Journal of Privacy and Confidentiality
Vol. 8 (1) 2018

Issue in honor of S. Fienberg

Submitted Dec. 9, 2018
Published December 2018

STATISTICAL DISCLOSURE LIMITATION: NEW DIRECTIONS AND CHALLENGES

NATALIE SHLOMO

University of Manchester, Manchester, United Kingdom
e-mail address: Natalie.Shlomo@manchester.ac.uk

Abstract. An overview of traditional types of data dissemination at statistical agencies is provided including definitions of disclosure risks, the quantification of disclosure risk and data utility and common statistical disclosure limitation (SDL) methods. However, with technological advancements and the increasing push by governments for open and accessible data, new forms of data dissemination are currently being explored. We focus on web-based applications such as flexible table builders and remote analysis servers, synthetic data and remote access. Many of these applications introduce new challenges for statistical agencies as they are gradually relinquishing some of their control on what data is released. There is now more recognition of the need for perturbative methods to protect the confidentiality of data subjects. These new forms of data dissemination are changing the landscape of how disclosure risks are conceptualized and the types of SDL methods that need to be applied to protect the data. In particular, inferential disclosure is the main disclosure risk of concern and encompasses the traditional types of disclosure risks based on identity and attribute disclosures. These challenges have led to statisticians exploring the computer science definition of differential privacy and privacy- by-design applications. We explore how differential privacy can be a useful addition to the current SDL framework within statistical agencies.

Key words and phrases: Inferential Disclosure, Disclosure Risk, Data Utility, Differential Privacy.

JPrivacy Confidentiality

www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.684

During this period of preparing for the special issue of the *Journal of Privacy and Confidentiality* in honour of Steve Fienberg, we received news of the tragic events that occurred at the Tree of Life Synagogue in Pittsburgh on October 27th, 2018 and the sudden senseless death of Joyce Fienberg. Whilst Steve was a great support and mentor to me as I embarked on my PhD research at the Hebrew University and the University of Southampton in 2004, he was married to an extraordinary woman who showed endless kindness to me and all of Steve's students and mentees. I had a wonderful visit to CMU during my sabbatical period in November 2011 spending much quality time with both Steve and Joyce.

As my mentor, Steve marked my PhD dissertation in 2007, provided me with advice and support as I embarked on an academic career and provided many recommendation and promotion letters over the years. I can honestly credit Steve with where I am today in my academic career. Steve was instrumental in bringing differential privacy to the forefront of research in statistical disclosure limitation and provided many opportunities to bring statisticians and computer scientists together for collaborations. Our most recent initiative was the *Data Linkage and Anonymisation Programme* at the Isaac Newton Institute of Mathematical Sciences at the University of Cambridge from July through December 2016. Steve was to participate in the programme but alas his illness took the better of him during that time. In fact, Steve was to participate in all three programmes that were running at the Institute: *Data Linkage and Anonymisation*, *Theoretical Foundations for Statistical Network Analysis* and *Probability and Statistics in Forensic Science* which demonstrates the breadth and depth of his research activities and achievements. He was sorely missed.

I can only hope that these words of devotion and appreciation will provide some comfort to Steve and Joyce's family. I end with a Hebrew blessing — זיכרונם לברכה (zichronam livracha) —- may their memory be a blessing.

Natalie Shlomo

DOI: 10.29012/jpc.695

1. Introduction

For many decades, statistical disclosure limitation (SDL) has been an important area of research for statistical agencies, data archives and other organizations responsible for the release of statistical data. These agencies have a legal obligation to maintain the confidentiality of statistical entities and in many countries there are codes of practice that must be strictly adhered to. In addition, statistical agencies have a moral and ethical obligation towards respondents that participate in surveys and censuses through confidentiality pledges presented to them prior to their participation. The key objective is to ensure public trust in official statistics production and hence ensure high response rates.

Traditionally, the types of data that are released by statistical agencies take the form of tabular data and microdata. Tabular data can contain frequency counts for whole populations such as from a census or register, weighted frequency counts from surveys and magnitude data containing totals and averages that are typically derived from business statistics. More recently, microdata from social surveys are also released usually through special license agreements or deposited into data archives where researchers can register and apply for access to data. Microdata from business surveys are generally not released because of their disclosive nature due to large sampling fractions including a 'take-all' strata and skewed distributions.

Statistical agencies must assess the disclosure risk in statistical data and if required choose appropriate SDL methods to apply to the data.

Measuring disclosure risk involves assessing and evaluating numerically the risk of re-identifying statistical units. SDL methods perturb, modify, or summarize the data in order to prevent re-identification by a potential attacker. Higher levels of protection through SDL methods however impact negatively on the quality of the data. The SDL decision framework involves finding the optimal balance between managing and minimizing disclosure risk to tolerable risk thresholds depending on how the data will be accessed and ensuring high utility and fit-for-purpose data.

With technological advancements and the increasing push by governments for open and accessible data, new forms of data dissemination including web-based applications are currently being explored by statistical agencies and data archives. On the other hand, the digitalization of all aspects of society means that personal information is often easily obtainable from the internet and there are increasing disclosure risks. This has changed the landscape of how disclosure risks need to be defined and the types of SDL methods that should be applied to protect the data from disclosures.

In Section 2 we first provide an overview of the SDL framework for traditional data dissemination of tabular data and microdata including the disclosure risks, approaches for confidentializing the data and disclosure risk and data utility measurement. In Section 3 we discuss the disclosure risk of inferential disclosure which encompasses traditional types of disclosure risks. Inferential disclosure risk is becoming more important as agencies are moving towards more flexible web-based modes of dissemination in the future. Section 4 then discusses new data dissemination strategies that are being applied or are under consideration by statistical agencies and how these generate new challenges on the measurement of disclosure risk and data utility. In particular, we examine whether the computer science standard of differential privacy can be integrated into the SDL framework to meet these challenges. We conclude in Section 5 with a discussion.

2. Traditional SDL Approaches

As mentioned in the introduction, traditional types of data releases are tabular data containing frequency counts, microdata from social surveys and magnitude tables. In this section, we present a brief overview of how disclosure risks are defined and measured, some common SDL methods that are applied and the measurement of data utility.

2.1 Types of Disclosure Risks

For traditional types of data releases, the two main disclosure risks are identity disclosure where a data subject can be identified based on a set of quasi-identifying variables and attribute disclosure where new information can then be learnt. Attribute disclosure can also occur without an identity disclosure. For example, sensitive information about a group of individuals may be revealed which could cause harm.

Identity disclosure for microdata from a social survey can arise if a data subject can be re-identified based on the set of quasi-identifiers in the data, for example by linking the microdata to an external data source containing information about the population where the quasi-identifiers are used as matching variables. The quasi-identifiers are typically visible and traceable categorical variables, such as sex, age, occupation, place of residence and marital status. When the quasi-identifiers are cross-classified, the cells formed by the cross-classification may have very small counts including many cells that have a value of zero. Once a re-identification is made, attribute disclosure then arises from the remaining survey target variables in the microdata, such as information about health, income and expenditures.

Therefore, for microdata arising from social surveys, the SDL methods are typically about reducing the risk of re-identification in order to avoid attribute disclosure.

Identity disclosure for tabular data containing whole population counts can occur if there are singleton cells in the table. Attribute disclosure comes from the marginal cells of the table. If there is a singleton on the margin of the table then that implies that a re-identification can be made on less variables defining the table and new information is learnt. In fact, it is the zero cells that cause attribute disclosure in frequency tables and this occurs when a row/column have all zero cell values except for one non-zero cell. Even if the marginal total has a large number but the given row/column contain only one non-zero cell value, this leads to group attribute disclosure. As mentioned, group attribute disclosure may cause harm and is avoided by statistical agencies.

2.2. MICRODATA FROM SOCIAL SURVEYS

Traditional methods of protecting microdata from social surveys include both 'safe data' and 'safe access' approaches. In terms of 'safe data', survey microdata is generally protected by coarsening the quasi-identifiers, deleting sensitive variables, such as low-level geographies, and top-coding sensitive variables such as the size of the household, income and expenditures. Since social surveys typically have very small sample fractions, statistical agencies generally assume that a potential attacker would not have response knowledge, meaning that the attacker would not know if an individual is included in the survey microdata or not. Sampling therefore provides an inherent level of protection and is considered an SDL method in itself. In terms of 'safe access', the survey microdata is generally released into data archives or under special licenses so that users need to undergo an application process and state the purpose of their request prior to obtaining access to the data.

Perturbative SDL methods might also be applied on the quasi-identifiers. In record swapping, variables(s), such as the geographic location, will be swapped between a select number of pairs of records having similar characteristics. Post-randomization (PRAM) introduces misclassification in the quasi-identifiers through a probability mechanism and the result of a random draw (Gouweleeuw, et al. 1998). Other perturbative SDL methods may be applied to the sensitive variables to reduce the risk of attribute disclosure, for example rounding or adding random noise to an income variable.

One of the first approaches to quantify disclosure risk in survey microdata was by record linkage (distance-based or probabilistic) where the confidentialized data was matched back to the original data based on the set of quasi-identifiers. The number of correct matches formed the basis for the quantification of the risk of re-identification. This approach however did not account for the protection afforded by the sampling. Bethlehem, et al. (1990) was among the first to describe a probabilistic modelling framework for estimating the risk of re-identification. The risk measures are based on the notion of population uniqueness on the cells defined by the cross-classification of the quasi-identifiers.

Denoting F_k the population size in cell k of a table defined by quasi-identifying variables having K cells and f_k the sample size and $\sum_{k=1}^{K} F_k = N$ and $\sum_{k=1}^{K} f_k = n$, the set of sample uniques, is defined as: $SU = \{k: f_k = 1\}$. The sample uniques are potential high-risk records since they may be population uniques. Individual per-record risk measures in the form of a probability of re-identification are estimated. These per-record risk measures are then aggregated to obtain global risk measures as follows (where I is the indicator function):

1. Number of sample uniques that are population uniques:
   $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$

2. Expected number of correct matches for sample uniques (i.e., a matching probability)
   $\tau_2 = \sum_k I(f_k = 1)\frac{1}{F_k}$.

The individual risk measure for τ_2, for example, is $1/F_k$ the match probability in cell k of a sample unique to the population.

When the population is unknown, Skinner and Holmes (1998) and Elamir and Skinner (2006) propose using a Poisson Distribution to estimate the disclosure risk measures with log-linear modelling to estimate population parameters inferred from the observed sample counts. Skinner and Shlomo (2008) developed goodness of fit criteria for determining the optimal log-linear model which produces unbiased estimates of the disclosure risk measures. Shlomo and Skinner (2010) adapt the estimation of risk measures to take into account measurement and perturbation errors. An extension of the probabilistic modelling by Reiter (2005a) accounted for the probability of re-identification weighted by suppositions on attacker knowledge regarding the methods of perturbation. More recently, Manrique-Vallier and Reiter (2012) used mixed membership models to estimate the probability of re-identification.

Utility measures for assessing the impact of the SDL methods on the quality of the data and whether the data remain fit-for-purpose are largely subjective and depend on the usage of the data. They include:

- Distance metrics between key parameters or distributions calculated from the original and confidentialized microdata using for example, a relative distance, the Hellinger's Distance or the Kullback-Leibler divergence. These metrics identify any bias that may have been introduced due to the SDL methods.
- A file level utility measure developed in Karr et al. (2006) and Woo et al. (2009) is a statistic based on a propensity score. Stacking the original and confidentialized microdata and defining an indicator of 1 for the confidentialized microdata and 0 for the original microdata, a propensity score is estimated using a logistic regression model. The test statistic is then $\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - 0.5)^2$where N is the size of the combined dataset. Snoke et al. (2018) provides a standardized version of the test statistic to facilitate testing and comparison of SDL approaches.
- Potential impact on statistical inference when using the confidentialized microdata compared to the original microdata through an evaluation of the differences in the variance of key parameter estimates, for example, the overlap of confidence intervals on means and regression parameters. In addition, it is important to understand if there is an impact on hypothesis testing and therefore useful utility measures include differences in test statistics such as the Chi-squared test statistic, rank correlations, R^2 and deviance for statistical modelling.

2.3 Frequency Tables for Whole Population Counts

We focus on frequency tables for whole population data, such as censuses and registers, since frequency tables containing weighted survey counts have little need for SDL methods and in fact, small survey counts in tables are often suppressed due to their low quality. There is generally strict control on what census/register based tables can be released due to the need to avoid sparse tables, differencing and linking tables. Statistical agencies devote much time and resources to the design of these tables with respect to the selection of variables and their categories defining the tables. The hard-copy frequency tables are generally made available on statistical agencies' websites and there may be specialized software available that can trawl and extract parts of the tables. Any special requests for tabulations from whole population counts are assessed against previous tabular releases.

There are two types of SDL methods for tabular data containing whole population counts: pre-tabular and post-tabular and combinations of both. Pre-tabular SDL methods are implemented on the microdata prior to the compilation of the tables. The United States and United Kingdom censuses use record swapping defined in Section 2.2 on their census microdata prior to tabulation. Post-tabular SDL methods are implemented on the cell values of the tables after they are generated and typically take the form of rounding or perturbing the cell values. The aim is to introduce ambiguity in the zero cell values of the tables so that it will not be known whether an observed zero in a table is a structural zero or a random zero. Random rounding rounds the value of each cell according to a probability mechanism and internal cells and marginal cells are rounded separately resulting in rows/columns of the tables that may not be additive. Controlled rounding ensures that the sum of rounded internal cells equal the rounded marginal total which is a desired property by users of the data. However, controlled rounding is too limiting for the large scale production of census/ register-based tables.

A more general case of random rounding is random cell perturbation based on a probability transition matrix which was first carried out at the Australian Bureau of Statistics (ABS) and described in Fraser and Wooton (2005). The approach is similar to PRAM described in Section 2.1 but in this case it is the values of the cells that are perturbed (or not perturbed) depending on the outcome of a random draw.

A probability transition matrix $\mathbf{P}$ is defined where:

$$p_{ij} = P(perturbed\ count = j | original\ count = i)$$

Shlomo and Young (2008) modified the method to preserve additivity in the tables (in expectation) by transformation of the probability transition matrix so that the frequencies of the cell values are preserved in the perturbed table. Let t be the vector containing the frequency counts of the original cell values: 0,1,2,3, etc. We place the condition of invariance on the probability transition matrix P such that $\mathrm{tP} = \mathrm{t}$ and the released table is a moment estimator of the original table.

Since the tabular data are based on whole population counts, disclosure risk measurement is straight-forward and there are general 'rules-of-thumb' that are followed: avoiding tables that are sparse having many small cell counts and ensuring that the row/columns do not contain only one or two non-zero cell counts. Since tables are released as hard-copy tables, disclosure by differencing and linking tables is generally not a problem since these are controlled by design.

Degenerate distributions in tables where rows/columns are mainly zero with few non-zero cells can be identified through disclosure risk measures grounded in Information Theory and developed in Antal, et al. (2014). The measures are based on the entropy and assign a value between 0 and 1 for the level of risk caused by degenerate distributions. In Antal, et al. (2015), the risk measures are expanded to account for the application of SDL methods through the conditional entropy which represents the amount of information needed to recover the original table given that we observe the confidentialized table.

To assess the impact on data utility for frequency tables of whole population counts, we can use similar utility measures as described for microdata in Section 2.2 since many of the measures for microdata are based on examining frequency distributions in the original data versus the confidentialized data. For example, the utility measures based on distance metrics between original and perturbed cell values in a table are relevant. In particular, the Hellinger's Distance is an often used utility measure as it allows for cell values that may contain a zero and in addition places more emphasis on small counts compared to the large counts. The impact on Chi-square testing for statistical associations between variables defining the table is also relevant. The aim is to ensure that the power of such tests is not impacted by the perturbation and there is no change in statistical inference.

2. 4 MAGNITUDE TABLES FROM BUSINESS STATISTICS

Magnitude tables are defined as tables where the cells contain sums or averages of a continuous variable such as total turnover, profits or revenue and the table is defined by quasi-identifying variables, such as region and economic activity. Potential attackers to this type of statistical data are other businesses that may be interested in learning sensitive commercial information about their competitors. Therefore, we assume that attackers are competing businesses in a cell of the table and that the identity of other businesses in the cell is known. In addition, we assume that the attackers also know the ranking of the businesses with respect to their size. The main concern is therefore one of attribute disclosure.

Disclosure risk measures are known as sensitivity measures and are based on whether a contributor in the cell of a table can learn the values of the target variable for the other contributors in the cell with sufficient precision. Since business surveys have large sampling fractions and in particular take-all strata for large businesses, we do not account for any protection afforded by sampling.

In the general framework, a table is defined by cross-classification of categorical variables. Let *X* denote a generic cell, *N(X)* denote the number of contributors in the cell and x_i denote the value of the target variable for contributor *i*. We define the total in cell *X* as $T(X) = \sum x_i$. Assume $x_i > 0$ for all $i = 1, \dots, N(X)$ and that the observations can be ordered so that: $x_1 \geq x_2 \geq \cdots \geq x_{N(X)} > 0$. Assuming that an attacker is a contributor in the same cell, we wish to avoid the attacker from being able to disclose an x_i value for other *i*. One sensitivity measure is the dominance rule: the (n, p) dominance rule classifies a cell as disclosive if $x_1 + \cdots + x_n \geq \frac{p}{100} T(X)$. This rule assumes that *n* businesses in a cell, say 2 businesses, can form a coalition to disclose a value for the third business in a cell. In addition, any cell having a small number of contributors, for example 3 contributors, is deemed disclosive. Another sensitivity measure is the *p*% rule. The most precise estimate by the second largest contributor for the value of the largest contributor in a cell is: $\hat{x}_1 = T(X) - x_2$. The percent error is: $100 \times (\hat{x}_1 - x_1)/x_1 = 100 \times (T(X) - x_1 - x_2)/x_1$. In the *p*% rule, the cell is disclosive if $100 \times (T(X) - x_1 - x_2)/ \mathrm{x}_1 \leq \mathrm{p}$. It has been established that if the parameters of the sensitivity measures are known to attackers, such as the *p* or *n,* they can be used to disclose sensitive information and hence the parameters are not released.

To protect magnitude tables containing business statistics, table design and cell suppression are generally used. Based on the sensitivity measures, disclosive cells are suppressed. These are called primary suppressions. Then, other cells need to be suppressed to ensure that the primary suppressions are not revealed through the marginal totals. These are called secondary suppressions. For a 2 by 2 table for example, at least 2 cells in a row and column, i.e. the vertices of a rectangle, need to be suppressed to ensure that the primary suppressions are safe and cannot be recalculated. To optimize secondary cell suppressions mathematical linear programming is used, for example in Tau-Argus (Salazar-Gonzalez et al., 2005) where an objective function $\sum C(X)$ is minimized. For *C(X)*=1 we minimize the total number of cells suppressed, for *C(X)=N(X)* we minimize the number of contributors suppressed and for *C(X)=T(X)* we minimise the total value of the target variable suppressed. The solution of the linear programming can be heavy (NP hard) so simplified and alternative solutions may be used. The constraints of the mathematical linear programming are the preservation of margins and ensuring non-negative values in the table. For more information on sensitivity measures and optimal cell suppression, see Willenborg and De Waal (2001), Duncan, et al. (2011) and Hundepool, et al. (2012).

2.5 Risk-utility map

The disclosure risk and utility measures can be used to produce a disclosure risk-data utility confidentiality map (Duncan, et al. 2001). We conceptualize the map in Figure 1.

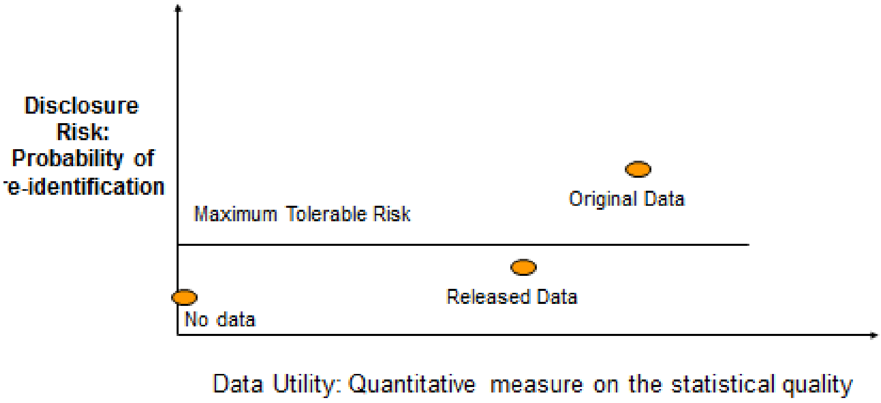

Disclosure Risk: Probability of re-identification
Maximum Tolerable Risk
Original Data
Released Data
No data
Data Utility: Quantitative measure on the statistical quality

Figure 1: Conceptualized Disclosure Risk-Data Utility Confidentiality Map

In the lower left hand quadrant of the map in Figure 1, we have low disclosure risk and low utility. In fact, not releasing data at all will have no utility although some disclosure risk remains as information about the disclosive nature of the data is leaked by not allowing its release. In the upper right hand quadrant of the map in Figure 1, we have high disclosure risk and high utility. We can see that the original data is above a maximal tolerable risk threshold determined by the statistical agency and hence SDL methods need to be applied. Thus SDL is an iterative process, where different SDL methods are applied with different parameterizations, the disclosure risk and data utility are quantified and mapped on to the Disclosure Risk-Data utility confidentiality map. The SDL method that is below the risk threshold and having the highest utility is selected. Note that the data points form a frontier on the map which allows the selection of the optimal SDL method.

3. Inferential Disclosure Risk and Differential Privacy

In Section 2, we reviewed the traditional SDL framework for microdata and tabular data. Disclosure risks were defined as identity and attribute disclosures. However, these types of disclosure risks are essentially components of a more general disclosure risk and that is inferential disclosure. Inferential disclosure risk is defined as the ability to learn new attributes with a high degree of confidence. For example, a regression model with a very high predictive power may cause inferential disclosure. Even if an individual is not in the dataset, there would still be disclosure which may cause harm to the individual. Another example of inferential disclosure is disclosure by differencing where census tables can be manipulated, linked and differenced and cause disclosures of sensitive information. In this case, even large cell counts can be disclosive.

Statistical agencies protect data releases from inferential disclosure by keeping strict control on the data that is released. For example, census tables are vetted to ensure that no two tables can be differenced and thus produce a disclosive table of small counts. Microdata is generally licensed and placed in data archives where it requires lengthy application procedures to obtain access to the data.

As statistical agencies are considering more dynamic approaches for releasing statistical data and relinquishing some of their control on statistical outputs, the methods described in Section 2 for reducing the risk of identity and attribute disclosures are not effective in handling inferential disclosure. In fact, most of the SDL approaches in Section 2 fail if the standard is provable defence against inferential disclosure.

This has led to statisticians exploring the potential of differential privacy for confidentializing statistical data in the framework of SDL. Differential privacy was developed by computer scientists as a standard for a perturbation mechanism for protecting outputs in a remote query-based system with the aim to specifically protect against inferential disclosure. See Dinur and Nissim (2003), Dwork, et al. (2006) and an overview book by Dwork and Roth (2014) for more details on differential privacy.

In differential privacy, a 'worst case' scenario is allowed for, in which the potential attacker has complete information about all the units in the database except for one unit of interest. The definition of a perturbation mechanism M satisfies ε-differential privacy if for all queries on neighbouring databases $a, a' \epsilon A$ differing by one individual and for all possible outcomes defined as subsets $S \epsilon Range(M)$ we have:

$$P(M(a) \in S) \leq e^{\varepsilon} P(M(a') \in S) \qquad (1)$$

This means that observing a perturbed output S, little can be learnt (up to a degree of e^{ε}) and the attacker is unable to decipher whether the output was generated from database a or a'. In other words, the ratio $\frac{P(M(a) \epsilon S)}{P(M(a\prime) \epsilon S)}$ is bounded and the probability in the denominator cannot be zero. The solution to guarantee differential privacy in the computer science literature is by adding noise/perturbation to the outputs of the queries under specific parameterizations and typically the noise is generated from the Laplace Distribution.

Shlomo and Skinner (2012) discuss differential privacy with respect to sampling and perturbation according to the SDL methods that were presented in Section 2. They found that sampling and other non-probabilistic forms of SDL methods are not differentially private since in these cases the denominator in the ratio based on two neighbouring datasets could take on a value of zero. However, for the kinds of large populations of individuals upon which social surveys are based, this failure of an unbounded ratio may occur with only a negligible probability. This leads to the definition of (ε, δ)- differential privacy:

$$P(M(a) \in S) \leq e^{\varepsilon} P(M(a') \in S) + \delta \qquad (2)$$

where δ is a small probability of an unbounded ratio. This relaxation of (ε, δ)- differential privacy allows for more utility under the probability mechanism M.

4. New Dissemination Strategies

In Section 2, we focused on traditional types of statistical data that are disseminated by statistical agencies: tabular data and microdata. However, with increasing demand for more open and accessible statistical data, statistical agencies are now considering alternative and more flexible dissemination strategies including web-based applications. In this section, we examine some of these strategies and how differential privacy can be embedded in current SDL practices through more rigorous and well defined perturbation mechanisms with privacy guarantees.

4.1 Web-based Applications

In recent years, there are two types of web-based dissemination applications that are being considered or are under development within statistical agencies: flexible table generators and remote analysis servers.

4.1.1. Flexible Table Generating Servers

Driven by demand from policy makers and researchers for specialized and tailored tables from statistical data, particularly census data, some statistical agencies are developing or considering online flexible table generating servers that allow users to define and generate their own tables. A good example is the Australian Bureau of Statistics (ABS) TableBuilder for disseminating census tables.

In flexible table generating, users access the servers via the internet and define their own table of interest from a set of pre-defined variables and categories typically from drop down lists. The generated table undergoes a series of ad-hoc SDL checks and if it passes the criteria, it is downloaded onto the user's PC without the need for human intervention. The SDL checks can easily be programmed within the system to determine whether tables can be released to the user. These SDL checks may include for example limiting the number of dimensions in the table, minimum population and sparsity thresholds, ensuring consistent and nested categories of variables to avoid disclosure by differencing. If the requested table does not meet the criteria, it is not released through the server and the user is advised to redesign the table.

For flexible table generating, the server has to quantify the disclosure risk in the original table, apply an SDL method and then reassess the disclosure risk. Obviously, the disclosure risk will depend on whether the underlying data is a whole population (census) and the zeros are real zeros, or the data are from a survey and the zeros may be random zeros. After the table is protected, the server should also calculate the impact on data utility by comparing the perturbed table to the original table. Measures based on Information Theory described in Section 2.3 can be used to assess disclosure risk and data utility in a flexible table generating server since they can be calculated on-the-fly.

Whilst the online flexible table generators have the same types of disclosure risks described in Section 2.3, the disclosure risks based on disclosure by differencing and disclosure by linking tables which form the basis for inferential disclosure need to be considered since there are no interventions or manual checks on what tables are produced or how many times tables are generated. Therefore, for online flexible generating servers, the statistical community has recognized the need for post-tabular perturbative methods on the generated tables to protect against disclosures (Shlomo, et al. 2015) and hence have explored the differential privacy standard.

The ABS approach of the table builder which uses the probability transition matrix **P** to perturb discrete cell counts as described in Section 2.3 has the potential of transforming into a differential privacy perturbation mechanism under certain restrictions. One characteristic of the ABS table builder is that for any cell that is generated from their census microdata, the perturbation of the cell value will always be the same. Fraser and Wooton (2005) describe the 'same cell-same perturbation' approach where each individual in the microdata is assigned a random number. Any time individuals are aggregated to form a cell in a table, their random numbers are also aggregated and this becomes the seed for the perturbation. Therefore, the same cell will always have the same perturbation. This reduces the chance of identifying the true cell value through multiple requests of the table and averaging out the perturbations.

According to this setting, all possible tables and all possible cells that can be generated in the flexible table generating server are essentially known in advance and hence can be protected under a given privacy budget ε in (1). This is known as a non-interactive mechanism in the theory of differential privacy and any post-processing of a differentially private output will still be differentially private.

Rinott et al. (2018) propose using a differentially private exponential mechanism (McSherry, et al. 2007) based on a utility function: $u(a, b)$ described as follows:

Given a list of all possible cells $k = 1, \ldots, K$: $a = (a_1, \ldots, a_K)\epsilon A$ choose output $b = (b_1, \ldots, b_K)\epsilon B$ with probability proportional to

$$\exp\left(\frac{\frac{\varepsilon}{2}u(a,b)}{\Delta u}\right) \qquad (3)$$

where ε is the privacy budget and the scale is defined as: $\Delta \mathrm{u} = \max_{b \in B} \max_{a,a\prime \in A} |u(a, b) - u(a', b)|$where a and a' are neighboring databases that differ by removing one individual. $\Delta \mathrm{u}$ is also known as the sensitivity in the differential privacy mechanism. The utility function is defined through a loss function: $u(a, b) = -l_1 = \sum_{k=1}^{K} |a_k - b_k|$. Under this definition, for a list of internal cells where an individual appears only once, the sensitivity Δu would be 1. If marginal totals are also included and an individual appears several times in the list then Δu will increase. This mechanism is essentially a discretized Laplace distribution and is optimal for the case of perturbing count data with respect to preserving utility.

Furthermore, bounding the perturbations such that $|a_k - b_k| \leq m$ for all k leads to (ε, δ)-differential privacy where δ is the probability in (2) at the cap m. Other implications under differential privacy for an online flexible table generating server compared to SDL methods are (1) zero cell values (unless they are true structural zeroes) must be perturbed; (2) the perturbation may cause negative values in the generated tables and in these cases, the perturbed cell value can be returned as a zero; (3) to preserve additivity, the margins can be perturbed separately (albeit with a larger Δu) and iterative proportional fitting can be carried out so that the sum of the perturbed internal cells equal the perturbed margins as this will not affect the differential privacy guarantee.

The notion of 'same cell-same perturbation' as described above which informs the seed for perturbation and underpins the definition of the non-interactive mechanism and a fixed privacy budget, fails in differential privacy where it is assumed that the attacker knows the entire database except for one target individual. In that case, the attacker can generate the same table on neighboring databases a and a' resulting in only one cell being changed and hence learn in which cell the individual belongs. Therefore the seed for perturbation will need to also account for the domain total of the table and this implies a privacy loss due to some lack of consistencies, i.e. same cells in tables with the same domain total will have the same perturbation but may change their perturbation across tables with different domain totals. More research is needed on the quantification of the privacy loss in this case.

In Figure 2, we show the utility of differential privacy through an examination of the Cramer's V statistic (a normalized Chi-square statistic) in the top row and an l_1 distance metric on the bottom row for a census table derived from the United Kingdom 2001 census spanned by 5 year age-groups and occupation classes for a specific area where the population is European born. We include different specifications of ε on the x-axis. We also add in a discretized Gaussian exponential mechanism for comparison in the right panel where: $u(a, b) = -l_2 =$

$\sum_{k=1}^{K}(a_k - b_k)^2$ (the perturbation cap m under the Gaussian exponential mechanism varied slightly to ensure the same values of δ as the discretized Laplace exponential mechanism).

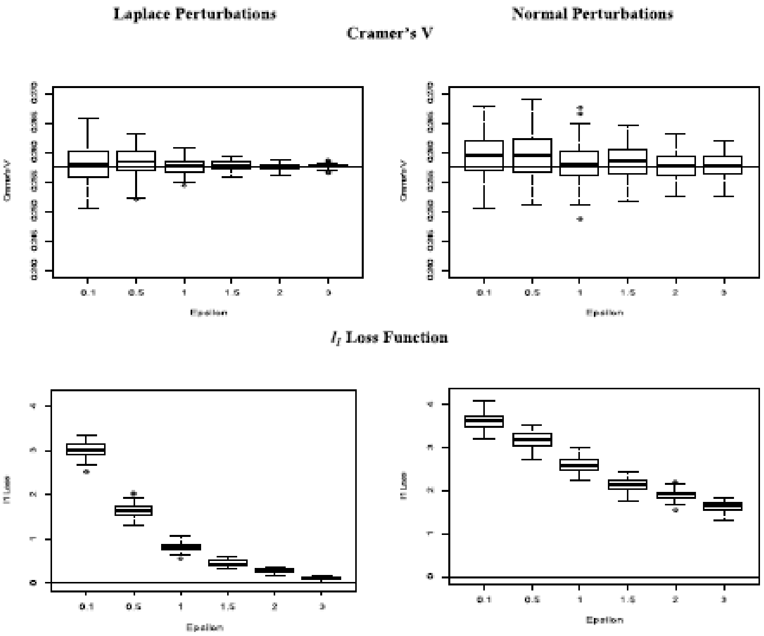


Figure 2: Values of Cramer's V and l_1 loss function over 100 perturbation repetitions for each ε for a UK 2001 Census table

As can be seen in Figure 2, the discretized Laplace perturbations outperformed the Gaussian perturbations with a lower loss function and a more accurate Cramer's V statistic for all levels of ε. For values of ε greater than 1, the utility is not severely impacted under the Laplace perturbations. Since differential privacy is a cryptographic approach and hence the probability mechanism for the perturbation is not secret and can be released, users are able to account for perturbation error in their statistical analysis. Rinott, et al. (2018) demonstrate how to use the parameters of the differential privacy mechanism to account for the perturbation in Chi-square testing for goodness of fit and independence.

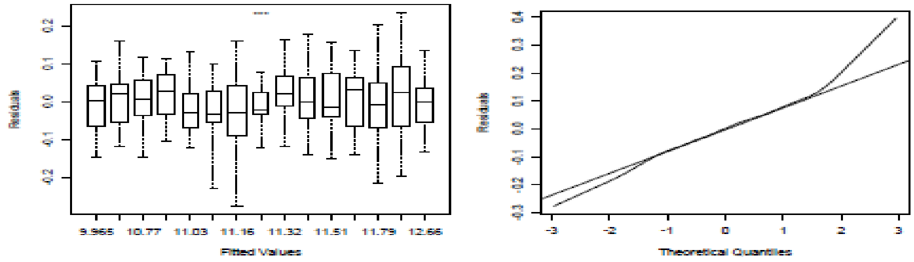
Shlomo et al. (2018) compared two standard SDL methods with differential privacy for a flexible table builder containing survey weighted counts. They showed that for the case of internal cells of tables and relatively large sample counts there was less perturbation required under differential privacy and higher utility compared to the SDL approaches. Other examples of perturbing counts in frequency tables in the computer science literature are Barak, et al. (2007), Yaroslavtsev, et al. (2013) and Qardaji, et al. (2014).

It is now being recognized that the differential privacy approach for protecting frequency tables can be a viable technique in the SDL framework at statistical agencies. Open questions remain and are subject to future research. Whilst the use of the non-interactive differential privacy mechanism will avoid depleting a privacy budget under multiple generation of tables, how to set this budget and determine the sensitivity of the mechanism given the large scale dissemination of tables containing internal and marginal cells need careful consideration. In addition, policy makers need to understand the consequences of the privacy parameters ε and δ and this work is ongoing.

4.1.2. Remote Analysis Servers

A remote analysis server is an online system which accepts a query from the researcher, runs it within a secure environment on the underlying data and returns a confidentialized output without the need for human intervention to manually check the outputs for disclosure risks. Similar to flexible table generating servers, the queries are submitted through a remote interface and researchers do not have direct access to the data. The queries may include exploratory analysis, measures of association, regression models and statistical testing. The queries can be run on the original data or confidentialized data and may be restricted and audited depending on the level of required protection. O'Keefe and Good (2008) describe regression modeling via a remote analysis server.

O'Keefe and Shlomo (2012) compared outputs based on original data and two SDL approaches: outputs from confidentialized microdata (where outliers were removed, additive noise added to the continuous variables and coarsening of the geography variable) and confidentialized outputs obtained from the original data via a remote analysis server. As an example, Figure 3 shows what residual plots would look like in a remote analysis server through a series of sequential box plots and a smoothed Normal QQ plot.



(a) Residuals by fitted values

(b) Normal QQ Plot of Residuals

Figure 3: Confidential Residual plot from a regression analysis in a remote analysis server

Under the confidentialized output approach in a remote analysis server, no single observations can be learnt and hence there are no maximum, minimum values or percentiles in the outputs. However, one can argue that there is more utility in the confidentialized output approach since distributions are not distorted due to the removal of outliers and other perturbation methods as would be the case when confidentializing the microdata.

Moreover, differential privacy can be applied to the underlying algorithms underpinning the remote analysis server to ensure a privacy-by-design approach. For example, a remote analysis server typically uses a robustified regression model in order to down-weight any outliers in the data. Alternatively, one might consider the approach proposed by Chipperfield and O'Keefe (2014) where a small unit-level noise is added to the estimating equations for the coefficients of the regression model. In addition, Laplace noise can be added to summary statistics in an exploratory analysis and this can be carried out consistently similar to the non-interactive mechanism approach proposed for flexible table builders so that the privacy budget will not be exhausted. These areas for implementation of differential privacy have yet to be explored.

4.2 Safe Data Enclaves and Remote Access

To meet increasing demands for high resolution data, many statistical agencies and data archives have set up data enclaves on their premises where approved researchers can go onsite and gain access to confidential statistical data. The secure servers within the enclave have no connection to printers or the internet and only authorized researchers are allowed to access them. To minimize disclosure risk, no data can be removed from the enclave and researchers undergo specialized training to understand the confidentiality guidelines. Researchers are generally provided with standard software within the system, such as STATA, SAS and R, but any specialized software would not be available. All information flow is controlled and monitored. Any outputs to be taken out of the data enclave are dropped in a folder and manually checked by experienced confidentiality officers for disclosure risks. Examples of disclosure risks in outputs are small cell counts in tables, residual plots from regression models which may highlight outliers and Kernel density estimation with small band-widths.

The disadvantage of the data enclave is the need to travel, sometimes long distances, to access confidential data. In recent years, some agencies have implemented remote access by extending the concept of the data enclave to a 'virtual' data enclave. These 'virtual' data enclaves can be set up at other government agencies, universities and even on a researcher's own laptop. Trusted approved users log on to secure servers via VPN connections to access the confidential data. All activity is logged and audited at the keystroke level and outputs are reviewed remotely by confidentiality officers before being sent back to the researchers via a secure file transfer protocol site. The technology also allows users within the same research group to interact with one another while working on the same dataset. An example of this technology is the Inter-university Consortium for Political and Social Research (ICPSR) housed at the University of Michigan. The ICPSR maintains access to data archives of social science data for research and operates both a physical on-site data enclave and a 'virtual' data enclave.

4.3. Synthetic Data

In recent years, there have been initiatives to produce synthetic microdata as public-use files which preserve some of the statistical properties of the original microdata. This allows freely available open data which can be used by researchers to plan their research questions and data analysis and prepare their code as well as for teaching purposes. The data elements are replaced with synthetic values generated from an appropriate probability model. The model is fit to the original data to produce synthetic populations through a posterior predictive distribution similar to the theory of multiple imputation. Several samples are drawn from the population to take into account the uncertainty of the model and to obtain proper variance estimates. See Raghunathan, Reiter and Rubin (2003) and Reiter (2005b) and references therein for more details of generating synthetic data. The synthetic data can be implemented on parts of data so that a mixture of real and synthetic data is released (Little and Liu, 2003) although this means that a thorough disclosure risk assessment is needed prior to releasing such data. In practice it is very difficult to capture all conditional relationships between variables and within sub-populations. If models used in a statistical analysis are sub-models of the model used to generate data, then the analysis of multiple synthetic samples should give valid inferences.

Synthetic values have also been proposed for magnitude tables arising from business statistics as described in Section 2.4. The traditional method of cell suppression in magnitude tables leads to a loss of information and there have been more recent initiatives to provide synthetic values for the suppressed cells. Controlled tabular adjustment (CTA) carries out cell suppression and replaces the suppressed cells with synthetic values that guarantee some statistical properties as well as the marginal totals (Dandekar and Cox, 2002).

The subject of using differential privacy in the production of synthetic data is still undergoing research. One early application which generated synthetic data using a differential privacy mechanism embedded in the Bayesian Multinomial- Dirichlet model is the US Census Bureau 'On the Map' available at: http://onthemap.ces.census.gov/. It is a web-based mapping and reporting application that shows where workers are employed and where they live according to the Origin-Destination Employment Statistics. More information is given in Abowd and

Vilhuber (2008). However, this application was limited in that the dataset only contained a set of counts. More research is needed on whether synthetic data can be generated from microdata containing many different types of variables. Some avenues to explore are to add the differentially private noise in the Bayesian predictive modelling (similar to 'On the Map') or to use a sequential regression modelling approach for generating synthetic data (See Ragunathan, et al. 2001 and Van Buuren, 2007) and adding differential private noise to the estimating equations in each iteration. The US Census Bureau is currently exploring reproducing census microdata from many tables that have been protected under the differential privacy mechanism. As mentioned, differentially private synthetic data is still an open area of research.

5. Discussion

In recent years, statistical agencies and data archives have been restricting access to statistical data due to their inability to cope with the large demand for data whilst ensuring the confidentiality of statistical units. However, with government initiatives for more open and accessible data, statistical agencies are exploring alternative means for disseminating statistical data which allows for more use of the internet. Given the rising concerns of inferential disclosure under these new dissemination strategies, this has led to fruitful collaborations between statisticians and computer scientists and initial research on whether the formal 'by-design' privacy guarantee of differential privacy can be embedded in the SDL framework.

The SDL framework for protecting against identity and attribute disclosures is still very much relevant at statistical agencies since it is part of the legal and ethical framework underpinning the dissemination of traditional types of statistical data and there is no move to stop current dissemination practices. However, when considering more flexible dissemination via the internet, it is necessary for statistical agencies to move towards perturbation as a viable way of protecting the confidentiality of data subjects. Additive noise perturbation under differential privacy within the SDL framework is still in the beginning stages of research. For count data, it has been shown to have good utility. Perturbative methods, however come at a cost in that researchers will have to cope with the perturbation when carrying out statistical analysis which may require more training. Clearly, the fact that the parameters of the differential privacy mechanism are not secret and can be used to correct statistical analysis of perturbed data provides a large incentive for introducing differential privacy into the SDL framework.

References

Abowd, J.M. and Vilhuber, L., (2008). How Protective Are Synthetic Data? In *PSD'2008 Privacy in Statistical Databases*, (Eds. J.Domingo-Ferrer and Y. Saygin), Springer LNCS 5262, 239-246.

Antal, L., Shlomo, N. and Elliot, M. (2014) Measuring Disclosure Risk with Entropy in Population Based Frequency Tables. In Privacy in Statistical Databases 2014, (Ed. J. Domingo-Ferrer), Springer LNCS 8744, pp. 62-78.

Antal, L., Shlomo, N., and Elliot, M. (2015) Disclosure Risk Measurement with Entropy in Two-Dimensional Sample Based Frequency Tables. *Joint UNECE/Eurostat work session on statistical data confidentiality, Helsinki, October 2015* https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/20150/Paper_14_Session_1_-_Univ._Manchester.pdf

Barak, B., Chaudhuri, K., Dwork, C., Kale,S., McSherry,F., and Talwar,K. (2007). Privacy, accuracy, and consistency too: a holistic solution to contingency table release. Symposium on Principles of database systems, ACM, 2007, 273-282.

Bethlehem, J., Keller, W., and Pannekoek, J. (1990) Disclosure limitation of Microdata. Journal of the American Statistical Association 85, 38–45.

Chipperfield, J.O. and O'Keefe, C.M. (2014). Disclosure-protected inference using generalised linear models, International Statistical Review, Vol. 82 (3), 371-391.

Dandekar, R.A. and Cox L. H. (2002). Synthetic Tabular Data: An Alternative to Complementary Cell Suppression. *Manuscript, Energy Information Administration*, U. S. Department of Energy.

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

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

.Duncan, G., Keller-McNulty, S., and Stokes, S. (2001) Disclosure Risk vs. Data Utility: the R-U Confidentiality Map. *Technical Report LA-UR-01-6428*. Statistical Sciences Group,Los Alamos, N.M.:Los Alamos National Laboratory.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography TCC* (eds. S. Halevi and R. Rabin). Heidelberg: Springer, LNCS 3876, 265-284.

Dwork, C. and Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9, 211-407.

Elamir, E. and Skinner, C.J. (2006) Record-Level Measures of Disclosure Risk for Survey Micro-data. Journal of Official Statistics, 22, 525–539.

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 Ran- 90 domisation for Statistical Disclosure limitation: Theory and Implementation. Journal of Official Statistics, 14, 463–478.

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 inSurvey Methodology. John Wiley & Sons, United Kingdom.

Karr, A., Kohnen, C. N., Organian, A., Reiter, J. P. and Sanil, A. P. (2006) A framework for evaluating the utility of data altered to protect confidentiality. *Am. Statistn*, **60**, 224–232.

Little, R.J.A., and Liu, F. (2003). Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata. *The University of Michigan Department of Biostatistics Working Paper Series.* Working Paper 6.

Manrique-Vallier, D. and Reiter, J.P. (2012). Estimating Identification Disclosure Risk Using Mixed Membership Models. Journal of the American Statistical Association, Vol. 107 (500), 1385-1394.

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, New York.

O'Keefe, C.M. and Good, N. (2008). A Remote analysis Server – What Does Regression Output Look Like? In *PSD'2008 Privacy in Statistical Databases*, (Eds. J.Domingo-Ferrer and Y. Saygin), Springer LNCS 5262, 270-283.

O'Keefe, C.M. and Shlomo, N. (2012). Comparison of Remote Analysis with Statistical Disclosure Control for Protecting the Confidentiality of Business Data. *Transactions on Data Privacy*, Vol. 5, Issue 2, 403-432.

Qardaji, W., Yang, W. and Li, N. (2014). Preview: practical differentially private release of marginal contingency tables. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data, ACM, 2014. 1435–1446.

Raghunathan T.E., Lepkowksi J.M., van Hoewyk J., Solenbeger P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, Vol. 27, 85-95.

Raghunathan, T.E., Reiter, J. and Rubin, D. (2003). Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics*, 19, No. 1, 1-16.

Reiter, J.P. (2005a) Estimating Risks of Identification Disclosure in Microdata. *Journal of the American Statistical Association* 100, 1103-1112.

Reiter, J.P. (2005b), Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study. *Journal of the Royal Statistical Society*, A, Vol.168, No.1, 185-205.

Rinott, Y., O'Keefe, C., Shlomo, N., and Skinner, C. (2018) Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. *Statistical Sciences*, Vol. 33, No. 3 (2018), 358-385.

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*.

Shlomo, N., Antal, L. and Elliot, M. (2015) Measuring Disclosure Risk and Data Utility for Flexible Table Generators, *Journal of Official Statistics,* Vol. 31, Issue 2, 305-324.

Shlomo, N., Krenzke, T. and Li, J. (2018) Confidentiality Protection Approaches for Survey Weighted Frequency Tables. Available at: http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2018/CMI_Working_Paper_Comparison_Post-tabular_Confidentiality_Approaches.pdf

Shlomo, N. and Skinner. C.J. (2012). Privacy Protection from Sampling and Perturbation in Survey Microdata. *Journal of Privacy and Confidentiality*, Vol. 4, Issue 1.

Shlomo, N. and Skinner, C.J. (2010). Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata. *Annals of Applied Statistics,* Vol. 4, No. 3, 1291-1310.

Shlomo, N. and Young, C. (2008). Invariant Post-tabular Protection of Census Frequency Counts. In *PSD'2008 Privacy in Statistical Databases*, (Eds. J.Domingo-Ferrer and Y. Saygin), Springer LNCS 5262, 77-89.

Skinner, C.J. and Holmes, D. (1998) Estimating the Re-identification Risk Per Record in Microdata. *Journal of Official Statistics* 14, 361–372

Skinner, C.J. and Shlomo, N. (2008) Assessing Identification Risk in Survey Microdata Using Log-linear Models. *Journal of American Statistical Association*, Vol. 103, Number 483, 989–1001.

Snoke, J., Raab, G.M., Nowok, B., Dibben C. and Slavkovic, A. (2018) General and Specific Utility Measures for Synthetic Data. *Journal of the Royal Statistical Society*. Series A, Vol 181, Issue 3, 663-688.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, Vol. 16(3), 219-242.

Willenborg, L. and De Waal, T. (2001) *Elements of Statistical Disclosure Control in Practice*. Lecture Notes in Statistics, 155. New York: Springer-Verlag.

Woo, M.-J., Reiter, J. P., Oganian, A. and Karr, A. F. (2009) Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, **1**, 111–124.

Yaroslavtsev, G., Cormode, G., Procopiuc, C.M., and Srivastava, D. (2013). Accurate and efficient private release of datacubes and contingency tables. In *ICDE,* 2013.