NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong

Working Paper

07-17

The Vulnerability of Multiplicative Noise Protection to Correlation-Attacks on Continuous Microdata

Yue Ma, Yan-Xia Lin and Rathindra Sarathy

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: anica@uow.edu.au

The Vulnerability of Multiplicative Noise Protection to Correlation-Attacks on Continuous Microdata

Yue Ma[1], Yan-Xia Lin[1] and Rathindra Sarathy[2]

[1]National Institute for Applied Statistics Research Australia
School of Mathematics and Applied Statistics
University of Wollongong, Australia
[2]Spears School of Business, Oklahoma State University
Stillwater OK 74078, USA

Abstract

In traditional noise multiplication masking method, it is frequently assumed that data intruders use the perturbed value (noise multiplied value) to estimate the unperturbed value (original value) of an observation. In this paper, we show that, to estimate the value of an original observation from its noise multiplied counterpart, data intruders could use another estimate instead of using the perturbed value. The new estimate, namely correlation-attack estimate, is obtained by exploiting the potentially high correlation between the perturbed data and the unperturbed data. We provide detailed comparison between the two estimates (perturbed value and the correlation-attack estimate) by comparing the mean square errors of the two underlying estimators, and we propose that data providers should assess the disclosure risk of the correlation-attack estimator for a considerable amount of original observations. We also propose a disclosure risk measure against the correlation-attack estimator, and we demonstrate how could the disclosure risk measure be used to help data providers with noise generating variable selection during the masking stage.

Keywords: Data Confidentiality, Noise Multiplication Masking, Continuous Microdata, Disclosure Risk, Attacking Strategy

1 Introduction

The role of data providers is to collect large amounts of data from different sources, and make the data available to users for analysis. On the one hand, data users need authentic data to obtain important statistics such as sample means or regression coefficients to learn information about the population; On the other hand in order to protect privacy of survey respondents, sensitive information, such as a person's income, or health condition, should not be revealed to the public. In order to satisfy both requirements, data providers apply data masking methods to the original data and release masked data to the public. Different data masking methods include generating synthetic data (Rubin, 1993), top and bottom coding (Hundepool et al. 2008), data shuffling (Muralidhar and Sarathy, 2006), noise addition (Kim, 1986) and noise multiplication (Hwang, 1986; Muralidhar et al., 1995; Evans, 1996).

Masked data should maintain a high analytical validity and a low disclosure risk, meaning that data users could still infer important aggregated statistical information from the masked data while data intruders could not learn private information of individuals from the masked data. In the context of masking continuous microdata, it is commonly assumed in the literature and implemented in practice that the analytical validity of first two moments estimations of the original data is preserved in the masked data, so that estimates of sample mean, covariances, and correlation coefficients could be obtained by analysing the masked data (Kim, 1990; Kim and Winkler, 2003; Brand, 2002; Yancey et al., 2002; Domingo-Ferrer et al.,2004; Oganian and Karr, 2011).

In this paper we consider using noise multiplication masking method to produce masked data. Noise multiplication masking method has been used in practice by the U.S. Energy Information Administration and the U.S. Bureau of Census (Kim and Jeong, 2008). The noise multiplication masking method works in the following way:

Suppose the j-th attribute of a microdata is sensitive. Then each individual observation in the attribute is masked by multiplicative noise. To do this, the data provider selects a noise generating variable C_j, and for each observation in the attribute, a random noise is drawn from C_j and multiplied with the observation to produce the noise multiplied observation. Mathematically, Denote the original j-th attribute values as $y_j = (y_{1j}, y_{2j}, \cdots, y_{nj})$. The data provider only releases the noise multiplied values $y^*_j = (y^*_{1j}, y^*_{2j}, \cdots, y^*_{nj})$ to the public, where $y^*_{ij} = y_{ij}c_{ij}$. The noise terms $\{c_{ij}\}_{i=1}^{n}$ are independently generated from C_j. When multiple attributes are masked by multiplicative noises, for instance the k-th attribute is also sensitive, then data providers use a noise generating variable C_k to mask the k-th attribute, and C_j and C_k are independent.

It is also commonly assumed in the literature to use noise multiplication masking method to protect the original data while to maintain analytical validity of the first two moments of the original data, provided that extra information about the noise generating variables used during masking stage is also released together with the noise multiplied data. Kim and Winkler (2003) showed that, if truncated normal noise generating variable is used, then it is sufficient for data providers to also release the mean and variance together with the truncation points to the public so that data users could derive estimates of mean and variance of the original data. More generally, Nayak et al. (2011) showed that, data users could obtain unbiased the first two moments estimates of the original data set if the variances of the noise generating variables used to mask different attributes during the masking stage are also released. Moreover, Lin and Wise (2012) showed that, releasing the variances of the noise generating variables together with the noise multiplied data set allows data users to obtain regression parameters estimates in subpopulations. Recent research shows that data providers could release the density function of C to the public to allow data users to carry out more sophisticated data analysis on noise multiplied data (Nayak et al.,2011; Sinha et al.,2011; Klein et al., 2014; Lin, 2014; Lin and Fielding, 2015). In this paper, **we only assume the more common and simpler case that data providers release the noise multiplied microdata together with the variances of noise generating variables to the public, so that data users could obtain the first two moments estimates of the original microdata**. Moreover, we assume all data entries and noise terms are positive and continuous. It is a reasonable assumption as many sensitive attributes, such as personal income, are non-negative and continuous, and the positivity nature is violated if negative noise is used. We do not consider the case where a data value is 0 as 0 values cannot be protected by multiplicative noise.

The mean of the noise generating variable C used to mask values of an attribute, denoted as μ_C, is not always assumed to be 1 in the literature, as in Kim and Winkler (2003). However, if μ_C is not 1 in some cases, data providers will also need to release μ_C together with σ_C^2 to the public so that valid inference of the original sample mean and sample variance could be obtained by data users. In that case, it is easy for data users to re-scale the noise multiplied data by dividing each noise multiplied observation by μ_C and make the corresponding adjustment to σ_C^2, so that it could be seen as the original data being masked by a noise generating variable with mean 1 and variance σ_C^2/μ_C^2, and the resulted noise multiplied data is the re-scaled noise multiplied data. In this paper we always **assume** $\mu_C = 1$. The assumption that $\mu_C = 1$ simplifies our discussion in the paper and could be generalised to other cases where noise mean is not 1.

Value disclosure risk is an important topic. Value disclosure occurs when the original value is reasonably inferred by data intruders from the perturbed data. Despite that data are masked before releasing, malicious data intruders may attempt to learn private information of individuals by deriving attacking strategies. For noise multiplied data, it is shown in the literature that the noise multiplied value is in fact an unbiased estimate of the original value. Based on this fact, many authors introduced different disclosure risk measures by assuming that data intruders attempt to estimate an original value using the noise multiplied value only. (Kim and Winkler, 2003; Evans, 1998; Lin and Wise, 2012).

For noise multiplication masking method, the selection of noise generating variable plays an important role in balancing data utility loss and disclosure risk. For data providers, one important job is to select an appropriate noise generating variable which provides adequate protection to the data without losing too much information. It turns out that, different noise generating variables perform differently in terms of data utility preservation and data protection. To understand the performance of different noise candidates, appropriate disclosure risk measure and utility loss measure should be defined, so that data providers could make appropriate decision about noise candidate selection. In order to appropriately define the disclosure risk measure, data providers need to be aware of potential attacking strategies that data intruders may use to reveal a protected value. Despite that the noise multiplied value has been recognised as an estimate of the original value, we found that this assumption is not sufficient.

In this paper, we show that, if the data provider releases the noise multiplied data together with variance of noise generating variable to the public, data intruders could infer another estimate, namely "correlation-attack estimate", from the released information. We will show that, the correlation-attack estimator could efficiently breach data confidentiality, leading to serious value disclosure risk. Consequently, we propose a new disclosure risk measure to help decision-makings of data providers. The application of using the new disclosure risk measure for noise generating variable selection is demonstrated in the paper.

The paper is organized as following: In Section 2 we introduce the noise multiplication masking method and first two moments estimations preservation. In Section 3 we show the correlation-attack strategy. In Section 4 we discuss the correlation-attack estimator and compare it with the noise multiplied value estimator. In Section 5 we introduce the disclosure risk and data utility loss measures. In Section 6 we discuss data intruders' motivation of using correlation-attack strategy to disclose original values. In Section 7 we present two simulation studies. In Section 8 we conclude the paper.

2 Noise multiplication to preserve the first two moments

In this section we describe the noise multiplication masking method and show how do data users obtain statistical inferences regarding original sample means, covariances and correlation coefficients by analysing the noise multiplied data.

Suppose in a positive continuous microdata, the values of the j-th attribute are $y_j = \{y_{ij}\}_{i=1}^{n}$, where $\{y_{ij}\}_{i=1}^{n}$ are independently drawn from an underlying random variable Y_j. The data provider selects a noise generating variable C_j to mask y_j, where $E(C_j) = 1$ and C_j and Y_j are independent. Then random noises $\{c_{ij}\}_{i=1}^{n}$ are independently drawn from C_j and are multiplied to $\{y_{ij}\}_{i=1}^{n}$ to produce a noise multiplied data $y^*_j = \{y^*_{ij}\}_{i=1}^{n}$, where $y^*_{ij} = y_{ij}c_{ij}$. Denote the underlying random variable of $\{y^*_{ij}\}_{i=1}^{n}$ to be Y^*_j, where $Y^*_j = Y_jC_j$. Similarly, suppose the k-th attribute is also masked by multiplicative noises. Denote the underlying random variable of the k-th attribute values as Y_k and noise generating variable used for the k-th attribute is C_k with $E(C_k) = 1$. We also denote the underlying random variable of y^*_{ik} to be Y^*_k, where $Y^*_k = Y_kC_k$. C_j and C_k are independent. The data provider releases the noise multiplied data y^*_j and y^*_k, together with the noise variances $\sigma^2_{C_j}$ and $\sigma^2_{C_k}$ to the public.

Data providers could use the same noise generating variable to mask all sensitive attributes (Kim and Winkler, 2003), or they could use different noise generating variables to mask different attributes (Nayak et al., 2011; Lin and Wise, 2012). In either case, the first two moments estimates of the original microdata could be unbiasedly recovered from the noise multiplied data. It follows from Nayak et al. (2011) that $E(Y_j) = E(Y^*_j)$ and $E(Y^2_j) = E(Y^{*2}_j)/(\sigma^2_{C_j} + 1)$. The first two sample moments estimates of noise multiplied data could easily be obtained by data users. Therefore, data users could obtain estimates of $E(Y_j)$ and $E(Y^2_j)$ by $\overline{y^*_j} = \frac{\sum_{i=1}^{n} y^*_{ij}}{n}$ and $\overline{y^{*2}_j} = \frac{\sum_{i=1}^{n} y^{*2}_{ij}}{n}$ respectively, where n is the total number of record in the microdata.

Estimates of the variance of Y_j and the covariance of Y_j and Y_k could also be obtained easily. Nayak et al. (2011) showed that the variance of Y_j, $\sigma^2_{Y_j}$, is

$$\sigma^2_{Y_j} = \frac{\sigma^2_{Y^*_j} - \mu^2_{Y^*_j}\sigma^2_{C_j}}{1 + \sigma^2_{C_j}} \tag{1}$$

(recall $\mu_{C_j} = 1$) and therefore data users could use $\hat{\sigma}^2_{Y_j}$ to estimate $\sigma^2_{Y_j}$, where

$$\hat{\sigma}^2_{Y_j} = (s^2_{y^*_j} - \overline{y^*_j}\sigma^2_{C_j})/(\sigma^2_{C_j} + 1), \tag{2}$$

where $s^2_{y^*_j}$ is the sample variance of y^*_j. The covariance between Y_j and Y_k, $cov(Y_j, Y_k)$, is the same as $cov(Y^*_j, Y^*_k)$, therefore data users could use $s_{y^*_j y^*_k}$ to estimate $cov(Y_j, Y_k)$, where $s_{y^*_j y^*_k}$ is the sample covariance between y^*_j and y^*_k.

The level of protection offered by multiplicative noise is frequently measured by data intruders' uncertainty of using the perturbed value to estimate the original value (Nayak et al.,2011; Kim, 2007; Kim and Jeong, 2008; Lin and Wise, 2012). The reason is that, as showed in Nayak et al.(2011), for each original value y_{ij}, since $E(Y^*_{ij}|Y_{ij} = y_{ij}) = y_{ij}$, therefore Y^*_{ij} is an unbiased estimator of

y_{ij} given that $Y_{ij} = y_{ij}$. This is a common assumption used in the literature. However, since data providers also release the variances of noise generating variables used to mask sensitive attributes during the masking process, this assumption may not adequately reflect all potential risks of value disclosure. The reason is that, data intruders could use the knowledge of the noise variance to obtain the correlation-attack estimates, which we will introduce in the next section.

3 The Correlation-attack strategy

In this section, we introduce the correlation-attack strategy which could be used by data intruders to obtain estimates of original values. The correlation-attack only requires the noise multiplied values of an attribute, and the variance of the noise used to mask the attribute. Therefore, the correlation-attack is based on univariate data only.

Suppose the original data of the j-th attribute j is $y_j = (y_{1j}, y_{2j}, \cdots, y_{nj})$, the corresponding noise multiplied data of is $y_j^* = (y_{1j}^*, y_{2j}^*, \cdots, y_{nj}^*)$. The noise generating variable used for masking y_j is C_j, with mean 1 and variance σ_{C_j}. Data providers release y_j^* and $\sigma_{C_j}^2$ to the public.

The correlation-attack follows from the idea that, if the sample correlation $r_{y_j y_j^*}$ is high, which indicates a strong linear relationship between y_j and y_j^*, then a linear regression model may adequately explains the relationship between the two sets of data, even though theoretically the relationship is not linear.

Now we model data intruders' behaviour of conducting the correlation-attack. Recall that since the original data y_j is not public, it is impossible for data intruders to obtain sample correlation coefficient $r_{y_j y_j^*}$. However, it could be reasonably estimated if sample size n is large. When n is large, $r_{y_j y_j^*}$ should be very close to $\rho_{Y_j Y_j^*}$, the population correlation coefficient between the two underlying random variables Y_j and Y_j^*, where $Y_j^* = Y_j C_j$. The theoretical correlation coefficient is given as

$$\rho_{Y_j Y_j^*} = \frac{Cov(Y_j^*, Y_j)}{\sqrt{Var(Y_j^*)Var(Y_j)}} = \sqrt{\frac{\sigma_{Y_j}^2}{\sigma_{Y_j}^2(\sigma_{C_j}^2 + 1) + \mu_{Y_j}^2 \sigma_{C_j}^2}} \tag{3}$$

where $\mu_{Y_j} = E(Y_j)$, $\sigma_{Y_j}^2 = Var(Y_j)$ and $\sigma_{C_j}^2 = Var(C_j)$.

For data intruders, although $\rho_{Y_j Y_j^*}$ is also not known, data intruders could use existing information to estimate this value. The reason is, data intruders could obtain estimates of μ_{Y_j} and $\sigma_{Y_j}^2$ by analysing the noise multiplied data, and the value of $\sigma_{C_j}^2$ is available as this information is directly released by the data provider. As introduced in Section 2, data intruders could use $\overline{y_j^*} = \frac{\sum_{i=1}^n y_{ij}^*}{n}$

to estimate μ_{Y_j}, and use Equation (2) to obtain $\hat{\sigma}^2_{Y_j}$ to estimate $\sigma^2_{Y_j}$. Consequently, data intruders could obtain $\tilde{r}_{y_j y^*_j}$ to estimate $\rho_{Y_j Y^*_j}$, which is

$$\tilde{r}_{y_j y^*_j} = \sqrt{\frac{\hat{\sigma}^2_{Y_j}}{\hat{\sigma}^2_{Y_j}(\sigma^2_{C_j}+1) + \overline{y^*_j}^2 \sigma^2_{C_j}}} \tag{4}$$

When n is large, the estimates $\overline{y^*_j}$ and $\hat{\sigma}^2_{Y_j}$ should be very close to the theoretical values, therefore $\tilde{r}_{y_j y^*_j}$ should be very close to $\rho_{Y_j Y^*_j}$. Recall that the sample correlation $r_{y_j y^*_j}$ should also be very close to $\rho_{Y_j Y^*_j}$ for large n. Therefore, data intruders could use $\tilde{r}_{y_j y^*_j}$ to estimate $r_{y_j y^*_j}$.

Theoretically, if both the original data y_j and noise multiplied data y^*_j are available and a linear regression model is built, the linear regression model should have the following form:

$$\hat{y}_{ij} = \hat{\alpha} + \hat{\beta} y^*_{ij},$$

where $\hat{y}_{ij}$ is the predicted value of y_{ij} given y^*_{ij}, $\hat{\beta} = r^2_{y_j y^*_j}$, $\hat{\alpha} = \overline{y_j} - \overline{y^*_j}\hat{\beta}$. Since in practice y_j is not known to data intruders, the exact values of $\hat{\beta}$ and $\hat{\alpha}$ are not known. However, for large n, as $r_{y_j y^*_j}$ could be reasonably estimated by $\tilde{r}_{y_j y^*_j}$, and $\overline{y_j}$ could be reasonably estimated by $\overline{y^*_j}$, data intruders could obtain estimates of $\hat{\beta}$ and $\hat{\alpha}$ by replacing the unknown terms with the corresponding estimates. Denote the estimates as $\tilde{\beta}$ and $\tilde{\alpha}$, then data intruders could obtain an estimated value of $\hat{y}_{ij}$, which is given as

$$\tilde{y}_{ij} = \tilde{\alpha} + \tilde{\beta} y^*_{ij} = (1 - \tilde{r}^2_{y_j y^*_j})\overline{y^*_j} + \tilde{r}^2_{y_j y^*_j} y^*_{ij}.$$

Therefore, data intruders may use $\tilde{y}_{ij}$ to estimate the original value y_{ij}. This estimate should be more accurate if the sample correlation between y_j and y^*_j is larger. Recall that, in the literature, it has been shown that the perturbed observation y^*_{ij} could also be used to estimate the original value y_{ij}. Therefore, data providers should compare these two estimates and find out under which condition is one estimate more accurate than the other. Data providers need to beware of the behaviour of both estimates and take corresponding precautions during masking stage. In next section, we compare these two estimates by comparing the two underlying estimators and discuss the accuracy of the correlation-attack estimator.

4 Accuracy of $\tilde{Y}_{ij}$ and comparison between Y^*_{ij} and $\tilde{Y}_{ij}$

In this section we base our discussion from the data providers point of view. For the data provider, the positive original data $y_j = (y_{1j}, y_{2j}, \cdots, y_{nj})$ is given, and the data provider needs to select a positive noise generating variable C_j to mask y_j to produce the noise multiplied data $y^*_j = (y^*_{1j}, y^*_{2j}, \cdots, y^*_{nj}) = (y_{1j}c_{1j}, y_{2j}c_{2j}, \cdots, y_{nj}c_{nj})$. The original sample mean of y_j is denoted as

$\overline{y_j}$ and original sample variance is denoted as $s^2_{y_j}$.

It is frequently assumed that, data intruders use the noise multiplied observation y^*_{ij} to estimate the original value of y_{ij} (Muralidhar et al., 1995; Nayak et al.,2011; Lin and Wise., 2012). The reason is that, for a given unmasked observation y_{ij}, the perturbed random variable $Y^*_{ij} = y_{ij}C_{ij}$ is an unbiased estimator of y_{ij}. The disclosure risk assessments currently proposed in the literature are measured against the unbiased estimator.

From last section, we introduced the correlation-attack estimate $\tilde{y}_{ij}$ to estimate y_{ij}, and its underlying random variable $\tilde{Y}_{ij}$ is another estimator of y_{ij}. It is important for data providers to understand the behaviour of both estimators and design disclosure risk measures accordingly. In this section, we will firstly discuss the accuracy of the estimator $\tilde{Y}_{ij}$. Secondly, we compare the performances of the two estimators Y^*_{ij} and $\tilde{Y}_{ij}$ and then draw some general conclusions.

The mathematically expression of $\tilde{Y}_{ij}$ cannot be easily find without making assumptions. By noting that for large-sized data, $\tilde{r}^2_{y_jy_j^*} \approx \rho_{Y_jY_j^*}$ and $\overline{y_j^*} \approx \overline{y_j}$, the expression of $\tilde{Y}_{ij}$ is reduced to

$$\tilde{Y}_{ij} = (1-\rho^2_{Y_jY_j^*})\overline{y_j} + \rho^2_{Y_jY_j^*} y_{ij}C_{ij}.$$

The expression of $\tilde{Y}_{ij}$ only make sense for large-sized samples where valid inferences of $\rho_{Y_jY_j^*}$ and $\overline{y_j}$ could be drawn. The simplified expression of $\tilde{Y}_{ij}$ greatly simplifies our discussion in this section. For data providers, it is important to know the behaviour of the two estimators Y^*_{ij} and $\tilde{Y}_{ij}$ and to understand how to reduce the disclosure risks against the two estimators.

First of all, we calculate the Mean Square Errors (MSE) of the two estimators. For the data provider, the original data y_j is available. Therefore, the MSE of the unbiased estimator Y^*_{ij} to estimate y_{ij} is given as:

$$MSE(Y^*_{ij}|y_{ij}) = E(C_jy_{ij} - y_{ij})^2 = y^2_{ij}E(C_j-1)^2 = y^2_{ij}\sigma^2_{C_j}.$$

The MSE of the correlation-attack estimator $\tilde{Y}_i$ is given as:

$$MSE(\tilde{Y}_{ij}|y_{ij}) = E[(1-\rho^2_{Y_jY_j^*})\overline{y_j} + \rho^2_{Y_jY_j^*}C_jy_{ij} - y_{ij}]^2 = (1-\rho^2_{Y_jY_j^*})^2(\overline{y_j} - y_{ij})^2 + \rho^4_{Y_jY_j^*}y^2_{ij}\sigma^2_{C_j}.$$

The MSEs of both estimators show that they become less accurate in general as the variance of noise $\sigma^2_{C_j}$ increase. It is straightforward to see it in Y^*_{ij}, but it is not straightforward to see it in $\tilde{Y}_{ij}$. To show this, we note that from Equation (3), $\sigma^2_{C_j} = \frac{1-\rho^2_{Y_jY_j^*}}{\rho^2_{Y_jY_j^*}(1+\mu^2_{Y_j}/\sigma^2_{Y_j})}$. Note that data providers may never know the value of μ_{Y_j} and $\sigma^2_{Y_j}$ but data providers always know $\overline{y_j}$ and $s^2_{y_j}$. Therefore, by replacing μ_{Y_j} and $\sigma^2_{Y_j}$ by $\overline{y_j}$ and $s^2_{y_j}$, we have

$$MSE(\tilde{Y}_{ij}|y_{ij}) = (k_1 - k_2)\rho^4_{Y_jY_j^*} + (k_2 - 2k_1)\rho^2_{Y_jY_j^*} + k_1,$$

where $k_1 = (\overline{y_j} - y_{ij})^2$, $k_2 = \frac{y^2_{ij}}{1+h}$, $h = \overline{y_j}^2/s^2_{y_j}$. So the MSE is a parabola in terms of $\rho^2_{Y_jY_j^*}$. The parabola is monotone in $[0,1]$ with $MSE(\tilde{Y}_{ij}|y_{ij}) = k_1$ when $\rho^2_{Y_jY_j^*} = 0$, and $MSE(\tilde{Y}_{ij}|y_{ij}) = 0$ when $\rho^2_{Y_jY_j^*} = 1$ in most cases, with the exception when $k_2 > 2k_1$ (under this condition the symmetric axis of $MSE(\tilde{Y}_{ij}|y_{ij})$ is within [0,1]). In that case it means when $y_{ij} \in (\frac{2\overline{y_j}(1+h)-\overline{y_j}\sqrt{2(1+h)}}{1+2h}$, $\frac{2\overline{y_j}(1+h)+\overline{y_j}\sqrt{2(1+h)}}{1+2h})$, the MSE of $\tilde{Y}_{ij}$ will increase slightly from k_1 as $\rho^2_{Y_jY_j^*}$ increases from 0 but will eventually decreases to 0 as $\rho^2_{Y_jY_j^*}$ goes to 1. Therefore, generally speaking, by controlling $\rho^2_{Y_jY_j^*}$ to be small through choosing a large variance of noise $\sigma^2_{C_j}$, data providers could reduce the disclosure risk of the correlation-attack estimator for most of original values.

An estimator predicts the unknown value better if it yields a smaller MSE. For data intruders, the original value y_{ij} is unknown. To predict the value of y_{ij}, logical data intruders should use the estimator which has a smaller MSE for prediction. To compare which estimator predicts y_{ij} better, we set $MSE(\tilde{Y}_{ij}|y_{ij}) - MSE(Y^*_{ij}|y_{ij}) < 0$, and we wish to find out which values of y_{ij} are more vulnerable to the correlation-attack. After solving the inequality, we let

$$\begin{aligned}
a &= \frac{s^2_{y_j} - s_{y_j}\sqrt{s^2_{y_j} + \overline{y_j}^2}}{\overline{y_j}^2}, \\
b &= \frac{s^2_{y_j} + s_{y_j}\sqrt{s^2_{y_j} + \overline{y_j}^2}}{\overline{y_j}^2}, \\
c &= \frac{\overline{y_j}\rho^2_{Y_jY_j^*}(s^2_{y_j} + \overline{y_j}^2) - \overline{y_j}s_{y_j}\rho_{Y_jY_j^*}\sqrt{(1+\rho^2_{Y_jY_j^*})(s^2_{y_j} + \overline{y_j}^2)}}{\rho^2_{Y_jY_j^*}\overline{y_j}^2 - s^2_{y_j}} \\
d &= \frac{\overline{y_j}\rho^2_{Y_jY_j^*}(s^2_{y_j} + \overline{y_j}^2) + \overline{y_j}s_{y_j}\rho_{Y_jY_j^*}\sqrt{(1+\rho^2_{Y_jY_j^*})(s^2_{y_j} + \overline{y_j}^2)}}{\rho^2_{Y_jY_j^*}\overline{y_j}^2 - s^2_{y_j}},
\end{aligned}$$

$e = min(c,d)$ and $f = max(c,d)$, then we have the following lemma which works for all large-sized positive continuous original data:

Lemma 1: *For any large-sized positive continuous original data $y_j = (y_{1j}, y_{2j}, \cdots, y_{nj})$ and noise generating variable C_j, when $\rho_{Y_jY_j^*} < a$ or $\rho_{Y_jY_j^*} > b$, y_{ij} is more vulnerable to $\tilde{Y}_{ij}$ when $y_{ij} \in (e,f)$; when $\rho_{Y_jY_j^*} \in (a,b)$, y_{ij} is more vulnerable to $\tilde{Y}_{ij}$ when $y_{ij} < e$ or $y_{ij} > f$; when $\rho_{Y_jY_j^*}$*

equals a or b, y_{ij} is more vulnerable to $\tilde{Y}_{ij}$ when $y_{ij} > \overline{y_j}/2$.

The above result means that, the correlation-attack estimator $\tilde{Y}_{ij}$ could be more accurate for certain original values, and therefore may yield a higher disclosure risk than the unbiased estimator Y^*_{ij} for those values. Therefore, for those original values, instead of measuring disclosure risk comes from the unbiased estimator, data providers should also measure disclosure risk against the correlation-attack estimator. When data providers calculate the values of a,b,c and d, $\rho^2_{Y_j Y^*_j}$ is not known as the Y_j is not known. But it could be reasonably estimated by replacing μ_{Y_j} by $\bar{y_j}$ and $\sigma^2_{Y_j}$ by $s^2_{y_j}$ into Equation (3).

Therefore, when selecting an appropriate noise generating variable during masking stage, it is important for data providers to evaluate potential value disclosure risk against the correlation-attack estimator. For noise multiplication masking method, the selection of noise generating variables plays an important role in balancing utility loss-disclosure risk tradeoffs. In the next section, we introduce the utility loss measure and disclosure risk measure under the context of this paper.

5 Data utility and disclosure risk assessment

For noise multiplication masking method, an important job for the data provider to do is to decide an appropriate noise generating variable to be used to generate noise multiplied data. To do this, the data provider needs to assess the resulted utility loss-disclosure risk tradeoffs of each potential noise candidate if it were used to mask the original data. Formal definitions of data utility loss and disclosure risk are needed in order for data providers to assess the quality of each noise candidate. In this section we introduce the definition of **data utility loss** and **disclosure risk** used in this paper.

Recall that we only assume that data providers seek to preserve the analytical validity of the first two moments estimates in the noise multiplied data. As the correlation-attack is only conducted on univariate data, we define data utility losses in terms of the first two moments estimates of one-dimensional original data.

Different utility loss measures were proposed in Domingo-Ferrer and Torra (2001) and Duncan et al., (2001; 2004). The measures proposed in Domingo-Ferrer and Torra are calculated based on the distances between perturbed estimates and unperturbed estimates, and the utility loss measures proposed in Duncan et al. are calculated based on data user's mean squared error in estimating the population parameters. In this paper, we adopt the utility measure used in Duncan et al., (2001; 2004), as they can be evaluated mathematically without simulations. That is, we define data utility loss to be data user's mean squared error in estimating the population parameters.

Suppose the original observations of the j-th attribute is $y_j = (y_{1j}, y_{2j}, \cdots, y_{nj})$, which are positive. A positive noise generating variable C_j with mean 1 and variance $\sigma^2_{C_j}$ is used to mask y_j, and the noise multiplied variable is given as $Y^*_j = (y_{1j}C_{1j}, y_{2j}C_{2j}, \cdots, y_{nj}C_{nj})$, C_{ij} are i.i.d with the same probability distribution of C_j for $i = 1, 2, \cdots, n$. Suppose that the original data $\{y_{ij}\}^n_{i=1}$ are independently drawn from the original random variable Y_j, then to infer the first two moments of Y_j, data users use the estimations based on noise multiplied variable. Specifically, data users use $\mu_{Y^*_j|y_j} = \frac{\sum^n_{i=1} y_{ij}C_{ij}}{n}$ to unbiasedly estimate the first population moment of Y_j, and use $\mu_{Y^{*2}_j|y_j} = \frac{\sum^n_{i=1} y^2_{ij}C^2_{ij}}{n(\sigma^2_{C_j}+1)}$ to unbiasedly estimate the second population moment of Y_j (Nayak et al., 2011). Therefore, the utility losses of first two moments estimations due to multiplicative noises are:

$$UL_1 = Var(\mu_{Y^*_j|y_j}) = \frac{\sigma^2_{C_j} \sum^n_{i=1} y^2_{ij}}{n^2}$$

and

$$UL_2 = Var(\mu_{Y^{*2}_j|y_j}) = \frac{[E(C^4_j) - (\sigma^2_{C_j} + 1)^2] \sum^n_{i=1} y^4_{ij}}{n^2(\sigma^2_{C_j} + 1)^2}$$

The definitions showed that, the noise generating variable C_j yields a higher data utility loss UL_1 if $\sigma^2_{C_j}$ is larger and a higher data utility loss UL_2 if $E(C^4) - (\sigma^2_{C_j} + 1)^2$ is larger. When selecting the best noise candidate in terms of utility preservation among a set of noise candidates with the same variances (so that UL_1 are the same), the one with the lowest fourth moment yields the lowest UL_2 and therefore the lowest overall utility loss.

In this paper, disclosure risk is defined in terms of "value disclosure" (Li and Sarkar, 2013), or predictive disclosure (Nayak et al., 2011). Mathematically, we define "disclosure" in the following way, which is also used in Lin and Wise (2012) and Klein et al. (2014):

Suppose the original observation y_{ij} is masked by a multiplicative noise. Based on available information, a data intruder has his own strategy to estimate the value of y_{ij}, and denote the estimate of y_{ij} as $\tilde{y}_{ij}$. Since the original data is continuous, it is almost impossible for the data intruder to infer the exact value of y_{ij} by $\tilde{y}_{ij}$. However, to classify the $\tilde{y}_{ij}$ as a good estimate of y_{ij}, it is sufficient for $\tilde{y}_{ij}$ to be reasonably close to y_{ij}. **Disclosure** of y_{ij} occurs if the data intruder has decided to use $\tilde{y}_{ij}$ to estimate y_{ij}, and $|\frac{\tilde{y}_{ij} - y_{ij}}{y_{ij}}| < \delta$, where δ is the **acceptance rule** as defined in Lin and Wise (2012) and is a small number. For instance, for a positive observation y_{ij}, if we set $\delta = 0.05$, we say that $\tilde{y}_{ij}$ discloses the value of y_{ij} if $0.95y_{ij} < \tilde{y}_{ij} < 1.05y_{ij}$.

The definition of disclosure is frequently used in the literature under other data perturbation context. For instance, it is used for measuring disclosure risk of compromising individual business values in a Remote System (Chipperfield et al., 2017; Ma et al., 2016).

Various disclosure risk measures have been introduced. For instance, Nayak et al. (2011) introduced a confidence interval measure to quantify disclosure risks. Specifically, let $Z_{ij} = y_{ij}C_{ij}$ with $y_{ij} > 0$. Then, for a given α, let $[a_1, b_1]$ be the shortest interval satisfying $P(a_1 < C_{ij} < b_1) = 1-\alpha$. So the $100(1-\alpha)$ confidence interval of y_{ij} based on y^*_{ij} is $(y^*_{ij}/b_1, y^*_{ij}/a_1)$, which is used to measure the disclosure risk of y_{ij}. A similar measure is used in Agrawal and Srikant (2000) under the context of noise addition masking method.

Another measure proposed in Lin and Wise (2012) takes into account the entire distribution of the noise generating variable and measures value disclosure risk by the probability that y_{ij} being disclosed by Y^*_{ij}, which is given by

$$R_{LW}(y_{ij}, \delta|Y^*_j) = P(|\frac{Y^*_{ij} - y_{ij}}{y_{ij}}| < \delta) = P(|C_j - 1| < \delta) \tag{5}$$

The disclosure risk measure evaluates the probability of an observation being disclosed by its noise multiplied counterpart, which is determined by the noise generating variable C_j. In the literature, many noise candidates have been proposed and used in practice to reduce the disclosure risk of noise multiplied data, such as truncated triangular distributed noise candidates (Kim, 2007; Kim and Jeong, 2008). The disclosure risk measure R_{LW} provides a mathematical reason of why the proposed noise candidates are better to use, as R_{LW} is very low if these noise candidates are used during masking stage. It is not so straight-forward to see if the confidence interval measure introduced above is used.

The disclosure risk measure R_{LW} assumes that the unbiased estimator Y^*_{ij} is used by data intruders to estimate the original value y_{ij}. However, this disclosure risk measure could easily be modified to measure disclosure risks from other estimators. For instance, it is used in Klein et al. (2014) to measure disclosure risk, with the unbiased estimator Y^*_{ij} being replaced by a generalised linear regression estimator which could be computed under the context of their paper. In this paper, as we introduced the correlation-attack estimator $\tilde{Y}_{ij}$ to estimate y_{ij}, we therefore propose the following disclosure risk measure $R_\rho(y_{ij}, \delta|Y^*_j)$ to evaluate the disclosure risk of each original value y_{ij} being disclosed by the corresponding correlation-attack estimator $\tilde{Y}_{ij}$. That is:

$$R_\rho(y_{ij}, \delta|Y^*_j) = P(|\frac{\tilde{Y}_{ij} - y_{ij}}{y_{ij}}| < \delta). \tag{6}$$

To calculate the disclosure risk R_ρ requires extensive simulation. It makes data providers uneasy to assess the disclosure risk against the correlation-attack estimator for a noise candidate C_j. However, we could avoid using extensive simulation by defining an alternative disclosure risk measure R_{cor}. Recall that in Section 4, for large sample size, we could let that $\tilde{Y}_{ij} = (1-\rho^2_{Y_jY_j^*})\bar{y_j} + \rho^2_{Y_jY_j^*}Y^*_{ij}$, so only randomness in $\tilde{Y}_{ij}$ is C_j. Therefore we could define the disclosure risk measure R_{cor} as following:

$$R_{cor}(y_{ij}, \delta|Y_j^*) = P(\frac{[-\delta+1]y_{ij} - k}{\rho^2_{Y_jY_j^*}y_{ij}} < C_j < \frac{[\delta+1]y_{ij} - k}{\rho^2_{Y_jY_j^*}y_{ij}}), \tag{7}$$

where $k = (1-\rho^2_{Y_jY_j^*})\bar{Y}_j$.

Normally speaking, data providers only know the original data without knowing the underlying original variable Y_j, so that $\rho^2_{Y_jY_j^*}$ is not known. In this case, data providers could obtain an estimate of $\rho^2_{Y_jY_j^*}$ by replacing corresponding original sample estimates of Y_j into Equation (3), so that data providers could use Equation (7) to evaluate the disclosure risk of any original observation y_{ij} being disclosed by the correlation-attack estimator without relying on simulations. A low $R_\rho(y_{ij}, \delta|Y_j^*)$ value sufficiently guarantees that y_{ij} is protected against the correlation-attack.

To simplify the notation, hereafter, we use notation $R_{LW}(y_{ij}, \delta)$ and $R_\rho(y_{ij}, \delta)$ instead of $R_{WL}(y_{ij}, \delta|Y_j^*)$ and $R_\rho(y_{ij}, \delta|Y_j^*)$.

In the literature, identity disclosure occurs if data intruders successfully identify a record through quasi-identifiers (see introduction in Li and Sarkar, 2013), or through record-linkage technique (Kim and Winkler, 1994). Given noise multiplied microdata, we may not know which records could be identified by data intruders. In this paper, we conservatively assume that all records are vulnerable to identity disclosure. Therefore, we aim to guarantee that, noise multiplied values could not be unmasked by data intruders, so that successfully identifying a record gives data intruders no benefit of learning private information of individuals. Based on our discussion in Section 4, we propose that data providers to use a disclosure risk measure $R(y_{ij}, \delta)$ to measure the disclosure risk of y_{ij}, where

$$R(y_{ij}, \delta) = \begin{cases} R_{LW}(y_{ij}, \delta), & \text{if } MSE(Y^*_{ij}|y_{ij}) < MSE(\tilde{Y}_{ij}|y_{ij}), \\ R_{cor}(y_{ij}, \delta), & \text{if } MSE(Y^*_{ij}|y_{ij}) \geq MSE(\tilde{Y}_{ij}|y_{ij}). \end{cases}$$

When $R(y_{ij}, \delta)$ is used for noise candidate selection, data providers may look at aggregated values to understand the overall disclosure risk of using each noise candidate. For instance, data providers could look at the average value $\overline{R(y_{ij}, \delta)} = \frac{\sum_{i=1}^n R(y_{ij},\delta)}{n}$ to understand the overall disclosure risk, and make comparisons regarding the level of protections among different noise candidates.

For illustration purpose of this paper, we propose that y_{ij} is sufficiently perturbed if $R(y_{ij}, \delta)$ is no greater than p_{thr}, where p_{thr} is a pre-specified threshold risk value determined by data providers. If we aim to protect all observations, it is sufficient to guarantee that $max\{R(y_{ij}, \delta), i = 1, 2, \cdots, n\} < p_{thr}$. Therefore, we define the criteria of disclosure risk assessment for noise candidates selection in the following way:

For a noise candidate, if it guarantees that $max\{R(y_{ij}, \delta), i = 1, 2, \cdots, n\} < p_{thr}$, then we say it offers an acceptable level of protection to the original data. Otherwise we say it could not adequately protect the original data.

We will use this criteria for noise generating variable selections in our simulation. We will also use R-U map (Duncan et al. 2001; 2004) to help us with the decision-makings.

6 Attacker's motivation to carry out correlation-attack

In this section we mention two potential motivations due to which data intruders may want to use the correlation-attack estimator to estimate the value of an original observation.

The first motivation is that, when data intruders have strong belief that the original data is masked by noises which follow a set of specific distributions. It is proposed in the literature that noise generating random variables following a few specific distributions provide better protections against the unbiased estimator to the original values. These distributions include truncated triangular distribution (Kim, 2007; Kim and Jeong, 2008), trapezoidal distributions (Kim, 2007), bi-modal normal distributions (Lin and Wise, 2012) and mixtures of uniform distributions (Klein et al., 2014). The truncated triangular distribution is also used by the U.S. Bureau of the Census for masking the Commodity Flow Survey data (Kim and Jeong, 2008).

The reason for proposing these distributions is that, noise terms following these distributions are unlikely to take values around 1, therefore most of original values are perturbed by a reasonable amount. It can be shown that (as we will show in Simulation 1), a noise generating random variable following one of these distributions has a very low disclosure risk if it is measured by R_{LW}. However, these distributions are proposed by assuming that data intruders estimate the original values using the perturbed values. When the correlation between the original data and the noise multiplied data is reasonably high and correlation-attack is used, there is no guarantee that the disclosure risk measured by R_ρ is also low. This point will be shown in Simulation 1.

Another motivation is that, even though data intruders only observe the noise multiplied value, data intruders could still know which original values are more vulnerable to the correlation-attack estimator, as the values of a,b,c and d defined in Lemma 1 of Section 4 could be reasonably estimated by data intruders using noise multiplied data. Even though data intruders may not know for sure whether the original value of a targeted noise multiplied data entry is more vulnerable to any particular estimator, the noise multiplied value itself may convey a strong information regarding whether any particular estimate should be used. This point is very evident in Simulation 2 in the next section. In the simulation, the original values range from 1 to 768742. If a noise candidate with variance 31/300 is used for data masking, it means that original values over 26317.6 are more vulnerable to the correlation-attack estimator. This information could be reasonably estimated by data intruders. The consequence is that, if a masked value is significantly greater than 26317.6, say 200000, it is very unlikely that the corresponding original value is less than 26317.6. Therefore, it is very likely that data intruders use the correlation-attack estimator to unmask this value, which is also very likely to be the wise and correct choice.

In fact, if a symmetric positive noise generating variable C_j, with mean 1, is used to generate noise multiplied data, which is often the case assumed in the literature, the range of C_j will be a subset of $[0, 2]$ and

$$E(\frac{1}{C_j}) = \lim_{\epsilon\to 0}\int_{\epsilon}^{2-\epsilon} \frac{1}{c_j} f_{C_j}(c_j)dc_j = \lim_{\epsilon\to 0}\int_{\epsilon}^{1} (\frac{1}{c_j} + \frac{1}{2-c_j}) f_{C_j}(c_j)dc_j$$

$$\geq 2\int_{\epsilon}^{1} f_{C_j}(c_j)dc_j \geq 1.$$

Thus, data intruders could have a rough idea about the expected value of the original observation y_{ij} based on its perturbed value y_{ij}^*. This is because $E(Y_{ij}|Y_{ij}^* = y_{ij}^*) = y_{ij}^* E(\frac{1}{C_j}) \geq y_{ij}^*$. If the expected original value belongs to the interval where the correlation-attack estimator is more effective, then it is incentive for data intruders to carry out correlation-attack to unmask the value, especially if the estimated correlation coefficient $\tilde{r}_{YY^*}$ is also very strong. The disclosure risk measure $R(y_{ij}, \delta)$ could efficiently help data providers to protect the original data against rational data intruders.

7 Simulations

In this section we present two simulations. The first simulation aims to show that, the correlation-attack could still lead to serious value disclosure risk even though R_{LW} is controlled to be 0. Moreover, we aim to show how different correlation coefficients between the original random variable Y_j and the noise multiplied random variable Y_j^* lead to different levels of disclosure risks, which are measured by R_{cor}. Finally we want to show that R_{cor} is very close to R_ρ for large sized data.

The second simulation aims to show how could data providers choose the best noise candidate among various noise candidates by using our proposed disclosure risk measure in a R-U map (Duncan et al., 2001; 2004). By fixing the variance of noise candidates, we use a R-U map to identify the best noise candidate which satisfies that the overall disclosure risk is controlled with the lowest utility loss. In both simulations, all the data are one-dimensional, therefore we denote the original data as $y = (y_1, y_2, \cdots y_n)$ without the subscript j indicating the j-th attribute values.

7.1 Simulation 1

In this simulation we aim to show three things: the first one is, even though R_{LW} could be easily controlled by choosing an appropriate noise generating variable, the disclosure risk from the correlation-attack may still be significant; the second one is, we aim to show how accurate it is of using $\tilde{r}_{yy^*}$ to estimate ρ_{YY^*} and using R_{cor} to estimate R_ρ; the third one is, we aim to show that, if several noise candidates have the same type of distribution (all normal, all uniform, etc), the noise candidate with a higher variance offers a better protection to the original data against the correlation-attack estimator.

Suppose one dimensional data contains 1000 observations, denoted as $y = (y_1, y_2, \cdots y_{1000})$. The records are independently drawn from $Y \sim U(100, 200)$. We consider the following four mixture of uniforms noise generating variables to mask the data:

C_1: $I_1U_{1,1} + (1 - I_1)U_{1,2}$, C_2: $I_2U_{2,1} + (1 - I_2)U_{2,2}$,

C_3: $I_3U_{3,1} + (1 - I_3)U_{3,2}$, C_4: $I_4U_{4,1} + (1 - I_4)U_{4,2}$,

where $\{I_i\}_{i=1}^4$ are independent Bernoulli random variables taking 0 and 1 with equal probability; $U_{1,1} \sim U(0.8, 0.9)$, $U_{1,2} \sim U(1.1, 1.2)$; $U_{2,1} \sim U(0.7, 0.9)$, $U_{2,2} \sim U(1.1, 1.3)$; $U_{3,1} \sim U(0.6, 0.9)$, $U_{3,2} \sim U(1.1, 1.4)$; $U_{4,1} \sim U(0.5, 0.9)$, $U_{4,2} \sim U(1.1, 1.5)$. The noise candidates C_1 and C_4 are used in Klein et al. (2014) for generating noise multiplied data.

We assume the acceptance rule $\delta = 0.1$. It can be easily shown that, all four noise generating variables have $R_{LW} = 0$ for all the original values. If R_ρ is not used to assess the potential value disclosure risks against the correlation-attack, then data providers may have an impression that all the noise generating variables provide strong protections to the original data.

Now we measure disclosure risks of using the proposed noise candidates to mask the original data y using R_ρ and R_{cor}. Recall that, calculating $R_\rho(y_i, 0.1)$ requires extensive simulation, while calculating $R_{cor}(y_i, 0.1)$ is fairly easy if the density function of the noise generating variable is known. For large-sized data, we propose that data providers to use R_{cor} to calculate the disclosure risk against correlation-attack as R_{cor} should be fairly close to R_ρ. In this simulation, as sample-size of y is fairly large, we aim to show that R_ρ and R_{cor} produce similar results.

To calculate R_ρ, we rely on software "R" to simulate the masking and then correlation-attack process 5000 times for each noise candidate. For instance, when calculating R_ρ of using C_1 as the noise generating variable to mask y, in each iteration we randomly simulate 1000 noises from C_1 and multiply them to y to produce the noise multiplied data y^*. Then, we perform the correlation-attack process introduced in Section 3. Through the process, we obtain $\tilde{r}_{yy^*}$, the estimated population correlation coefficient between Y and Y^*, and we also obtain 1000 correlation-attack estimates of the original values. Then, we create a binary vector of size 1000 indicating whether an original value is disclosed by the corresponding correlation-attack estimate. For instance, if the correlation-attack estimate $\tilde{y}_i$ discloses the value of y_i, i.e. $0.9y_i < \tilde{y}_i < 1.1y_i$, then we put 1 in the i-th entry of the binary vector indicating that y_i is disclosed by the correlation-attack estimate. If $\tilde{y}_i$ did not disclose the value of y_i, we put 0 in the i-th place of the binary vector. After 5000 iterations, we combine all 5000 binary vectors to estimate $R_\rho(y_i, 0.1)$ for $i = 1, 2, \cdots, 1000$. To compute $R_\rho(y_i, 0.1)$, we count the number of iterations in which the value of y_i is disclosed by the correlation-attack estimate by looking at the binary vectors. If y_i is disclosed in N iterations, then we divide N by 5000 to estimate $R_\rho(y_i, 0.1)$. Moreover, as we have also obtained the $\tilde{r}_{yy^*}$ in each iteration (in total we have 5000 of them), we compute the average and the standard deviation of the 5000 items, denoted as $\widehat{\tilde{r}_{YY^*}}$ and $sd(\tilde{r}_{yy^*})$. We report the summary statistics of $\{R_\rho(y_i, 0.1)\}_{i=1}^{1000}$ for each noise candidate as well as $\widehat{\tilde{r}_{YY^*}}$ and $sd(\tilde{r}_{yy^*})$ in Table 1.

Table 1 gives the summary statistics of $\{R_\rho(y_i, 0.1)\}_{i=1}^{1000}$. To interpret the table, for instance, the first quartile Q_1 of $\{R_\rho(y_i, 0.1)\}_{i=1}^{1000}$ when noise generating variable is C_1 is 0.496, meaning that 75% of the original values have probabilities of more than 49.6% of being disclosed by the corresponding correlation-attack estimates. Similarly, Q_2 of $\{R_\rho(y_i, 0.1)\}_{i=1}^{1000}$ is 0.503 when noise generating variable is C_2, indicating that 50% observations have more than 50.3% probabilities of being disclosed.

Table 1 also shows that, the value of $\tilde{r}_{yy^*}$ is very consistent in each iteration ($sd(\tilde{r}_{yy^*})$ is very small) and is very close to the theoretical correlation ρ_{YY^*} (shown in Table 2). Table 1 also shows

Table 1: Summary statistics of $\{R_\rho(y_i, 0.1)\}_{i=1}^{1000}$ if y is masked by C_1, C_2, C_3 and C_4, respectively.

noise	$\widehat{\tilde{r}_{yy^*}}$	$sd(\tilde{r}_{yy^*})$	Min	Q_1	Q_2	$mean$	Q_3	Max
C_1	0.768	0.0085	0.0864	0.496	0.504	0.531	0.583	0.709
C_2	0.661	0.0139	0	0.492	0.503	0.477	0.574	0.663
C_3	0.570	0.0181	0	0.383	0.498	0.447	0.584	0.713
C_4	0.496	0.0216	0	0.252	0.494	0.422	0.591	0.785

Table 2: Summary statistics of $\{R_{cor}(y_i, 0.1)\}_{i=1}^{1000}$ if y is masked by C_1, C_2, C_3 and C_4, respectively.

noise	ρ_{YY^*}	Min	Q_1	Q_2	$mean$	Q_3	Max
C_1	0.778	0.241	0.500	0.500	0.529	0.571	0.652
C_2	0.672	0	0.500	0.500	0.476	0.567	0.608
C_3	0.581	0	0.381	0.500	0.443	0.572	0.655
C_4	0.507	0	0.250	0.500	0.415	0.579	0.723

a clear trend between ρ_{YY^*} and the mean of $\{R_\rho(y_i, 0.1)\}_{i=1}^{1000}$. If ρ_{YY^*} is high, then the mean of $\{R_\rho(y_i, 0.1)\}_{i=1}^{1000}$ shows that overall disclosure risk is also high in general. **That means, if the distributions of noise candidates are the same, the noise candidate with a large variance offers a better protection against the correlation-attack, which has also been showed in our discussion about the MSE of the correlation-attack estimator in Section 4**.

We also computed the results of ρ_{YY^*} and summary statistics of $\{R_{cor}(y_i, 0.1)\}_{i=1}^{1000}$ in Table 2. The aim is to show how close the results of $\tilde{r}_{yy^*}$ and $\{R_\rho(y_i, 0.1)\}_{i=1}^{1000}$ are to ρ_{YY^*} and $\{R_{cor}(y_i, 0.1)\}_{i=1}^{1000}$. Recall that calculating $\{R_{cor}(y_i, 0.1)\}_{i=1}^{1000}$ does not rely on simulation and therefore we propose that data providers to compute $\{R_{cor}(y_i, 0.1)\}_{i=1}^{1000}$ instead of $\{R_\rho(y_i, 0.1)\}_{i=1}^{1000}$ when assessing the disclosure risk of a noise candidate.

Overall, Simulation 1 shows that, the disclosure risk measure R_ρ captures serious value disclosure risk, even when $R_{LW} = 0$. Therefore, it is important for data providers to evaluate disclosure risk using R_ρ as well during noise candidate selection. We also showed that, data providers could reduce the disclosure risk by choosing a noise candidate with a larger variance. Finally, we showed that data providers could use the disclosure risk measure R_{cor} to assess the disclosure risk against correlation-attack as it is easier to calculate than R_ρ.

8 Simulation 2

In this section we will show how could data providers select an appropriate noise generating variable using the disclosure risk measure $R(y_i, \delta)$ we proposed in Section 4 in a R-U map.

Even though the primary concern for data providers is to control the disclosure risk, data providers cannot choose a noise generating variable with a too large variance, as doing so may significantly reduce the analytical validity of the noise multiplied data. Moreover, if unbounded noise generating variable is used, such as normally distributed noise, then a too large noise variance may result in negative noise terms (Sinha et al., 2011) which may violate the positivity nature of an attribute. Therefore, data providers need to carefully select the noise generating variable which satisfies the overall disclosure risk requirement, while maintains as high data utility as possible. The disclosure risk measure $R(y_i, \delta)$ could help data providers to assess the overall disclosure risk of any noise candidate, and the utility loss measures introduced in Section 5 could help data providers to decide the noise candidate which maintains a better analytical validity among a set of noise candidates.

We use the public use data from the 2000 Current Population Survey (CPS) March supplement (available from http://www.census.gov/cps/). The data set is also used in Klein et al. (2014). The entire data set contains household, family, and individual records. In this paper, we only consider the one-dimensional attribute **household income**, and we consider all positive household income values as the original data (Recall that data entries with value 0 cannot be protected by multiplicative noises and therefore we do not consider 0 incomes). The data provider needs to select an appropriate noise candidate among a set of noise candidates with the aid of a R-U map. The data provider only seeks to preserve the first two moments estimations in the noise multiplied variable.

The original data contains 50661 positive observations ranging from 1 to 768742, with mean 53007 and variance 2411407246. The data is skewed to the right. In the following we define the data as $y = (y_1, y_2, \cdots, y_{50661})$. The histogram of the data is given in Figure 1.

In Klein et al. (2014), four mixture of uniforms noise candidates were proposed. As a side note, the potential disclosure risk from the correlation-attack was not considered in their paper. The disclosure risk they considered in the paper was from the fact that data intruders could obtain a generalised linear regression model from noise multiplied data set using their proposed method, and the fitted values obtained from the regression line were considered as estimates of original household income values. Our simulation showed that the correlation-attack could be carried out in the context of their paper and we showed that the disclosure risk from the correlation-attack tends to

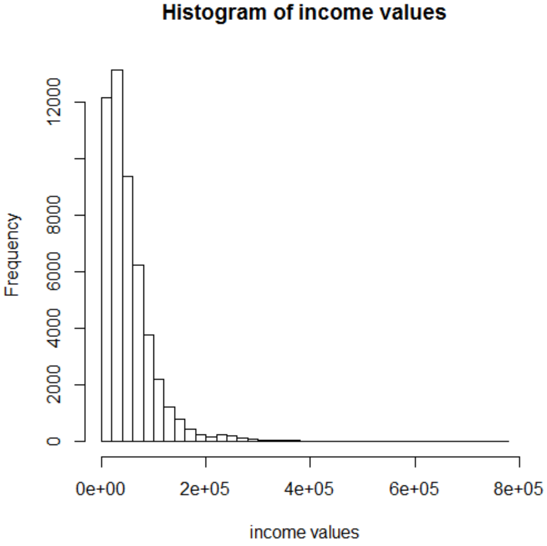


Figure 1: Histogram of income data.

be higher than the disclosure risk came from the fitted values. However, the context we consider in Simulation 2 is different from the context considered in their paper, so we do not present the results in the paper.

For illustration purposes, in this section we consider noise candidate C_4 (used in Simulation 1) as the benchmark noise generating variable. We also consider the following noise candidates which have the same variances as C_4. The noise candidates are: $C_5 \sim U(1-0.5\sqrt{93/75}, 1+0.5\sqrt{93/75})$; $C_6 \sim N(1, 31/300)$; $C_7 \sim I_5N_1 + (1-I_5)N_2$, where $P(I_5 = 0) = P(I_5 = 1) = 0.5$, $N_1 \sim N(0.7, 4/300)$ and $N_2 \sim N(1.3, 4/300)$; $C_8 \sim I_6T_1 + (1-I_6)T_2$, where $P(I_6 = 0) = P(I_6 = 1) = 0.5$, T_1 and T_2 are triangular random variables with three parameters $(1.1 - \sqrt{\frac{9.6}{4}}, 0.9, 0.9)$ and $(1.1, 1.1, 0.9 + \sqrt{\frac{9.6}{4}})$ respectively.

The distributions of these noise generating variables have been used in the literature for producing noise multiplied data. For instance, C_6 follows a normal distribution, which was considered in Sinha et al., (2011). C_7 follows a bi-modal normal distribution, which was proposed in Lin and Wise (2012) and C_8 follows a truncated triangular distribution which was proposed in Kim (2007).

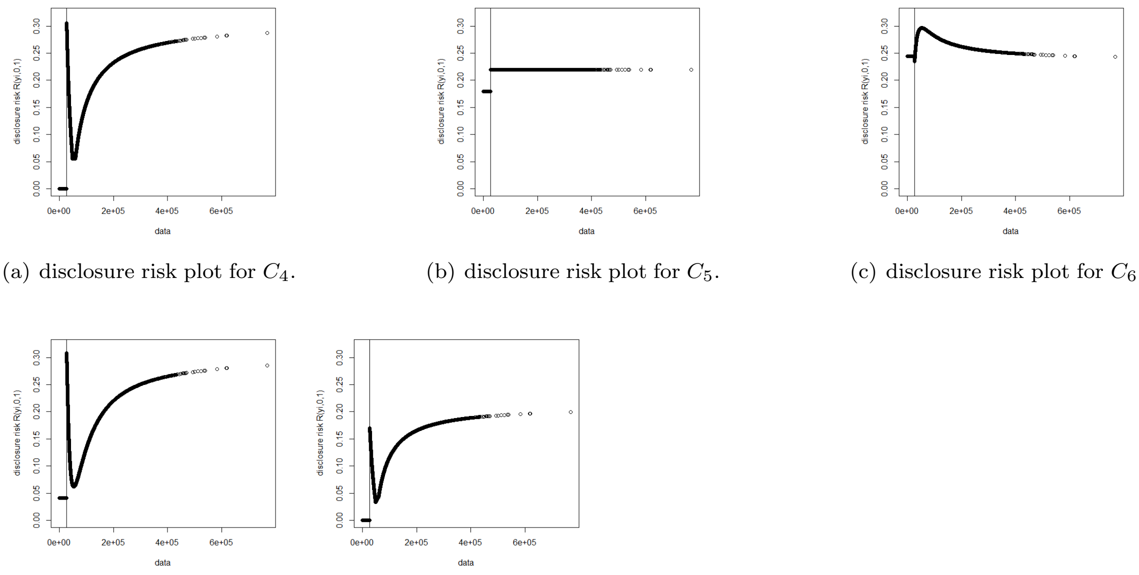
Now we assume the role of the data provider and we aim to find an appropriate noise candidate to use during masking process. As the noise candidates have the same variances, initial analysis shows that, these noise candidates will result in roughly 0.903 correlation coefficients between the original data and the noise multiplied data. Furthermore, using these noise candidates, the values of a, b, c and d defined in Lemma 1 in Section 4 are -0.405, 2.121, 26317.6 and -3749526, respectively, meaning that original values greater than 26317.6 are more vulnerable to the correlation-attack estimator. Recall that in Section 5, we propose to use $R(y_i, 0.1)$ to measure disclosure risk of y_i. In this application, we have $R(y_i, 0.1) = R_{LW}(y_i, 0.1)$ for $y_i < 26317.6$ and $R(y_i, 0.1) = R_{cor}(y_i, 0.1)$ for $y_i \geq 26317.6$. As we assume that all records are vulnerable to identity disclosure, we need to guarantee that $max\{R(y_i, \delta), i = 1, 2, \cdots, n\} < p_{thr}$. We let $p_{thr} = 0.3$ for illustration purposes, so that a noise generating variable is satisfactory in terms of disclosure risk control if it guarantees that the probability of any observation being disclosed is less than 0.3 if the noise generating variable was used to mask y.

The fact that all the noise candidates have the same variances means that they all result in the same amount of UL1 (see Section 5). Therefore, a noise candidate with the lowest fourth moment has the lowest UL2 (see Section 5) and hence the highest data utility preservation in terms of first two moment estimations. Using a R-U map, we would be able to find out the noise candidate which satisfies the requirement of disclosure risk while preserves the most data utility.

The disclosure risk plots of $\{R(y_i, 0.1)\}_{i=1}^{50661}$ for all five noise generating variables are given in Figure 2. The vertical line of each plot corresponds to the horizontal value 26317.6, meaning that the disclosure risk plot on the left of the line was created by R_{LW} and the disclosure risk plot on the right of the line was created by R_{cor}. We can see that, except a few data points in the case of $C6$, the disclosure risks measured by R_{cor} for data values greater than 26317.6 are larger than that measured by R_{LW}, meaning that the correlation-attack estimator is more accurate in terms of estimating original values greater than 26317.6. It coincides with our conclusion in Section 4.

Figure 2 also showed how different noise candidates protect original values differently. For instance, C_5 offers uniform protections against both the unbiased estimator and the correlation-attack estimator. C_6 offers the worst protection against the unbiased estimator but it protects some very large original values against the correlation-attack estimator reasonably well (the larger the original value, the lower the disclosure risk). In practice, data providers could choose the noise candidate which meets their needs. For instance, if extremely large original values are more vulnerable to identity disclosure, then data providers may choose C_6 to protect the original data as it may offer a good protection to extreme values.

The result of R-U map is shown in Figure 3. The noise candidate satisfying the requirement of disclosure risk control with the lowest UL2 value is the ideal noise candidate under our context.



(a) disclosure risk plot for C_4. (b) disclosure risk plot for C_5. (c) disclosure risk plot for C_6.

(d) disclosure risk plot for C_7. (e) disclosure risk plot for C_8.

Figure 2: Disclosure risk plots for different noise candidates.

Figure 3 shows that, noise candidates C_4 and C_7 cannot guarantee the requirement of disclosure risk that the disclosure risks of all original values are controlled to be below 0.30. Noise candidate C_8 offers the best overall protection but it also offers a low data utility preservation. We propose that C_5 is the best noise candidate to use during masking stage as it satisfies the requirement of disclosure risk control while it offers the lowest utility loss.

To sum up, in this simulation we showed how could the disclosure risk measure $R(y_i, \delta)$ be used to help data providers to make decision about noise generating variable selection during masking stage. The noise generating variable which preserves the highest data utility and satisfies the requirement of disclosure risk control could be considered as the best noise candidate for masking the original data.

9 Conclusion and Future work

In this paper we showed the correlation-attack estimator which could be used by data intruders to breach noise multiplied data. The comparison between the correlation-attack estimator and the unbiased estimator which is frequently used in the literature for estimating original values were made and the result showed that the correlation-attack estimator results in more disclosure risks

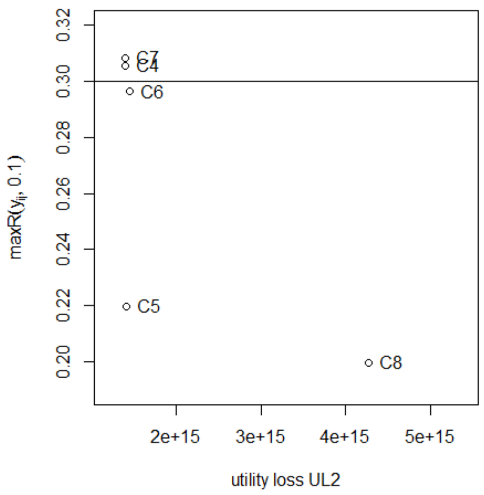


Figure 3: R-U plot of noise candidates.

for some original observations. Therefore, we proposed a disclosure risk measure for data providers to evaluate value disclosure risk against the correlation-attack estimator. The proposed disclosure risk measure could be used to help data providers with decision-makings on noise candidates selection during masking stage.

In the literature, various distributions which noise generating variables should follow have been proposed such that the noise multiplied value is unlikely to reveal the original value. It may also be interesting to find out the best noise distribution which significantly reduce the disclosure risk against the correlation-attack estimator without losing too much data utility. Finding a set of noise distributions which behave well against the correlation-attack requires further study.

References

[1] Agrawal, R. and Srikant, R. (2000). Privacy preserving data mining, in *Proceedings of the ACM SIGMOD*, pp. 439-450.

[2] An, D., and Little, R. J. A. (2007). Multiple imputation: An alternative to top coding for statistical disclosure control. *Journal of Royal Statistical Society*, Series A, **170**, 923-940.

[3] Brand, R. (2002). Microdata protection through noise addition. In *Inference Control in Statistical Databases*, vol. 2316 of LNCS. Springer Berlin Heidelberg, pp. 61-74.

[4] Domingo-Ferrer, J., Sebé, F., and Castellà-Roca, J. (2004). On the security of noise addition for privacy in statistical databases. *Lecture notes in computer science*, **3050**, 149-161.

[5] Domingo-Ferrer, J. and Torra, V. (2001): 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

[6] 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*, Los Alamos National Laboratory, Statistical Sciences Group, Los Alamos, New Mexico.

[7] Duncan, G., Keller-McNulty, S., and Stokes, S. (2004). Database security and confidentiality: examining disclosure risk vs. data utility through the R-U confidentiality map. *Techncal Report Number 142*, National Institute of Statistical Science.

[8] Evans, T. (1996). Effects on trend statistics of the use of multiplicative noise for disclosure limitation. U.S. Bureau of the Census, http://www.census.gov/srd/sdc/papers.html.

[9] Guo, S., Wu, X., and Li, Y. (2006). Deriving private information from perturbed data using IQR based approach. *Proceedings of the 22nd international conference on data engineering workshops*, Atlanta, 92-101.

[10] Hundepool, A., van de Wetering, A., Ramaswamy, R., de Wolf, P.P., Franconi, L., Brand, R., and Domingo-Ferrer, J. (2008). *μ-ARGUS version 4.2 Software and User's Manual*, Statistics Netherlands, Voorburg NL. (available from http://neon.vb.cbs.nl/casc/mu.htm).

[11] Hwang, J. T. (1986). Multiplicative errors-in-variables models with applications to recent data released by the U.S. Department of Energy. *Journal of American Statistical Association*, **81**, 680-688.

[12] Kargupta, H., Datta, S., Wang, Q., and Sivakumar, K. (2003). On the privacy preserving properties of random data perturbation techniques. *Proceedings of the 3rd IEEE international conference on data mining*, Melbourne, 99-106.

[13] Klein, M., Mathew, T., and Sinha, B. (2014). Noise multiplication for statistical disclosure control of extreme values in log-normal regression samples. *Journal of Privacy and Confidentiality*, **6**, 77-125.

[14] Kim, J. J. (1986). A method for limiting disclosure in microdata based on random noise and transformation. In *Proceedings of the American Statistical Association*, Section on Survey Research Methods, 303-308.

[15] Kim, J. J. (1990). Subpopulation estimation for the masked data. Proceedings of the Section on Survey Research Methods. *American Statistical Association*, 456-461.

[16] Kim, J. J. (2007). Application of truncated triangular and trapezoidal distributions for developing multiplicative noise. Proceedings of the Survey Methods Research Section, *American Statistical Assoication*, CD Rom.

[17] Kim, J., Jeong, D. M. (2008). Truncated triangular distribution for multiplicative noise and domain estimation. Section on Government Statistics-JSM 2008, 1023-1030.

[18] Kim, J.J., Winkler, W. E. (2003). Multiplicative noise for masking continuous data. Statistical Research Division, Research Report Series(Statistics #2003-01). U.S. Census Bureau.

[19] Li, X. B., Sarkar, S. (2013) Class-restricted clustering and microperturbation for data privacy. *Management Science*, 59(4), 796–812.

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

[21] Massell, P. and Russell, N. (2006). Protecting confidentiality of commodity flow survey tabular data by adding noise to underlying microdata. presented at a Washington Statistical Society Seminar, October 24, 2006.

[22] Muralidhar, K., Batra, D., and Kirs, P. J. (1995). Accessibility, security, and accuracy in statistical databases: the case for the multiplicative fixed data perturbaton approach. *Management Science*, **41**, No.9, 1549-1564.

[23] Muralidhar, K., and Sarathy, R. (2006). Data Shuffling-A New Masking Approach for Numerical Data. *Management Science*, 52(5), 658-670.

[24] Nayak, T. K., Sinha, B., and Zayatz, L. (2011). Statistical properties of multiplicative noise masking for confidentiality protection. *Journal of Official Statistics*, **27**, No.3, 527-544.

[25] Oganian, A. and Karr, A. (2011). Masking methods that preserve positivity constraints in microdata. *Journal of Statistical Planning and Inference*, **141**, 31-41.

[26] Rubin, D. B. (1993). Discussion of statistical disclosure limitation. *Journal of Official Statistics*, **9**, 461-468.

[27] Sinha, B., Nayak, T.K., and Zayatz, L. (2011). Privacy protection and quantile estimation from noise multiplied data. *Sankhya B*, **73**, No. 2, 297-315.

[28] Yancey, W.E., Winkler, W.E., and Creecy, R.H. (2002). Disclosure risk assessment in perturbative micro-data protection. *Inference Control in Statistical Databases (ed. J. Domingo-Ferrer)*, New York: Springer, 135-151.