2016

A new algorithm for protecting aggregate business microdata via a remote system

Yue Ma
University of Wollongong, ym894@uowmail.edu.au

Yan-Xia Lin
University of Wollongong, yanxia@uow.edu.au

James O. Chipperfield
University of Wollongong

John Newman
Australian Bureau Of Statistics

Victoria Leaver
Australian Bureau of Statistics, vll881@uowmail.edu.au

Publication Details

Ma, Y., Lin, Y., Chipperfield, J., Newman, J. & Leaver, V. (2016). A new algorithm for protecting aggregate business microdata via a remote system. Lecture Notes in Computer Science, 9867 210-221.

A new algorithm for protecting aggregate business microdata via a remote system

Abstract

Releasing business microdata is a challenging problem for many statistical agencies. Businesses with distinct continuous characteristics such as extremely high income could easily be identified while these businesses are normally included in surveys representing the population. In order to provide data users with useful statistics while maintaining confidentiality, some statistical agencies have developed online based tools to allow users to specify and request tables created from microdata. These tools only release perturbed cell values generated from automatic output perturbation algorithms in order to protect each underlying observation against various attacks, such as differencing attacks. An example of the perturbation algorithms has been proposed by Thompson et al. (2013). The algorithm focuses largely on reducing disclosure risks without addressing much on data utility. As a result, the algorithm has limitations, including a limited scope of applicable cells and uncontrolled utility loss. In this paper we introduce a new algorithm for generating perturbed cell values. As a comparison, The new algorithm allows more control over utility loss, while it could also achieve better utility-disclosure tradeoffs in many cases, and is conjectured to be applicable to a wider scope of cells.

Keywords

protecting, aggregate, algorithm, business, microdata, system, via, remote

Disciplines

Engineering | Science and Technology Studies

Publication Details

Ma, Y., Lin, Y., Chipperfield, J., Newman, J. & Leaver, V. (2016). A new algorithm for protecting aggregate business microdata via a remote system. Lecture Notes in Computer Science, 9867 210-221.

A New Algorithm for Protecting Aggregate Business Microdata via a Remote System

Yue Ma[1], Yan-Xia Lin[1], James Chipperfield[12],
John Newman[2] and Victoria Leaver[2]

[1] National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong
[2] Australian Bureau of Statistics. Disclaimer: Views expressed in this paper are those of the author(s) and do not necessarily represent those of the Australian Bureau of Statistics. Where quoted or used, they should be attributed clearly to the author.

Abstract. Releasing business microdata is a challenging problem for many statistical agencies. Businesses with distinct continuous characteristics such as extremely high income could easily be identified while these businesses are normally included in surveys representing the population. In order to provide data users with useful statistics while maintaining confidentiality, some statistical agencies have developed online based tools to allow users to specify and request tables created from microdata. These tools only release perturbed cell values generated from automatic output perturbation algorithms in order to protect each underlying observation against various attacks, such as differencing attacks. An example of the perturbation algorithms has been proposed by Thompson et al. (2013). The algorithm focuses largely on reducing disclosure risks without addressing much on data utility. As a result, the algorithm has limitations, including a limited scope of applicable cells and uncontrolled utility loss. In this paper we introduce a new algorithm for generating perturbed cell values. As a comparison, The new algorithm allows more control over utility loss, while it could also achieve better utility-disclosure tradeoffs in many cases, and is conjectured to be applicable to a wider scope of cells.

Keywords: Business data; Output Perturbation; Remote Access; Continuous Tabular Data; Statistical Disclosure Control

1 Introduction

Disseminating data containing confidential information is a challenging issue for many statistical agencies. On the one hand, the released data should not reveal confidential information to the public; on the other hand, the released data should carry enough statistical information to reflect behaviours of the population. To achieve these two conflicting objectives, statistical agencies release confidentialised data to data users. The confidentialised data conceals sensitive information from the public at the expense of some data utility.

However, it is not easy to confidentialise business data. Typically, some industries will be dominated by large businesses whose information is difficult to conceal by existing data masking methods. Non-perturbative data masking, such as top coding (Klein et al. 2014), suppression (Salazar-González 2005) and micro-aggregation (Defays and Nanopoulos 1993), significantly reduce information of continuous data items such as turnover or profit, which are of key interest to data users. Perturbation methods, such as data swapping (Moore 1996), synthetic data (Rubin 1993) and noise addition (Kim and Winkler 1995), cannot efficiently protect businesses with distinct continuous-valued characteristics. As a result, most statistical agencies have taken a cautious approach to releasing business data, and the majority of business data is still released in the form of broad-level tables.

The emergence of remote access may provide a solution to releasing business microdata. Remote access (Blakemore 2001; Reiter 2004) is a virtual system that provides a data analyst with access to a remote system built by a data agency. The data agency stores microdata in the remote system, and the data analyst communicates with the remote system through a query system. The analyst is restricted from viewing the underlying microdata. Instead, the analyst could only obtain statistical outputs of underlying microdata through the following model (see O'Keefe and Chipperfield 2013; Chipperfield and O'Keefe 2014): (1) an analyst submits a query (i.e. request for a table) to the remote system; (2) the remote system modifies or restricts estimates using an automatic algorithm; (3) the system sends the modified output to the analyst. An example of remote access system is American FactFinder (Hawala et al. 2004), which releases confidentialised tabulations of census data to data users.

The reason for a remote system to release confidentialised statistics is to prevent disclosure of confidential values via various methods of attack, the most significant of which is a differencing attack (Lucero et al 2009, Sect. 4.1). A differencing attack reveals a confidential value by taking the difference of two cell totals whose contributing values differ by one. A differencing attack could be very effective on a remote system as the attacker is able to obtain statistical outputs of different underlying microdata with a high degree of freedom.

In this paper we base our discussion on releasing confidentialised totals from business data through a remote system. We assume a remote system could allow users to specify and request tables created from business microdata. Each cell of a table contains a perturbed survey estimate of total computed from a set of surveyed business values as specified by a data user. The algorithm in Thompson et al. (2013) has been shown to perform particularly well for confidentialising totals from business-typed microdata, and hence we have investigated the prospect of implementing it on remote systems to perturb business totals. The algorithm works by adding a perturbation amount to the unperturbed cell estimate to produce a perturbed cell estimate. The perturbation amount follows a parametric distribution which could be adjusted according to the distribution of underlying business values to produce the best result.

It has been proposed that an algorithm of generating perturbed statistical estimates should satisfy the definition of ϵ-differential privacy (Dwork et al. 2006). Such algorithms include adding Laplace distributed noises (Dwork et al. 2006) and other similar variations (Soria-Comas and Domingo-Ferrer 2013; Nissim et al. 2007) to perturb statistical estimates. These algorithms generally sacrifice a large degree of data utility for data confidentiality (see, for example, Sarathy and Muralidhar 2011). The algorithm in Thompson et al. is not designed to achieve ϵ-differential privacy; however, it achieves good utility-disclosure trade-offs for many cells. The details of the algorithm are introduced in Section 2.

A distinct feature of the algorithm in Thompson et al. is that, for a given cell of a table, the algorithm achieves its best performance if the optimal set of parameters for perturbing the cell is used. As developing a program of searching the optimal set of parameters to be used to perturb each cell value is non-trivial, a recent study investigated the outcomes of using one set of parameters to perturb all cells. The set of parameters was selected upon satisfying the requirement of disclosure risks for a few benchmark cells, and the study examined its impact on utility losses of different cells through empirical studies (see reference [22]).

However, there are issues with this configuration. The issues are: 1. The algorithm cannot always generate legitimate cell estimates which fulfill the requirements of both utility loss and disclosure risk. 2. The algorithm could still produce a very perturbed cell value even though the requirement of utility loss is satisfied. 3. The way it trades data utility for data confidentiality may not be the most efficient one. As a result, we are looking for alternative algorithms which could help to solve these issues.

It need to be mentioned that the first issue is not easy to solve completely. The reason is that cells contributed to by a small number of business values, some of which strongly dominate the cell value, are very difficult to perturb in a reasonable manner. Methods for confidentialising such cells require further studies. In this paper, we do not consider this kind of cells. Instead, we focus on perturbing common cells which do not contain strong dominant contributors.

In this paper, a new algorithm for perturbing business totals is proposed. The main point of the new algorithm is to limit the loss of utility, and simulation results show that the new algorithm addresses the issues mentioned above more effectively than the algorithm in Thompson et al.. The new algorithm allows a better control over cell utility losses, achieves better utility-disclosure tradeoffs for many cells, and is conjectured to be able to legitimately perturb a wider range of cells.

This paper is organized as follow. Section 2 describes the algorithm in Thompson et al. of generating perturbed cell values. Section 3 introduces measures of disclosure risk and utility loss. Sections 4 describes the new algorithm. Section 5 discusses advantages of the new algorithm compared with the algorithm in Thompson et al. through simulations. Section 6 concludes the paper.

2 The algorithm in Thompson et al.

Consider any particular cell in a table and let there be n sample units contributing to the cell, where the units are indexed by $i = 1, 2, \cdots n$. Define a continuous valued characteristic (e.g. income or turnover) for the ith unit (e.g. business) by y_i, the estimation weight of y_i by w_i, and the survey estimate of the total is $\hat{s} = \sum_{i=1}^{n} w_i y_i$. We assume $y_1 w_1 \geq y_2 w_2 \geq \cdots \geq y_n w_n$. We call the weighted business values $(y_1 w_1, \cdots y_n w_n)$ **contributor values** to the cell value $\hat{s}$. The algorithm in Thompson et al. (2013) generates a perturbed cell value in the following way:

1. The algorithm identifies the parameters (K, m) to be used.

2. The algorithm generates a perturbation amount p^* from a random variable P^*, and add p^* to the total $\hat{s}$ to generate a perturbed cell value $\hat{s}^*$.

The random variable P^* has the expression $P^* = \sum_{i=1}^{K} (m_i D_i^* H_i^*) y_i w_i$, where K is the number of top contributors in the cell that are used in calculation of P^*; $m = (m_1, \cdots, m_K)$ is a magnitude vector; D_i^* is a random variable taking the value -1 and 1 with equal probability; H_i^* is a random variable centred on 1 and for the purpose of this paper we set H_i to have a symmetric triangular probability density function centered at 1 with width 0.6.

The optimal set of parameters to be used for each cell depends on the distribution of contributing values to the cell estimate. The optimal set of parameters guarantees that the perturbed estimates have the lowest average utility loss subject to having an acceptable disclosure risk. Examples of the optimal choices of magnitude vector when $K = 3$ for different contributor values are given in Table 1.

As mentioned in the Introduction, developing a program of searching the optimal set of parameters to perturb each cell value is non-trivial. One possible remedy is to use a fixed set of parameters to perturb all cells. However, this configuration certainly limits the efficacy of the algorithm as the choice of parameters is not always the optimal one for perturbing many cells.

To evaluate the validity of perturbed estimates generated by an algorithm, we need to define measures of utility loss and disclosure risk. It is important that the perturbed estimates should satisfy both an acceptable level of utility loss and an acceptable level of disclosure risk. In next section, we introduce these measures.

3 Measuring Disclosure Risk and Utility Loss

3.1 Differencing Attack

We measure the disclosure risk with respect to a 'Differencing Attack'. Throughout the paper, without loss of generality, we assume the attacker's target is the largest contributor value y_1w_1, and the attacker also knows the weight of y_1 is equal to one ($w_1 = 1$). The reason for these assumptions is that, normally speaking, the largest contributor value y_1w_1 has the highest disclosure risk against differencing attack than any other contributor value.

Differencing Attack: The attacker uses the difference between two perturbed cell estimates, $\hat{y}_1 = \hat{s}^* - \hat{s}^*_{-1}$ as an estimate of y_1, where $\hat{s}^*_{-1}$ is defined as the same as $\hat{s}^*$ except that the attacker's target, y_1, is dropped from the cell.

To define disclosure risk, we conservatively assume that: 1. the target is in the sample. and 2. the attacker could uniquely identify the target in terms of a set of quasi-identifiers. So the only protection available in a remote system is perturbation. Consequently, perturbation is the focus of how disclosure risk is measured.

3.2 Defining Disclosure

We first describe the process of conducting a differencing attack on a remote system.

Consider the following scenario: suppose a continuous valued characteristic of the ith sample unit is y_i and there are n sample units, and the estimation weight of y_i is w_i. Define $y = (y_1, y_2, \cdots, y_n)$ and $y_1w_1 \geq y_2w_2 \geq \cdots \geq y_nw_n > 0$ with $w_1 = 1$. The attacker estimates y_1 by taking the difference of two perturbed estimates. The estimate of y_1 is a realization of $\hat{Y}_1 = \hat{S}^* - \hat{S}^*_{-1} = y_1 + P^* - P^*_{-1}$, where $\hat{S}^*$ is the underlying random variable for the cell consisting of $(y_1w_1, \cdots, y_nw_n)$ from which a perturbed cell value is drawn; P^* is the perturbation random variable for perturbing the cell; $\hat{S}^*_{-1}$ is the underlying random variable for the cell consisting of $(y_2w_2, \cdots, y_nw_n)$; and P^*_{-1} is the perturbation random variable for perturbing the cell.

Disclosure occurs if the realization of $\hat{Y}_1$ reveals the value of y_1. It is not necessary for the realization of $\hat{Y}_1$ to be exactly equal to y_1- the degree of accuracy required for disclosure must be determined by the statistical agency. The following definition of disclosure risk we adopt is similar to that used by Lin and Wise (2012) and Klein et al.(2014).

Disclosure Risk: We say that the disclosure risk against a differencing attack is the probability that a realization of $\hat{Y}_1$ is within $100\alpha\%$ of the true value y_1. If we define disclosure risk of attacking target value y_1 as $D(y_1)$, then

$$D(y_1) = P(|P^* - P^*_{-1}| < \alpha y_1)$$

We say that α is the definition of disclosure and R is the acceptable disclosure risk. Different values of (R, α) could be justified on the basis of whether the attack is likely to occur. We say that perturbed cell estimates have an acceptable disclosure risk if $D(y_1)$ is less than R.

3.3 Defining Utility Loss

We define the **utility loss** of perturbing a cell as the relative distance between the perturbed cell value and unperturbed cell value. Measuring utility loss by percentage difference between the perturbed estimate and the unperturbed estimate has been widely used in many applications. It is formally introduced by Domingo-Ferrer and Torra (2001) and widely used by other authors in their studies (see Kim and Winkler 1995; Yancey et al. 2002).

As an algorithm produces a perturbed cell value randomly, in order to assess the general performance of the algorithm to perturb a cell in terms of utility loss, we look at the **average utility loss**, which is the expected utility loss of perturbing the cell using the algorithm. It is preferable that the average utility loss to be as low as possible given that $D(y_1) < R$.

4 A New Algorithm to Generate Perturbed Cell Estimates

Now we introduce a new algorithm to generate perturbed cell estimates. Suppose an ordinary cell has contributor values $(y_1w_1, y_2w_2, \cdots, y_nw_n)$, $w_1 = 1$. The new algorithm perturbs the cell value as follows:

1. The statistical agency sets a parameter value of β.

2. Define $\lambda = \hat{s}\beta$, where $\hat{s}$ is the cell value. If n is an even number, the new algorithm generates a perturbed amount $p^* = d_1z_1$ and adds it to $\hat{s}$ to produce a perturbed cell estimate, where d_1 and z_1 are random samples drawn from random variables D_1 and Z_1, respectively. The random variable D_1 takes values -1 or 1 with equal probabilities

and random variable Z_1 is distributed as U_1 or U_2 with equal probabilities, where $U_1 \sim U(0, 0.5\lambda)$ and $U_2 \sim U(1.5\lambda, 2\lambda)$. If n is an odd number, then the new algorithm generates a perturbed amount $p^* = z_2 d_2$ and adds it to $\hat{s}$ to produce a perturbed cell estimate, where z_2 and d_2 are samples drawn from random variables Z_2 and D_2, respectively. The random variable Z_2 has distribution $U(0.5\lambda, 1.5\lambda)$ and D_2 has the same distribution as D_1.

The value of the parameter β is actually the average utility loss of using the algorithm to perturb a cell. Therefore, the value of β could be set according to the requirements of the statistical agency. The disclosure risk of y_1 could also be mathematically determined and the mathematical expressions are given in Tables 2 and 3.

The advantage of splitting up odd and even cases is addressing the differencing attack by guaranteeing that the counts of contributors will go from odd to even or even to odd. It makes it much harder for the perturbation under the set of n contributors and the set of $n-1$ contributors to cancel out if the largest contributor is not strongly dominating the cell.

In next section, we discuss the advantages of using the new algorithm compared to the algorithm in Thompson et al..

5 Discussion of the New Algorithm Against the algorithm in Thompson et al.

5.1 Controlled Utility Loss of Perturbing a Cell

For a given cell with cell value $\hat{s}$, it can be easily seen that the perturbation amount p^* generated by the new algorithm is bounded in $(-2\beta\hat{s}, 2\beta\hat{s})$. As a result, the utility loss of perturbing the cell through the new algorithm is bounded in $(0, 2\beta)$. As a result, both the average utility loss and the maximum utility loss of perturbing the cell could be controlled.

To illustrate this advantage, suppose the contributor values to a cell are $(y_1 w_1, \cdots y_8 w_8) = (25, 25, 25, 25, 25, 25, 25, 25)$. From Table 1, the optimal magnitude values are $(m_1, m_2, m_3) = (0.5, 0.4, 0.3)$. If the cell is perturbed by the algorithm in Thompson et al. with the optimal magnitude values, the average utility loss is 7.54%. However, simulation results show that it is possible that the algorithm generates a extremely high perturbation amount which leads to a 22.3% utility loss to the cell. A perturbed result with such a high utility loss may mislead some data users. In contrast, perturbing the cell by the new algorithm with $\beta = 0.0754$ also gives an average utility loss of 7.54% while the maximum utility loss of perturbing the cell is only 15.08%.

5.2 A Conjectured Wider Applicability

It is conjectured that the new algorithm could legitimately perturb a wider range of cells. A legitimately perturbed cell value should satisfy the requirements of both disclosure risk and utility loss. Recall that we assume the attacker's target is the largest contributor value y_1, and $w_1 = 1$. In the following, we say that a cell could be legitimately perturbed by an algorithm if the average utility loss is less than T, and the disclosure risk of y_1 against differencing attack is less than R given a specified definition of disclosure α. We illustrate this conjectured wider applicability by comparing the performances of the two algorithms on different cells.

Suppose the statistical agency set (T,R,α) to be $(10\%, 15\%, 11\%)$ for a perturbed cell estimate to be legitimate. Recall that in a recent study, we use one set of parameters for the algorithm in Thompson et al. for perturbing all cells as it is more practical. We stick to this way in this section and the parameters were set to be $K = 3$, $m = (0.4, 0.3, 0.2)$. The parameter of the new algorithm was set to be $\beta = 0.1$. Recall that $\beta = 0.1$ guarantees a 10% utility loss for each cell.

Cell 1: The contributor values of cell 1 consist of $(y_1w_1, y_2w_2, \cdots, y_6w_6) = (30, 30, 30, 10, 5, 5)$. Using the algorithm in Thompson et al., the average utility loss and disclosure risk of releasing perturbed cell estimates are 12.4% and 9.4%, respectively. It means that, even though the perturbed cell estimates satisfy a required level of disclosure risk, they do not carry enough data utility as required by the statistical agency. Using the new algorithm, the average utility loss and disclosure risk of releasing perturbed cell estimates are 10% and 6.5%. It means that, both the requirements of utility loss and disclosure risk are satisfied, and it is legitimate to release a perturbed cell value generated by the new algorithm.

Cell 2: The contributor values of cell 2 consist of $(y_1w_1, y_2w_2, \cdots, y_7w_7) = (25, 25, 25, 25, 1, 1, 1)$. Using the algorithm in Thompson et al., the average utility loss and disclosure risk of releasing perturbed cell estimates are 10.9% and 12.0%, respectively. That means perturbed cell estimates do not carry enough data utility as required by the statistical agency. Using the new algorithm, the average utility loss and disclosure risk of releasing perturbed cell estimates are 10% and 11.4%. Both the requirements of utility loss and disclosure risk are satisfied, and it is legitimate to release a perturbed cell value generated by the new algorithm.

The above two cells are used to show that the new algorithm could help to generate legitimate cell estimates that are not achievable by the algorithm in Thompson et al.. However, we next show that it is possible that, when a cell contains a dominant contributor value, the algorithm in Thompson et al. is better. For illustration, we set (T,R,α) to be

$(15\%, 12\%, 11\%)$, and $\beta = 0.15$ for the next cell.

Cell 3: The contributor values of cell 3 consist of $(y_1w_1, y_2w_2, \cdots y_9w_9) = (60, 20, 20, 15, 15, 10, 10, 10, 10)$. The average utility loss and disclosure risk of releasing perturbed cell estimates generated by the algorithm in Thompson et al. are 14.1% and 9.5%, while the counterparts generated by the new algorithm are 15% and 13.1%. In this case the algorithm in Thompson et al. is the better algorithm perturbing the cell.

The reason the new algorithm is not favourable for **Cell 3** is that, when the ratio $y_1w_1/\hat{s}$ gets large, the disclosure risk of using the new algorithm goes up dramatically. To see this, without loss of generality, we assume n is even, $\beta \geq 2\alpha$, $y_1w_1 < min(\frac{\lambda}{1.5\beta+\alpha}, \frac{\lambda}{2\alpha})$. From Table 3, the disclosure risk is $P_{C11} = \frac{1}{2\lambda\lambda_1}\alpha y_1^2 w_1^2 \beta$. It is evident that the ratio $y_1w_1/\hat{s}$ largely impact the value of disclosure risk. Possible future research would be to use either the algorithm in Thompson et al. or the new algorithm to perturb a cell estimate subject to a condition involving the value of $y_1w_1/\hat{s}$.

5.3 Better Utility-Disclosure Trade-offs

We compare the utility-disclosure tradeoffs of the two algorithms on different cells through simulations. In order to obtain utility-disclosure plots, we gradually changed the values in the magnitude vector m used by the algorithm in Thompson et al. and the parameter β used by the new algorithm. We recorded the average utility losses and disclosure risks given different parameter values. Moreover, we provide utility-disclosure plots for $\alpha = 0.11$ and $\alpha = 0.18$, respectively.

Simulation 1: The contributor values of a cell are $(y_1w_1, y_2w_2, \cdots, y_8w_8) = (25, 25, 25, 25, 25, 25, 25, 25)$. We set the magnitude vector to be $m = (0.3 + 0.01i, 0.2 + 0.01i, 0.1 + 0.01i)$, where $i = 1, 2 \cdots 40$. We recorded the utility loss and disclosure risk of releasing perturbed cell estimates generated by the algorithm in Thompson et al. for each value of i for generating the utility-disclosure plot. When $i = 20$, $m = (0.5, 0.4, 0.3)$, which is the optimal magnitude vector as shown in Table 1. Similarly we set $\beta = 0.04 + i/400, i = 1, 2 \cdots 40$; and we obtained the utility-disclosure plot for the new algorithm. We use a box-plot to represent the utility-disclosure plot of the algorithm in Thompson et al. and a dotted plot to represent utility-disclosure plot of the new algorithm and these symbols also apply to Figures 2 and 3 discussed in Simulations 2 and 3. The utility-disclosure plots for $\alpha = 0.11$ and 0.18 are provided in Fig. 1.

Simulation 2: The contributor values of a cell are $(y_1w_1, y_2w_2, \cdots, y_9w_9) = (40, 20, 20, 15, 15, 10, 10, 10, 10)$. We set the magnitude vector to be $m = (0.15 + 0.01i, 0.1 + 0.01i, 0.05 + 0.01i)$, where $i = 1, 2 \cdots 40$. The parameter of new algorithm is set to be $\beta = 0.03 + i/300, i = 1, 2 \cdots 40$. We follow the same procedure as in

Simulation 1 to obtain the utility-disclosure plots of the two algorithm for $\alpha = 0.11$ and 0.18. The plots are given in Fig. 2.

Simulation 3: The contributor values are $(y_1w_1, y_2w_2, \cdots, y_{49}w_{49}) = (60, 20, 10, 10, \cdots, 10)$. We set the magnitude vector to be $m = (0.05+0.05i, 0.05i, 0.05i)$, where $i = 1, 2 \cdots 40$. The parameter of the new algorithm is set to be $\beta = 0.01 + i/150, i = 1, 2 \cdots 40$. The utility-disclosure plots for $\alpha = 0.11$ and 0.18 are given in Fig. 3.

From Fig. 1, we see that the new algorithm leads to a better utility-disclosure trade-off when the contributor values to a cell estimate are uniformly distributed. From Fig. 2, we see that this advantage is reduced when the largest contributor value dominates the cell estimate. From Fig. 3, we see that the new algorithm again offers a better utility-disclosure trade-off even though the largest contributor value is significantly larger than all other contributor values, as in this case the largest contributor value does not dominate the cell estimate.

6 Conclusion

In this paper we introduced a new algorithm to generate perturbed cell estimates. The advantages of the new algorithm are discussed compared with the algorithm in Thompson et al.. It is conjectured that the new algorithm could be widely used in many remote systems for creating tables from business microdata. Possible future research would be on combining the new algorithm with the algorithm in Thompson et al. to perturb survey estimate of population totals from business microdata.

Acknowledgements We sincerely thank Sybille McKeown and all the others in the Australian Bureau of Statistics for providing so many good feedbacks for the paper.

References

Blakemore, M.: The potential and Perils of Remote Access, in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes(eds.), Amsterdam: Elsevier Science B. V., 315-340 (2001)

Chipperfield J. O. and O'Keefe C. M.: Disclosure-protected Inference Using Generalised Linear Models. International Statistical Review, 82, 3, 371-391. doi:10.1111/insr.12054 (2014)

Decays D. and Nanopoulos P.: Panels of enterprises and confidentiality: the small aggregates method. Proceedings of 92 Symposium on Design and Analysis of Longitudinal Surveys, pp.195-204. Statistics Canada, Ottawa (1993)

Domingo-Ferrer, J. and Torra, V.: Disclosure Protection Methods and Information Loss for Microdata. In Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies (eds. Doyle P., Lane J.I., Theeuwes J.J.M. and Zayatz L.), pp.91-110. North-Holland, Amsterdam (2001)

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

Hawala, S. Zayatz, L. and Rowland, S.: American FactFinder: Disclosure Limitation for the Advanced Query System. Journal of Official Statistics, Vol.20, No.1, pp. 115-124 (2004)

Kim, J.J. and Winkler, W.E.: Masking Microdata Files, American Statistical Association, Proceedings of the Section on Survey Research Methods, 114-119 (1995)

Klein, M., Mathew, T., and Sinha, B.: Noise Multiplication for Statistical Disclosure Control of Extreme Values in Log-normal Regressopm Samples. Journal of Privacy and Confidentiality, 6, 77-125 (2014)

Lin, Y. X. and Wise, P.: Estimation of regression paremeters from noise multiplied data. Journal of Privacy and Confidentiality, 4, 61-94 (2012)

Lucero, J. Singh, L. and Zayatz, L.: Recent Work on the Microdata Analysis System at the Census Bureau. Research Report Series(Statistics #2009-09) (2009)

Moore R.: Controlled data swapping techniques for masking public use microdata sets. U. S. Bureau of the Census, Washington, DC (1996) Available at: http://www.census.gov/srd/papers.pdf.rr96-4.pdf.

Nissim, K., Raskhodnikova, S., and Smith, A.: Smooth sensitivity and sampling in private data analysis, in: D.S.Johnson, U.Feige(Eds.), 39th ACM Symposium on Theory of Computing-STOC 2007, ACM, pp.75-84 (2007)

O'Keefe C. M. and Chipperfield J. O.: A summary of attack methods and confidentiality protection measures for fully automated remote analysis systems. International Statistical Review, 0,0, 1-30 doi: 10.1111/insr.12021 (2013)

Reiter, J. : New approaches to data dissemination: A glimpse into the future (?). Chance, 17, 12–16(2004)

Rubin,D.B.: Discussion: Statistical disclosure limitation. Journal of Official Statistics,9:461-468 (1993)

Salazar-González, Juan-José: A Unified Mathematical Programming Framework for different Statistical Disclosure Limitation Methods. Operations Research, Volume 53, Issue 5 (2005)

Sarathy, R. and Muralidhar, K.: Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data. Transactions on Data Privacy, 4, 1-17 (2011)

Soria-Comas, J. and Domingo-Ferrer, J.: Optimal data-independent noise for differential privacy, Information Science 250, 200-214 (2013)

Thompson, G., Broadfoot, S., and Elazar, D. :Methodology for the Automatic Confdentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics. UNECE Work Session on Statistical Data Confidentiality, p 28-30, Ottawa, October (2013)

Yancey, W.E., Winkler, W.E., and Creecy, R.H.: Disclosure Risk Assessment in Perturbative Micro-data Protection. In: Inference Control in Statistical Databases (ed. J. Domingo-Ferrer), New York: Springer, 135-151 (2002)

Chipperfield, J., Newman, J., Thompson, G., Ma, Y., Lin, Y.X. : Prospects for Protecting Aggregate Business Microdata via a Remote Server. (working paper)

Appendix

Table 1. Magnitude values that guarantee 15% disclosure risk given $\alpha = 0.11$ and minimise the average utility loss for different distributions of top contributor values.

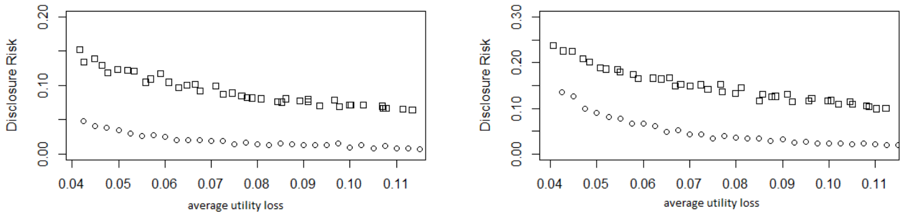
Distribution	Relative Size of Top Contributors				Optimal Magnitude Values		
	1st	2nd	3rd	4th	m1	m2	m3
1	90	5	5		0.15	0.1	0.1
2	80	10	5	5	0.15	0.1	0.1
3	70	20	10		0.15	0.1	0.1
4	60	20	10	10	0.2	0.1	0.1
5	60	40			0.25	0.15	0.1
6	50	20	20	10	0.25	0.15	0.1
7	40	30	30		0.3	0.2	0.1
8	30	30	30	10	0.4	0.3	0.2
9	25	25	25	25	0.5	0.4	0.3

Table 2. Probability expressions for Table 3. $m_1 = y_1 w_1$ and $\lambda_1 = \hat{s}_{-1}\beta$.

P_{C11}	$\frac{1}{2\lambda\lambda_1}\alpha m_1^2\beta$
P_{C12}	$\frac{1}{2\lambda\lambda_1}(0.125\lambda^2 + 0.5\lambda\alpha m_1 + 0.5\alpha^2 m_1^2 - 0.25\lambda\lambda_1 + 0.125\lambda_1^2 - 0.5\alpha m_1\lambda_1)$
P_{C13}	$\frac{1}{2\lambda_1\lambda}(1.125\lambda_1^2 + 1.5\lambda_1\alpha m_1 + 0.5\alpha^2 m_1^2 - 2.25\lambda\lambda_1 + 1.125\lambda^2 - 1.5\alpha m_1\lambda)$
P_{C21}	$\frac{\alpha m_1}{2\lambda}$
P_{C22}	$\frac{1}{2\lambda\lambda_1}(-0.5\alpha^2 m_1^2 - \lambda_1 m_1\alpha + 3\lambda\lambda_1 - 2\lambda_1^2 - 1.125\lambda^2 + 1.5\lambda\alpha m_1)$
P_{C23}	$\frac{3\alpha m_1^2\beta}{2\lambda\lambda_1}$
P_{C24}	$\frac{1}{2\lambda\lambda_1}(0.125\lambda_1^2 + 0.5\alpha^2 m_1^2 + 0.5\alpha m_1\lambda_1 - 0.25\lambda\lambda_1 + 0.125\lambda^2 - 0.5\alpha m_1\lambda)$
P_{C25}	$\frac{1}{2\lambda\lambda_1}(0.75\lambda\lambda_1 + 0.5\alpha m_1\lambda_1 - 0.875\lambda_1^2)$
P_{C26}	$\frac{1}{2\lambda\lambda_1}(1.125\lambda^2 + 0.5\alpha^2 m_1^2 + 1.5\alpha m_1\lambda - 2.25\lambda\lambda_1 + 1.125\lambda_1^2 - 1.5\lambda_1\alpha m_1)$

Table 3. Disclosure risk of perturbed estimates generated by the new algorithm.

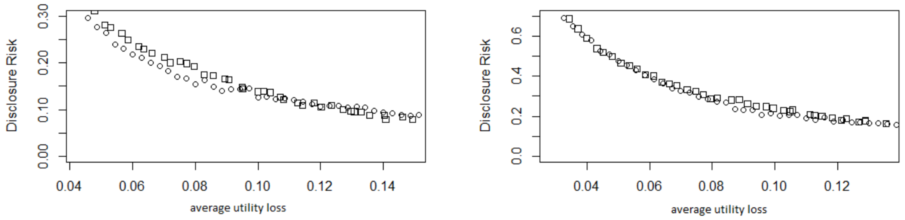
	$\beta \geq 2\alpha$	$\frac{\alpha}{1.5} < \beta < 2\alpha$	$\beta \leq \frac{\alpha}{1.5}$
n is odd, $\frac{\lambda}{4\beta-2\alpha} < y_1 w_1 < min(\frac{\lambda}{1.5\beta+\alpha}, \frac{\lambda}{2\alpha})$	P_{C21}	$P_{C24} + P_{C21}$	not possible
n is odd, $\frac{\lambda}{4\beta+2\alpha} < y_1 w_1 < \frac{\lambda}{4\beta-2\alpha}$	P_{C22}	$P_{C24} + P_{C22}$	not possible
n is odd, $\frac{\lambda}{4\beta+2\alpha} < y_1 w_1 < min(\frac{\lambda}{1.5\beta+\alpha}, \frac{\lambda}{2\alpha})$	P_{C22}	$P_{C24} + P_{C22}$	$P_{C24} + P_{C25}$
n is odd, $y_1 w_1 < \frac{\lambda}{4\beta+2\alpha}$	P_{C23}	$P_{C24} + P_{C23}$	$P_{C24} + P_{C26}$
n is even, $y_1 w_1 < min(\frac{\lambda}{1.5\beta+\alpha}, \frac{\lambda}{2\alpha})$	P_{C11}	P_{C12}	$P_{C12} + P_{C13}$



(a) Utility-Disclosure plots for Simulation 1 with $\alpha = 0.11$

(b) Utility-Disclosure plots for Simulation 1 with $\alpha = 0.18$

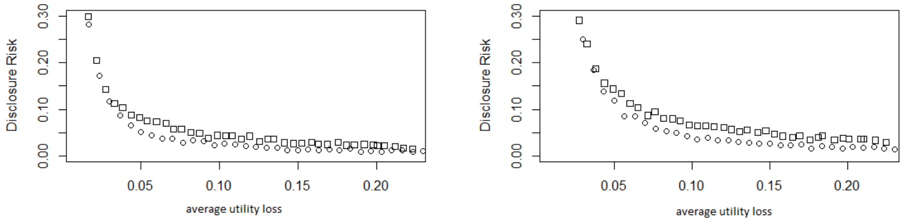
Fig. 1. Utility-disclosure plots for Simulation 1 with different α values. The box-plot represents results generated by the Thompson et al. algorithm and the dotted plot represents results generated by the new algorithm.



(a) Utility-Disclosure plots for Simulation 2 with $\alpha = 0.11$

(b) Utility-Disclosure plots for Simulation 2 with $\alpha = 0.18$

Fig. 2. Utility-disclosure plots for Simulation 2 with different α values.



(a) Utility-Disclosure plots for Simulation 3 with $\alpha = 0.11$

(b) Utility-Disclosure plots for Simulation 3 with $\alpha = 0.18$

Fig. 3. Utility-disclosure plots for Simulation 3 with different α values.