q_i}\right) = 1 - \sum_{i=1}^{K} \sqrt{p_i \cdot q_i} = HD^2(P,Q)$. Therefore Hellinger distance is also an f-divergence.

We can also apply the Hellinger Distance to two vectors of frequencies $F = (F_1, F_2, \ldots, F_K)$ and $G = (G_1, G_2, \ldots, G_K)$ where $\sum_{i=1}^{K} F_i = N$ and $\sum_{i=1}^{K} G_i = M$.

$$HD(F,G) = \frac{1}{\sqrt{2}} \cdot \left\|\sqrt{F} - \sqrt{G}\right\|_2 = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(\sqrt{F_i} - \sqrt{G_i}\right)^2} \quad (3)$$

The Hellinger Distance shows the magnitude of the cells since the difference between the square roots of two 'large' numbers is higher than the difference between two 'small' numbers, even if these pairs have the same absolute difference. Naturally, while the lower bound remains zero, the upper bound of this distance of counts changes:

$$HD(F,G) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(\sqrt{F_i} - \sqrt{G_i}\right)^2} = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(F_i + G_i - 2 \cdot \sqrt{F_i \cdot G_i}\right)} =$$

$$\frac{1}{\sqrt{2}} \cdot \sqrt{N + M - 2 \cdot \sum_{i=1}^{K} \sqrt{F_i \cdot G_i}} \leq \sqrt{\frac{N+M}{2}}.$$

We can also show the further inequality: $2 \cdot HD^2(P,Q) \leq D(Q \parallel P)$, and in terms of frequencies: $2 \cdot HD^2(F,G) \leq \frac{1}{2} \cdot \left(D(F \parallel G) + D(G \parallel F)\right)$.

5.2 An Information Theory Disclosure Risk Measure

A small level of entropy can indicate few non-zero cells in a row/column or table. The fewer the number of non-zero cells, the more likely that attribute disclosure occurs. We use the frequency based entropy as defined in (1). To produce a disclosure risk measure between 0 and 1, we define the risk measure as: $1 - \frac{H\left(\frac{F}{N}\right)}{\log K}$.

The entropy however does not take into account the magnitude of the cells counts or the number of zeros in the table (or row/column of the table) which both contribute to identity disclosure. Let A be the set of zeros in the table and $|A|$ the number of zeros in the set. We define a disclosure risk measure as a weighted average of different components, each component being a measure between 0 and 1 as follows:

$$R(F, w_1, w_2) = w_1 \cdot \left[\frac{|A|}{K}\right] + w_2 \cdot \left[1 - \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N \cdot \log K}\right] - (1 - w_1 - w_2) \cdot \left[\frac{1}{\sqrt{N}} \cdot \log \frac{1}{e\sqrt{N}}\right] \quad (4)$$

The first measure in (4) is the proportion of zeros which is relevant for attribute disclosure, the more zeros in a table, the more risk of learning new attributes after an identification. The second measure in (4) is the risk based on the entropy as shown in (1) which is the core of the risk measure. The third measure in (4) allows us to differentiate between tables with different magnitudes. As the population size *N* gets larger, the measure converges to zero. The weights w_1 and w_2 should be chosen depending on the data protector's choice of how important each of the terms are in contributing to the disclosure risk.

As can be seen, the final disclosure risk measure in (4) can be calculated 'on the fly' by the flexible table generating server without the need to see the table beforehand. In order to emphasize the risk of small counts (ones and twos) which still remain in the table for some of the SDC methods, we split the entropy measure as shown in (1) (and the second term in (4)) into two parts, small counts up to 3 and larger counts 4 and more, and provide different weights for each part. For this study, we use weights: $w_1 = 0.1$, $w_{2Part1} = 0.7$, $w_{2Part2} = 0.1$, and $w_3 = 0.1$ which provides the largest weight to the entropy based on small counts.

5.3 Adapting Disclosure Risk After Perturbation

The disclosure risk measure in (4) does not take into account perturbation methods. Random rounding, for example, eliminates ones and twos by introducing more zeros and threes in the table, and seemingly increases the risk of attribute disclosure. Although the extra zeros in the table are random and not real zeros, the disclosure risk as measured by the entropy in (1) (and the second term in (4)) will not reflect the noise introduced into the table and may even produce a higher disclosure risk estimate. So, in order to take into account the perturbation, we propose to modify the first two terms of the risk measure in (4) as follows:

1. We generalize the first term of the proportion of zeros in (4) in order to compare the number of zeros in the original and perturbed table. From (4), *A* is the set of zeros in the original table and |*A*| is the number of zeros in the set. Similarly, let *B* be the set of zeros in the perturbed table and |*B*| the number of zeros in the set. We denote $A \cup B$ as the union of the sets of zeros in the original and perturbed table and $A \cap B$ as the intersection of the sets of zeros in the original and perturbed table. The revised measure, which takes into account that non-zero cells may be transformed into zero counts and vice versa, is defined

as: $\left(\frac{|A|}{K}\right)^{\frac{|A\cup B|}{|A\cap B|}}$.

To control the rate of convergence to zero we may replace the power term $\frac{|A\cup B|}{|A\cap B|}$ with a square root: $\sqrt{\frac{|A\cup B|}{|A\cap B|}}$.

2. We assume that the possible values in the table are: $0, 1, 2, \ldots, L$ and the frequency of frequencies of these values is denoted by: $(n_0, n_1, n_2, \ldots, n_L)$ and that the table is perturbed according to a perturbation mechanism (for example, using the perturbation matrix as shown in Table 2). Let the frequency of frequencies of the perturbed values be denoted by: $\left(n_0^{'}, n_1^{'}, n_2^{'}, \ldots, n_L^{'}\right)$. The contribution to the total for value $j, j = 0, 1, \ldots, L$ in the perturbed table is: $\sum_{i=0}^{L} i \cdot n_i \cdot p_{ij}$.

   We replace the observed perturbed values of value j by the term: $\frac{\sum_{i=0}^{L} i\cdot n_i\cdot p_{ij}}{n_j^{'}}$. As an example, assume the SDC method of random rounding. We replace the zero cells in the perturbed table by: $\left[0 \cdot n_0 + 1 \cdot \frac{2}{3} \cdot n_1 + 2 \cdot \frac{1}{3} \cdot n_2\right] / n_0'$ and replace the cells of size three in the perturbed table by: $\left[1 \cdot \frac{1}{3} \cdot n_1 + 2 \cdot \frac{2}{3} \cdot n_2 + 3 \cdot n_3 + 4 \cdot \frac{2}{3} \cdot n_4 + 5 \cdot \frac{1}{3} \cdot n_5\right] / n_3'$. The procedure ensures the same overall total of the original and adjusted vector of counts. After replacing the values in the perturbed table, we calculate the entropy as shown in (1) (and the second term in (4)).

5.4 Data Utility Measure

For the data utility measure we use the distance metric defined by the Hellinger Distance in (3) where $\sqrt{G} = \left(\sqrt{G_1}, \sqrt{G_2}, \ldots, \sqrt{G_K}\right)$ is the vector of square roots of the perturbed counts:

$$HD(F, G) = \sqrt{\frac{1}{2} \cdot \sum_{i=1}^{K} \left(\sqrt{F_i} - \sqrt{G_i}\right)^2} \quad (5)$$

Since all the SDC methods applied to the table produce approximately the same total *N* due to the controlled methods of perturbation, we can compare the Hellinger Distance across the methods as it is bounded by 0 and approximately $\sqrt{N}$.

6. Results

We report in Table 4 the disclosure risk measure in (4) and the Hellinger Distance in (5) for the table defined in Section 4 based on SDC methods on the input hypercube (record swapping, semi-controlled random rounding and stochastic perturbation) described in Section 3. In addition, we report the measures when implementing the SDC method of semi-controlled random rounding applied directly on the output table generated from the original hypercube.

Regarding the number of small cells of size 1 and 2, there were a total of 6,140 small cells in the hypercube (2.5%). The stochastic perturbation changed only 6.9% of the small cells, the random rounding to base 3 changed 100% of the small cells and the random record swapping changed 16.2% of the small cells.

From Table 4, it is clear that the method of record swapping when applied to the hypercube did little to reduce the disclosure risk in the final output table. This was due to the fact that most of the small cells remained unperturbed in the final table. On the other hand, record swapping provides the smallest distance metric (highest data utility) between the original and

perturbed table compared to the other pre-tabular methods. From among the input perturbation methods on the hypercube, the stochastic perturbation provided the most protection against disclosure but at the cost of a low data utility with the highest distance metric between the original and perturbed table. Removing the small cells entirely and rounding the other cells provided lower disclosure risk as seen in the measures for the semi-controlled random rounding but had less of an impact on the data utility. Comparing the pre-tabular and post-tabular semi-controlled random rounding procedure, we see slightly lower disclosure risk based on the post-tabular rounding but much improvement in data utility since the SDC method is not compounded by aggregating rounded cells. The semi-controlled random rounding on the final output table would be the preferred method based on the results of the study.

Table 4: Disclosure Risk and Data Utility for the Generated Table

	Disclosure Risk	**Hellinger Distance**
Original	**0.352**	**-**
Perturbed Input		
Record Swapping	**0.351**	**6.469**
Semi-controlled Random Rounding	**0.237**	**7.970**
Stochastic Perturbation	**0.230**	**14.120**
Perturbed Output		
Semi-Controlled Random Rounding	**0.233**	**5.902**

7. Concluding Remarks

In this paper, we describe a simulation study comparing the application of SDC methods at different stages of generating tables in a flexible table generating server. For the pre-tabular methods, record swapping had little impact on reducing our measure of disclosure risk and therefore we would not recommend it in a flexible table generating server of census data.

Semi-controlled random rounding offers more protection since every cell in the table is perturbed and by preserving consistency of cells across tables, it is more difficult to 'attack' the rounding to obtain the original table. The stochastic perturbation can also be refined to improve data utility by adapting the transition matrix in Table 2 but this will come at the cost of higher disclosure risk. However, for a flexible table generating server, we have seen that the post-tabular SDC method achieves nearly as good a disclosure risk impact as the pre-tabular stochastic perturbation method whilst achieving the best level of data utility.

We also propose new measures for disclosure risk and data utility based on Information Theory which are particularly suited for assessing disclosure risk arising from attribute disclosure in tables and can easily be embedded in a flexible table generating server.

Post-tabular stochastic perturbation combined with preliminary SDC rules may also provide more protection. This method can also be adapted to guarantee differential privacy according to the Computer Science definitions. Further research needs to be directed to improving stochastic post-tabular SDC methods whilst preserving additivity and consistency of user-defined tables.

References

Cover, T. M. and Thomas, J. A. (2006). Elements of Information Theory, 2nd ed., New York: Wiley.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientarium Mathematicarum Hungarica,* 2, 299-318.

Csiszár, I. and Shields, P. C. (2004). Information Theory and Statistics: A Tutorial. *Foundations and Trends in Communications and Information Theory,* 1, Issue 4.

Dalenius, T. and Reiss, S.P. (1982). Data Swapping: A Technique for Disclosure Control. *Journal of Statistical Planning and Inference*, 7, 73-85.

Dinur, I. and Nissim, K. (2003). Revealing Information While Preserving Privacy. *PODS 2003*, pp. 202-210.

Fienberg, S.E. and McIntyre, J. (2005). Data Swapping: Variations on a Theme by Dalenius and Reiss. *Journal of Official Statistics*, 9, 383-406.

Fraser, B. and Wooton, J. (2005). A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing. Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, 9-11 November.

Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.P. (1998). Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics*, 14, 463-478.

Salazar-Gonzalez, J.J., Bycroft, C. and Staggemeier, A.T. (2005). Controlled Rounding Implementation. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, 9-11 November.

Serre, D. (2010). *Matrices: Theory and Applications*. Graduate Texts in Mathematics: 216, 2nd Edition. New York: Springer.

Shlomo, N. (2007). Statistical Disclosure Control Methods for Census Frequency Tables. *International Statistical Review*, Vol. 75, Number 2, pp. 199-217.

Shlomo, N. and Young, C. (2008). Invariant Post-tabular Protection of Census Frequency Counts. In In *PSD'2008 Privacy in Statistical Databases*, (Eds. J. Domingo-Ferrer and Y. Saygin), Springer LNCS 5261, pp. 77-89.