Statistical Disclosure Control: New Directions and Challenges

Natalie Shlomo[1]
[1] University of Manchester, Manchester, UK
e-mail: Natalie.Shlomo@manchester.ac.uk

1. Introduction

Traditionally, statistical agencies generally release outputs in the form of microdata and tabular data. Microdata contain data from social surveys and tabular data contain either frequency counts, such as for census dissemination, or magnitude data typically arising from business surveys, eg. total revenue. For each of these traditional outputs, there has been much research on how to quantify disclosure risk, optimal statistical disclosure control (SDC) methods and how to assess the impact on data utility.

For traditional outputs, the two main disclosure risks are identity disclosure where a statistical unit can be identified based on a set of identifying variables and attribute disclosure where new information can be learnt about an individual or a group of individuals. One disclosure risk that is often overlooked in traditional statistical outputs is inferential disclosure. This disclosure risk has to do with learning new attributes with high probability. 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 from this type of disclosure risk. Another example of inferential disclosure is disclosure by differencing when multiple releases are disseminated from one data source. For example, census tables can be differenced or manipulated to reveal individual units. For traditional hard-copy census tables, disclosure by differencing is controlled by having a fixed set of variables and categories which disallow differencing non-nested groups of individuals.

In Section 2 of this paper we provide more discussion of inferential disclosure and define differential privacy as developed in the computer science literature for the protection of outputs in on-line query systems. Inferential disclosure is now a key risk that statistical agencies need to consider when developing new online and remote strategies for disseminating statistical outputs. Section 3 describes recent advances in data dissemination with some examples. Section 4 concludes with a discussion.

2. Disclosure Risk in Query Systems and Differential Privacy

The theory of differential privacy for protecting outputs in a remote query-based system has been widely presented in the computer science literature and it is closely related to the concept of inferential disclosure. See Dinur and Nissim, 2003 and Dwork, McSherry, Nissim, and Smith, 2006, for more details on differential privacy. Shlomo and Skinner, 2012 discuss differential privacy with respect to sampling and perturbation. A ‘worst case’ scenario is allowed for, in which the intruder has complete information about all the units in the database except for one unit of interest. Under this assumption, let x denote a cell value, taking possible values $1,\ldots,k$ where the table is formed by cross-classifying all variables in the database. The population database is denoted by $\mathbf{x}_U=(x_1,....,x_N)$ where N denotes the size of the population U={1,…,N}. We assume that sampling would be part of the SDC mechanism to protect the population database and that the intruder would not have response knowledge of who is in the sample. Let $\tilde{x}_i$ denote the cell value of unit i in the database after an SDC method has been applied and define $\tilde{f}_j=\sum_{i\in s} I(\tilde{x}_i=j)$ as the observed count in cell j where s denotes a sample drawn randomly from the population. We can view the released data as the vector of counts: $\tilde{\mathbf{f}}=(\tilde{f}_1,\tilde{f}_2,...,\tilde{f}_k)$. Let $\Pr(\tilde{\mathbf{f}}\,|\,\mathbf{x}_U)$ denote the probability of $\tilde{\mathbf{f}}$ with respect to an

SDC method, which includes sampling and/or perturbation, and where $\mathbf{x}_U$ is treated as fixed. In this framework, the definition of differential privacy is as follows:

Definition (Dwork et al.,2006)*:* ε - differential privacy holds if: $\max\left|\ln\left(\frac{\Pr[\tilde{\mathbf{f}} \mid \mathbf{x}_U^{(1)}]}{\Pr[\tilde{\mathbf{f}} \mid \mathbf{x}_U^{(2)}]}\right)\right| \le \varepsilon$ for some $\varepsilon > 0$, where the maximum is over all pairs $(\mathbf{x}_U^{(1)}, \mathbf{x}_U^{(2)})$, which differ in only one element and across all possible values of $\tilde{\mathbf{f}}$.

Differential privacy therefore aims to avoid inferential disclosure by ensuring that an intruder cannot make inference about a single unit when only one of its value is changed given that all other units in the population are known. This definition would control for disclosure by differencing and highly predictive models which are now problematic when considering online query systems to disseminate statistical data compared to traditional hard-copy outputs. 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.

3. New Dissemination Strategies

3.1 Data Enclaves and Remote Access

In the last decade, many statistical agencies have set up data enclaves on their premises. These are largely motivated by the statistical agencies lack of means to continue to meet demands for large amounts of data and still ensure the confidentiality of individuals. The data enclave is a secure environment where researchers go on-site and gain access to confidential data. The secure servers have no connection to printers or the internet and only authorized users are allowed to access them. To minimize disclosure risk, no data is allowed to be removed from the enclave and researchers undergo training to understand the security rules. Researchers are generally provided with software in the system, such as STATA and R. All information flow is controlled and monitored. Any outputs to be taken out of the data enclaves 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 which may denote 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 very recent years, some statistical agencies have been extending the concept of the data enclave to remote access through a ‘virtual’ data enclave. These ‘virtual’ data enclaves have been set up at organizations and Universities and allow users within the same group to interact with one another while working with the data. Users log on to secure servers to access the confidential data and all activity is logged and audited at the keystroke level. Generally, the secure data lab must be approved by the agencies and outputs are reviewed remotely by confidentiality officers before being sent back to the researchers via a secure file transfer protocol site.

3.2 Web-based Applications

There are two types of web-based applications that are being considered for disseminating statistical outputs: flexible table generators and remote analysis servers.

3.2.1 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 considering the development of 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 and the Israel Central Bureau of Statistics developed a server to disseminate tables from the Social Survey. 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.

When selecting the SDC method to apply to the output table, 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 since it will compound the SDC impact and overprotect the data whilst reducing data utility. The post-tabular approach is also motivated by the computer science definition of differential privacy as discussed in Section 2.

For flexible table generating, the server has to quantify the disclosure risk in the original table, apply an SDC 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 can be used to assess disclosure risk and data utility in a table generating server. For measuring disclosure risk, the entropy takes into account whether the distributions in the tables are skewed or uniform leading to the risk of attribute disclosure. The entropy can be adapted depending on whether the data is from a census or a survey. Combining the entropy with other measures which account for the proportion of zeros (real or random) in the table and the size of the population in the table provides further information on the disclosure risks (Shlomo, Antal and Elliot, 2013). For data utility, the Hellinger Distance between the perturbed and original table provides insight on the impact of the SDC method.

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 tables that should not be released. These SDC rules may include limiting the number of dimensions in the table, minimum population thresholds, ensuring consistent and nested categories of variables to avoid disclosure by differencing, etc.

Pre-tabular SDC methods may include record swapping where attributes are swapped between two records having similar characteristics through a set of control variables. Post-tabular methods may include cell perturbation such as random rounding or the post-randomization method which perturbs cell counts based on a probability transition matrix. The SDC method should ensure that sufficient statistics are preserved, such as the marginal and overall totals. When carrying out post-tabular stochastic perturbation methods, consistency across same cells generated in different tables is necessary to avoid the possibility of ‘unpicking’ the SDC method. This can be carried out by microdata keys. For each record in the microdata, a random number is defined which when combined with other records to form a cell of a table defines the seed for the perturbation. Records that are aggregated into same cells will always have the same seed and therefore a consistent perturbation (Fraser and Wooton, J, 2005, Shlomo and Young, 2008).

In table 1 we compare different SDC methods for a census table defined in one region of the United Kingdom according to banded age groups, education qualification and occupation. The table contained 2,457 cells where 62.4% were real zeros. The underlying data in the flexible table generating server was a very large hypercube instead of the original census microdata. The hypercube provides a priori protection since no units below the level of the cells of the hypercube are disseminated. We compare three pre-tabular methods on the hypercube: record swapping, semi-controlled random rounding and a stochastic perturbation, and a post-tabular method of semi-controlled random rounding applied directly to the output table. The measures are based on Information Theory where the upper bound for the disclosure risk measure is one and the upper bound for the Hellinger distance is the square root of the population size in the table and therefore is comparable across methods.

Table 1: Disclosure Risk and Information Loss for the Generated Table

	Disclosure Risk	Hellinger Distance
Original	0.3520	-
Perturbed Input		
Record Swapping	0.3505	6.469
Semi-controlled Random Rounding	0.2374	7.970
Stochastic Perturbation	0.2303	14.12
Perturbed Output		
Semi-Controlled Random Rounding	0.2327	5.902

From table 1, it is clear that the method of record swapping when applied to the input data did little to reduce the disclosure risk in the final output table. This was due to the fact that the small cells remain unperturbed in the table. Record swapping provides the smallest distance metric between the original and perturbed table compared to the other pre-tabular methods as expected given that it had the lowest level of protection against disclosure risk. From among the input perturbation methods, the stochastic perturbation provided the most protection against disclosure but at the cost of low data utility with the highest distance metric between the original and perturbed table. Removing the small cells entirely by rounding provides lower disclosure risk but better utility than the stochastic perturbation. Comparing the pre-tabular and post-tabular semi-controlled random rounding procedure, we see slightly lower disclosure risk according to the post-tabular rounding but more 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 and can be adapted to guarantee differential privacy.

3.2.2 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 generators, 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, 2009 describe regression modeling via a remote analysis server.

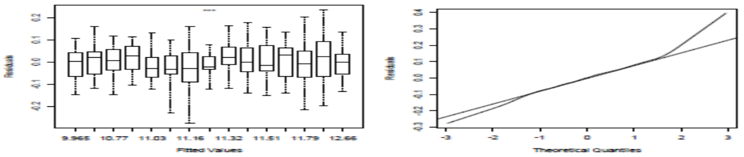
O’keefe and Shlomo, 2012 compared outputs based on original data and two SDC approaches: outputs from confidentialized microdata and confidentialized outputs obtained from the original data via a remote analysis server. The comparison was carried out on a dataset from the 1982 survey of the sugar cane industry in Queensland, Australia (Chambers and Dunstan,1986). The dataset corresponds to a sample of 338 Queensland sugar farms and contained the following variables: region, area, harvest, receipts, costs and profits (equal to receipts minus costs). The dataset was confidentialized by deleting large outlier farms, coarsening the variable area and adding random noise to harvest, receipts, costs and profits. Figure 1 shows what the residual plot would look like in a remote analysis server where the response variable is receipts and the explanatory variables: region, area, harvests and costs. Figure 2 presents the comparison for the univariate analysis of receipts.

3.3 Synthetic Data

Basic confidential data is a fundamental product of virtually all statistical agency programs. These lead to the publication of public-use products such as summary data, microdata from surveys, etc. Confidential data may also be used for internal use within data enclaves. In recent years, there has been a move to produce synthetic microdata as public-use files which preserve some of the statistical properties of microdata. The data elements are replaced with synthetic values sampled 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 variance estimates. See Raghunathan, Reiter and Rubin, 2003, and Reiter, 2005 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). One application which uses partially synthetic data 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.

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. In addition, partially synthetic datasets may still have disclosure risks and need to be checked prior to dissemination.

For tabular data there are also techniques to develop synthetic magnitude tables arising from business statistics. Controlled tabular adjustment (CTA) carries out cell suppression and replaces the suppressed cells with imputed values that guarantee some statistical properties (Dandekar and Cox, 2002).



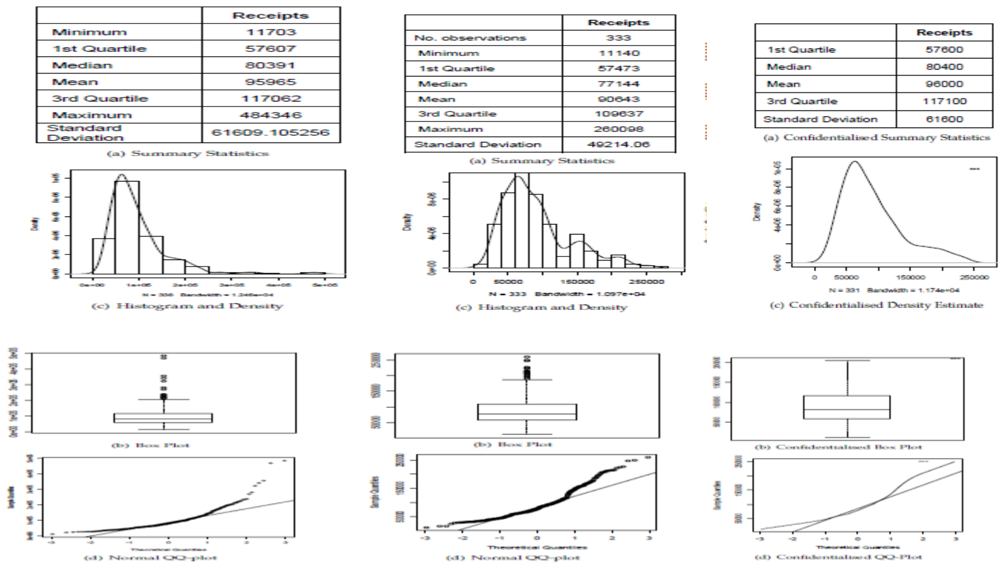
(a) Residuals by fitted values

(b) Normal QQ Plot of Residuals

Figure 1: Confidential Residual plot from a regression analysis on receipts

4. Discussion

In recent years, statistical agencies have been restricting access to statistical data due to the inability to cope with the large demand for data whilst ensuring the confidentiality of statistical elements. However, with government initiatives for ‘open data’, new ways to disseminate statistical data are being explored. This has led to more cooperation with computer scientists who have developed formal definitions of disclosure risk, particularly for inferential disclosure, and SDC methods that guarantee privacy. These methods come at a cost in that researchers will have to cope with perturbed data when carrying out statistical analysis which may require more training. In addition, statistical agencies need to release the parameters of the SDC methods so that researchers can account for the measurement error in their analysis.



	Receipts
Minimum	11703
1st Quartile	57607
Median	80391
Mean	95965
3rd Quartile	117062
Maximum	484346
Standard Deviation	61609.105256

(a) Summary Statistics

	Receipts
No. observations	333
Minimum	11140
1st Quartile	57473
Median	77144
Mean	90643
3rd Quartile	109637
Maximum	260098
Standard Deviation	49214.06

(a) Summary Statistics

	Receipts
1st Quartile	57600
Median	80400
Mean	96000
3rd Quartile	117100
Standard Deviation	61600

(a) Confidentialised Summary Statistics

(c) Histogram and Density

(c) Histogram and Density

(c) Confidentialised Density Estimate

(b) Box Plot

(b) Box Plot

(b) Confidentialised Box Plot

(d) Normal QQ-plot

(d) Normal QQ-plot

(d) Confidentialised QQ-Plot

Original **Confidential Input** **Confidential Output**

Figure 2: Univariate analysis of receipts for the Sugar Canes dataset

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.

Chambers, R. and Dunstan, R. (1986). Estimating Distribution Functions from Survey Data. *Biometrika*, 73, 597–604.

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.

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.

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.

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.

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.

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. (2005), 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.

Shlomo, N. Antal, L. and Elliot, M. (2013). Disclosure Risk and Data Utility in Flexible Table Generators**.** *Proceedings of the NTTS2013 Conference*, Brussels, March 5-7.

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 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.