'Easy to Implement' is Putting the Cart before the Horse: Effective Techniques for Masking Numerical Data

Krish Muralidhar
Gatton Research Professor
University Of Kentucky
Lexington KY 40513

Rathindra Sarathy
Ardmore Professor
Oklahoma State University
Stillwater OK 74078

1. Introduction

In a recent paper, William Winkler (2006) of the Census Bureau observed the following:

> "Statistical agencies have typically adopted masking methods because they are *easy to implement*. The easiest-to-implement methods seldom, if ever, have been justified in terms of preserving one or two analytical properties and preventing re-identification. In extreme situations, the crude application of masking methods may yield a file that cannot be used for analyses yet still allows some re-identification." (page 4)

We agree strongly with this statement, particularly for numerical confidential data. In recent years, there have been significant developments in masking techniques for numerical data. While there have been similar advances for protecting categorical confidential data, our analysis in this paper is restricted to numerical, confidential data. One key aspect of some of the newer techniques is that their performance is theoretically predictable. This is true for both information loss and disclosure risk. This is not to say that earlier methods did not provide this ability; obviously, we could assess the extent to which there was information loss using the additive noise approach. However, there were other techniques (micro-aggregation and data swapping for instance) where, while it is possible to make some generalizations ("data swapping results in correlation attenuation" or "the micro-aggregation results in variance attenuation"), it was difficult to *predict* their performance characteristics for a given data set.

Yet, there seems to be a tendency among statistical agencies to use micro-aggregation and data swapping more frequently than other techniques. There are several explanations for this state of affairs, the primary one being the opinion expressed by Mr. Winkler and is the title of this paper. We believe that since some techniques are easy to explain and easy to implement, they are preferred over other more sophisticated techniques that are, in relative terms, more difficult to implement and explain. After all, it is very easy to understand, explain, and implement data swapping than it is to understand, explain, and implement data shuffling. Yet, in terms of performance, there is no doubt that data shuffling is superior to data swapping; *data shuffling provides lower information loss and lower disclosure risk than data swapping*.

In this study, we demonstrate the superiority of theoretically sound techniques over "easy to implement" techniques such as micro-aggregation and data swapping, for masking numerical data. Based on a theoretical evaluation, we identify two techniques (sufficiency-based linear models perturbation and data shuffling) that have the best performance characteristics. We then evaluate their performance on two data sets (one simulated and one real data set). In addition, we also evaluate the ability of these techniques to maintain characteristics of sub-domains of the data, something that has not been evaluated previously. We hope that the results of this paper will provide some impetus for the adoption of these techniques.

2. Assessing Disclosure Risk and Information Loss

In our paper titled "A Theoretical Basis for Perturbation Methods" (Muralidhar and Sarathy, 2003), we developed specific measures for assessing disclosure risk and data utility. The primary objective of that paper was to provide, at least in theory, the optimal method for perturbation. In that paper, we showed that if we are able to generate the perturbed values from the *true conditional distribution* of the confidential variables given the non-confidential variables, the resulting perturbed values

provided both the highest level of security (lowest level of disclosure risk) and the highest possible level of data utility (lowest possible information loss). These theoretical derivations have important implications for practice.

2.1. Disclosure Risk

There are two important issues relating to disclosure risk. First, the assessment of disclosure risk must be partitioned into two separate categories: disclosure risk attributable to the release of the non-confidential variables (if any) and disclosure risk attributable to the masking technique. Without such partitioning, it would be impossible to accurately assess the extent to which the making technique results in increased disclosure. For instance, if there are several non-confidential categorical variables that uniquely identify an individual in the data set, then regardless of the technique chosen to mask the numerical variables, an intruder would be able to identify that individual. In this case, the resulting identity disclosure must be attributed to the non-confidential variables. For the sake of argument, assume that a data provider first releases summary information (such as the mean vector and covariance matrix) for all the variables (both confidential and non-confidential). Subsequently, the same data provider also releases microdata on the non-confidential variables (but not the confidential variables). At this stage, no masking technique has been employed and no confidential microdata has been released. Yet, an intelligent intruder would be able to analyze the released data to identify a record as belonging to an individual or to predict the value of a confidential variable through techniques such as regression analysis using only the non-confidential variables. The level of disclosure that can be attributed to the non-confidential variables forms the baseline disclosure risk. After the release of the masked data set, the intruder will use the masked variables in addition to the non-confidential variables, to estimate the confidential values. This second step provides an assessment of the *incremental* disclosure that can be attributed to the masking technique, over and above that resulting from the non-confidential variables alone. Obviously, in cases where there are no non-confidential variables, the benchmark disclosure risk resulting from non-confidential variables will be zero. A good masking technique will result in little or no additional disclosure risk than that resulting from the non-confidential variables alone.

The second issue relates to the *extent* of disclosure risk of a masking technique. From the discussion in the previous paragraph, in order for a masking technique to provide the highest possible level of security, it is necessary that the disclosure risk after releasing the masked data not be higher than that from releasing only the non-confidential and summary data. This can be achieved if the masked values of the confidential variables are generated as a function only of the non-confidential variables and an independent noise term. Our derivations show that when the masked values are generated in this manner, the incremental disclosure risk resulting from the release of the masked values is zero. This can be illustrated using simple regression analysis. Consider a data set with a set of confidential variables $\mathbf{X}$ and a set of non-confidential variables $\mathbf{S}$. Assume that a data provider has released aggregate information on both $\mathbf{X}$ and $\mathbf{S}$ (in the form of mean vectors and covariance matrices). In addition, the data provider has also released microdata on $\mathbf{S}$. Using this information alone, an intruder would be able to predict the value of $\mathbf{X}$ using $\mathbf{S}$ as:

$$\mathbf{X} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{S}. \tag{1}$$

Assume that the proportion of variability explained in $\mathbf{X}$ using the above equation is $R^2(\mathbf{X}|\mathbf{S})$. Since no confidential microdata has been released, none of this disclosure can be attributed to the masking technique itself. Thus, $R^2(\mathbf{X}|\mathbf{S})$ becomes the benchmark for assessing the resulting increase in disclosure (if any) resulting from the masking technique.

Now let us assume that the masked values $\mathbf{Y}$ have been generated as:

$$\mathbf{Y} = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1\mathbf{S} + \boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{\varepsilon}$ is a noise term independent of both $\mathbf{X}$ and $\mathbf{S}$. Now consider an intruder who attempts to predict the value of the confidential variables $\mathbf{X}$ using both $\mathbf{S}$ and $\mathbf{Y}$. The resulting regression equation would be of the form:

$$\mathbf{X} = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1\mathbf{S} + \boldsymbol{\gamma}_2\mathbf{Y} = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1\mathbf{S} + \boldsymbol{\gamma}_2(\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1\mathbf{S} + \boldsymbol{\varepsilon}). \tag{3}$$

We can easily show that in this case, since $\boldsymbol{\varepsilon}$ is independent of both $\mathbf{X}$ and $\mathbf{S}$, the resulting regression model reduces to equation (1). Hence, the proportion of variability in $\mathbf{X}$ explained by both $\mathbf{S}$ and $\mathbf{Y}$, $R^2(\mathbf{X}|\mathbf{S},\mathbf{Y})$, is the same as the proportion of variability explained in $\mathbf{X}$ using $\mathbf{S}$ alone ($R^2(\mathbf{X}|\mathbf{S})$). Thus, using $R^2(\mathbf{X}|\mathbf{S})$ as the benchmark, releasing $\mathbf{Y}$ does not result in *no increase in disclosure risk*. Hence, the disclosure risk attributable to the masking technique is actually zero.

In the above example, **Y** was generated as a linear function of **S**. In the *Statistics & Computing* paper, we show that as long as **Y** is generated as *any function* of **S** and **ε**, where **ε** is *independent* of **S** and **X**, the incremental disclosure risk resulting from the release of the masked microdata will be zero. Conversely, it also follows that when the masked values are generated as a function of **X**, releasing the masked values *will* resulting increased risk of disclosure. Finally, it is important to note that the results provided here are independent of the characteristics of the data set and the type of masking technique used; disclosure risk will be minimized (no more than that explicable using only **S**) and security will be maximized as long as the masked values are generated as a function of **S** and some noise term **ε** and independent of **X**.

In situations where it is necessary to compare the performance of masking techniques that do not satisfy the minimum disclosure risk criterion, security has to be assessed empirically. Typically, such assessments are performed based on the extent to which the masked data results in disclosure of identity and disclosure of value. Identity disclosure is usually assessed by using record linkage and similar techniques. Value disclosure is usually assessed by computing the proportion of variability explained in the confidential variable(s) using the released data.

2.2. Information Loss

In our *Statistics & Computing* paper we showed that the released data set (consisting of masked and non-confidential variables) maintain the characteristics of the original data set (consisting of the confidential and non-confidential variables) when the masked data are generated from the conditional distribution of **X** given **S**, f(**X**|**S**). In most practical situations however, the true conditional distribution f(**X**|**S**) is never known, and may not even be possible to be estimated. Hence, most masking techniques rely on an approximate model of the joint distribution of **X** and **S** from which f(**X**|**S**) is estimated. The information loss is directly related to the extent to which the assumed (modeled) joint distribution deviates from the true joint distribution. Note that, as discussed earlier, we can still assure minimum disclosure risk even if we do not know the true conditional distribution f(**X**|**S**). Many masking techniques (simple noise addition, Kim's method, and multiple imputation) implicitly assume a multivariate normal joint distribution, resulting in a linear model for generating the value of **Y**. Other masking techniques (copula based perturbation and data shuffling) use a more complex model of the joint distribution of **X** and **S**. Yet other approaches (data swapping, micro-aggregation, to name a few), make no explicit assumption regarding the joint distribution of **X** and **S**.

While disclosure risk can be theoretically assessed, information loss assessment must be performed empirically since there is no general theoretical approach for estimating deviations from the true (possibly unknown) joint distribution. Thus, it makes sense to focus on the information loss from the viewpoint of how the masked data set will be used, in place of the original data. For a vast majority of situations, data released by statistical agencies is meant for statistical analyses such as estimation of means, standard deviations and variances, and other parameters and relationships. In all these cases, the interest is in developing masking procedures that can reproduce statistics and statistical relationships that exist in the original data. Consequently, the most sensible measure of information loss would be one that provides an assessment of deviations between the same statistics that are obtained from the original and masked data. In this study, we adopt this approach to define information loss as a *statistical* information loss. One disadvantage of this approach is that we need to assess information loss for each statistic of interest. Given the typical types of statistical analysis that is performed with masked data, we first assess the extent to which the masked variable maintains the marginal distribution of the original variables. This satisfies the requirements of those users who are interested in performing univariate analysis on the data. Next, we assess information loss related to relationships among variables. Typically linear relationships are measured by the covariance matrix (or product moment correlation matrix) among the variables. In addition, we also assess the rank order correlation among the variables to assess the extent to which the masked variables maintain monotonic non-linear relationships among the variables. These assessments are also performed for sub-groups created by non-confidential categorical variables.

Our measures of information loss differ from that proposed by Domingo-Ferrer and Torra (2001). In their case information loss is measured by the extent to which the masked value of a variable for a given record differs from the original value. In other words, the more different the masked value is from the original value, the greater the information loss. We believe that this measure is misguided, since it is not necessary for the masked values to be "close" to the original values, to avoid information loss in the statistics of interest. This measure does not reflect how the masked data is generally used by statistical agencies. Additionally, this measure introduces a false trade-off between information loss and disclosure risk. As we shall show later, it is possible to provide a high level of protection against disclosure while ensuring practically no (statistical) information loss. .

3. An Assessment of Masking Techniques

In this section, we perform an assessment (that is by no means comprehensive) of common masking techniques for numerical confidential data. We focus our attention on those techniques that have received attention in the literature and in practice. We first provide a quick description of these techniques and an assessment of the resulting disclosure risk and information loss.

3.1. Noise Addition

The noise addition model adds random noise to the original confidential variable as follows (using the same notation as in the previous section):

$$\mathbf{Y} = \mathbf{X} + \boldsymbol{\varepsilon}. \tag{4}$$

In the original form of noise addition, the noise term $\boldsymbol{\varepsilon}$ was independent of $\mathbf{X}$ and was typically drawn from a normal distribution with mean vector $\mathbf{0}$ and a covariance structure that was diagonal (and all non-diagonal values were zero). This implied that noise terms were uncorrelated among themselves. The values of the diagonal terms represent the variance of the noise to be added, usually specified as a percentage of the variance of $\mathbf{X}$.

As we have shown above, since $\mathbf{Y}$ is not independent of $\mathbf{X}$, releasing the masked data provides the intruder with additional information, resulting in increased disclosure risk. In terms of information loss in the marginal distribution of the masked variable, the variance of $\mathbf{Y}$ will be higher than that of the original variable $\mathbf{X}$. In addition, since this is an additive model, if the original variable $\mathbf{X}$ is non-normal, the masked variable will be less skewed (and closer to normal) than the original variable. In terms of information loss in relationships, due to the random noise added, the relationship among the $\mathbf{Y}$ variables is different from those among the $\mathbf{X}$ variables and the relationship between ($\mathbf{Y}$ and $\mathbf{S}$) is different from that between ($\mathbf{X}$ and $\mathbf{S}$).

A simple but important variation of this approach was suggested by Kim (1986). In this variation, the covariance structure of $\boldsymbol{\varepsilon}$ ($\boldsymbol{\Sigma}_{\mathbf{ee}}$) specified to be $d\boldsymbol{\Sigma}_{\mathbf{XX}}$ where $\boldsymbol{\Sigma}_{\mathbf{XX}}$ is the covariance matrix of the original $\mathbf{X}$ variables and d is a constant (typically between 0 and 1). The advantage of this specification is that the resulting relationships among $\mathbf{Y}$ are the same as that between the $\mathbf{X}$ variables. However, this procedure still results in higher variance for $\mathbf{Y}$ than that of $\mathbf{X}$ and linear relationships between the non-confidential and masked variables are attenuated. Further, since $\mathbf{Y}$ is not independent of $\mathbf{X}$, this method does not minimize disclosure risk. Generally, the higher the value of d, the greater the information loss and lower the disclosure risk and vice versa. A further modification was proposed by Tendick and Matloff (1994) resulting in the variance of $\mathbf{Y}$ being the same as that of $\mathbf{X}$. With this exception, the performance characteristics do not change.

3.2. Linear Model Based Approaches

Unlike models that add noise only to the confidential variables $\mathbf{X}$, approaches based on a linear model generate the perturbed values $\mathbf{Y}$ using some variation of the following model:

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{S} + \boldsymbol{\beta}_2\mathbf{X} + \boldsymbol{\varepsilon}. \tag{5}$$

Several such models have been proposed including Muralidhar et al. (1999, 2001) and Franconi and Stander (2002). The latter authors propose an empirical model whose specific form is based on the characteristics of the data set being masked. By contrast, Muralidhar et al. (1999) originally proposed a model of the form as shown in (5), but with the requirements that the covariance matrix of the released data ($\mathbf{S}$ and $\mathbf{Y}$) be the same as that of ($\mathbf{S}$ and $\mathbf{X}$). This specification imposes a specific structure on the covariance of $\boldsymbol{\varepsilon}$. In order to improve the disclosure risk characteristics, Muralidhar et al. (2001) proposed a modified model of the form

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{S} + \boldsymbol{\varepsilon}, \tag{6}$$

where $\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{\mathbf{XS}}\boldsymbol{\Sigma}_{\mathbf{SS}}^{-1}$, $\boldsymbol{\beta}_0 = \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{XS}}\boldsymbol{\Sigma}_{\mathbf{SS}}^{-1}\boldsymbol{\mu}_{\mathbf{S}}$, and $\boldsymbol{\Sigma}_{\mathbf{ee}} = \left(\boldsymbol{\Sigma}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XS}}\boldsymbol{\Sigma}_{\mathbf{SS}}^{-1}\boldsymbol{\Sigma}_{\mathbf{SX}}\right)$, all of which are estimated using the original data. With these specifications, for large data sets, the mean vector and covariance matrix of the released data ($\mathbf{S}$ and $\mathbf{Y}$) will be the same as that of the original data ($\mathbf{S}$ and $\mathbf{X}$). However, there is some information loss in estimates of the covariance matrix, due to sampling error in smaller data sets. Since $\mathbf{Y}$ is generated as a function of $\mathbf{S}$ and $\boldsymbol{\varepsilon}$ this procedure also minimizes the risk of both identity and value disclosure.

3.3. Sufficiency Based Linear Models

An important variant of the linear model was suggested by Burridge (2003). In this approach, by appropriately generating the values of $\boldsymbol{\varepsilon}$, it is possible to ensure that the mean vector and covariance matrix of the released data are *identical* to that of the original data. Hence, for all statistical analyses for which the mean vector and covariance matrix are sufficient statistics, the results of the analysis using the masked data will yield *identical results* to that using the original data. That is, the (statistical) information loss will be *zero*. Note that for most traditional statistical analyses (including, but not limited to, comparison of means, ANOVA, regression analysis, and even such multivariate procedures such as canonical correlation analysis), the mean vector and covariance matrix serve as sufficient statistics. Hence, if this procedure is employed to mask the data, a user who analyzes the masked data will get exactly the same results as using the original unmasked data. In addition, this procedure also minimizes disclosure risk. A similar approach was also suggested by Ting et al. (2005) when the entire data set is confidential.

Muralidhar and Sarathy (2007) recently proposed a further modification of this linear model in equation (5) with the following restrictions:

$$\boldsymbol{\beta_0} = (\mathbf{1} - \boldsymbol{\beta_2})\boldsymbol{\mu_X} - \boldsymbol{\beta_1}\boldsymbol{\mu_S}, \tag{7}$$

$$\boldsymbol{\beta_1} = (\mathbf{I} - \boldsymbol{\beta_2})\boldsymbol{\Sigma_{XS}}\boldsymbol{\Sigma_{SS}^{-1}}, \text{ and} \tag{8}$$

$$\boldsymbol{\Sigma_{\varepsilon\varepsilon}} = \left(\boldsymbol{\Sigma_{XX}} - \boldsymbol{\Sigma_{XS}}\boldsymbol{\Sigma_{SS}^{-1}}\boldsymbol{\Sigma_{SX}}\right) - \boldsymbol{\beta_2}\left(\boldsymbol{\Sigma_{XX}} - \boldsymbol{\Sigma_{XS}}\boldsymbol{\Sigma_{SS}^{-1}}\boldsymbol{\Sigma_{SX}}\right)\boldsymbol{\beta_2^T}. \tag{9}$$

With the above specifications and appropriately selecting the value of $\boldsymbol{\beta_2}$ it is possible to ensure that the masked data (**Y** and **S**) has exactly the same mean vector and covariance matrix as the original data (**X** and **S**). That is, this approach preserves sufficient statistics underlying linear relationships. Consequently, there is zero information loss in estimating any of the linear relationships among the different variables. When $\boldsymbol{\beta_2}$ is zero, this model reduces to the model shown in equation (6). For non-zero $\boldsymbol{\beta_2}$ the resulting masked variables *do not* provide the lowest possible level of disclosure risk. However, these masked values may be more acceptable to some users who may have reservations using the more "synthetic" data generated from the model in equation (6).

While the sufficiency-based linear models (SBLM) procedure in equations (5, 7, 8, and 9) provides significant advantages over other perturbation procedures, it is not without problems. First and foremost, this procedure results in information loss in the marginal distribution of the masked variable (**Y**). The exact form of the distribution of **Y** would depend on the selection of the distribution for the error term $\boldsymbol{\varepsilon}$. However, unless **X** was normally distributed, the marginal distribution of **Y** will be different from that of **X**. One common problem that arises as a consequence is that the masked values may consist of negative values whereas the original data may be all positive. In addition, while this procedure maintains linear relationships among all variables, non-linear relationships are not preserved in the masked data. Thus, while this procedure is a *complete solution* to the masking problem when the joint distribution of (**X** and **S**) is multivariate normal, resulting in zero information loss and zero incremental disclosure risk, in other cases, it has some shortcomings.

3.4. Multiple Imputation

Originally proposed for missing data, multiple imputation was suggested as a possible alternative for masking data by Rubin (1993). Since then, several researchers have investigated the effectiveness of multiple imputation for masking numerical microdata. In its basic form, multiple imputation essentially uses a linear model as in equation (5). Using the available data, the intercept, slope coefficients, and error variance are estimated. Up to this point, both the linear models approach and multiple imputation are identical. In the traditional linear model approach, a data set would be generated using the estimated coefficients. In other words, the estimated coefficients are treated as population parameters and the only variability arises from the error variance. In multiple imputation, additional variability is introduced by treating the intercept and slope coefficients as sample statistics. Further, several sets of masked data are generated (perhaps as many as 100). Each set of imputed values are based on newly generated values of intercept, slope, and error variance. The user is required to analyze each imputed set and finally aggregate the results (Raghunathan et al. 2003). The effectiveness of the procedure improves when the number of imputed data sets is larger.

The difficulty of using multiple imputation in practice is obvious. Requiring users to analyze multiple data sets (as many as 100) and then aggregating the results imposes a significant burden on the users. Rubin (1987) has clearly indicated that, in

order for this procedure to be effective, it is necessary that the data be imputed multiple times. Furthermore, in order to ensure a high level of security, this procedure also requires that the data be imputed for those non-confidential values for which corresponding confidential values do not exist. In other words, the data set should consist of some data points for which both **S** and **X** have been observed and some for which only **S** has been observed. This may not be true for all data sets.

In terms of performance, multiple imputation behaves very similar to the linear model procedure suggested by Muralidhar et al. (1999). If the values of **Y** are generated independent of the values of **X**, then multiple imputation also minimizes disclosure risk. More importantly, if we use the modification suggested by Burridge (2003) to maintain sufficient statistics, the resulting masked data is actually superior to multiple imputation. By ensuring that the results of analyses using the masked data are the same as that using original data, the liner model procedure allows the user to analyze a single data set and get better results than what they would get by analyzing and aggregating multiple data sets (Muralidhar and Sarathy 2006a).

3.5. Micro-aggregation

Micro-aggregation is often suggested as an attractive procedure for data masking because of its simplicity (Domingo-Ferrer et al. 2002). In its simplest form, micro-aggregation works as follows. A set of k observations are identified as the "closest" observations. The values of the confidential variables for these observations are aggregated. The aggregated values are released in place of the original values. The selection of the "closest" observations can be performed on a variable by variable basis (univariate micro-aggregation) or for multiple variables (multivariate micro-aggregation). While micro-aggregation may be easy to implement, as we discuss below, its performance characteristics are very poor.

In terms of security, univariate micro-aggregation has a very high risk of both identity and value disclosure (Winkler 2002). Proponents of multivariate micro-aggregation often contend that it satisfies k-anonymity (that is, there are at least k observations that have exactly the same value). This may provide security against identity disclosure, but not against value disclosure. Consider the situation where the values of all k observations are close together. Then, releasing the aggregated value is, for all practical purposes, the same as releasing the original value. Releasing the aggregated value will enable an intruder to predict the original values with a great deal of accuracy. That there are k observations that are similar is moot since the intruder is able to estimate the values of the confidential variables for all k observations.

The information loss resulting from micro-aggregation can also be very high (Muralidhar and Sarathy 2006b). We can show theoretically that there is attenuation in the variance of the masked variables compared to the original variables. We can show that this attenuation in variance may be as high as 50%. The marginal distribution of the masked variables is different from those of the original variables. Due to the attenuation in variance, the correlation between the masked variables is higher than that of the original variables (Muralidhar and Sarathy 2006b).

3.6. Data Swapping

Like micro-aggregation, data swapping is often proposed as an effective masking technique because of its simplicity. Originally proposed for categorical variables, data swapping has since been adopted for numeric variables (Moore 1996). In data swapping, values of a particular variable within a specified proximity are exchanged. The process is repeated for every observation and every variable. The resulting masked data set retains the same marginal distribution as the original confidential variables. However, data swapping results in attenuation of product moment correlation as well as very high disclosure risk. It is also evident that, since the values of the variables are swapped randomly between a specified proximity, it will result in attenuation of the rank order correlation as well. Like micro-aggregation, it is difficult to theoretically specify the extent to which relationships among variables are affected by data swapping. An empirical assessment is provided by Moore (1996) and Muralidhar and Sarathy (2006c). Fienberg and McIntyre (2005) provide an excellent discussion of the variations of the data swapping procedure.

3.7. Data Shuffling

Data Shuffling is a new patented procedure (US Patent # 7200757) developed by Muralidhar and Sarathy (2006d). It is a hybrid procedure where the original variables are first perturbed using the copula based perturbation approach (Sarathy et al. 2002). The resulting perturbed values are then reverse-mapped on to the original values, resulting in the shuffled data set. Superficially, data shuffling can be considered to be a multivariate version of data swapping since it is performed on the entire data set rather than on a variable by variable basis. For a complete description of data shuffling, please refer to Muralidhar and Sarathy (2006d). Data shuffling is a more general version of the LHS procedure suggested by Dandekar et al. (2002).

Data shuffling has the following desirable properties. First and foremost, the perturbed values are generated independent of **X** (given **S**) and hence have no incremental disclosure risk. Second, like data swapping, the shuffled values are actually the original values of the confidential variables assigned to a different observation. Hence, the marginal distribution of the masked data is identical to the marginal distribution of the original data. Third, the use of the copula-based perturbation approach enables data shuffling to maintain the rank order correlation of the masked data to be the same as that of the original data. This implies that data shuffling results in minimal information loss in linear and monotonic non-linear relationships among variables. It does not maintain non-monotonic non-linear relationships.

3.8. Summary of Comparison of Masking Techniques

In this section, we provide a brief summary of the techniques and their capabilities. It is important to note that, among these techniques, it is possible to assess, theoretically, the disclosure risk and information loss characteristics of all the techniques *except* micro-aggregation and data swapping. Ironically, these are also the techniques that are often employed in practice because they are "easy to implement", and not because they perform well. We can theoretically show that micro-aggregation does not maintain the marginal distribution, attenuates the variance, and accentuates correlation. Similarly, data swapping results in correlation attenuation even for linear relationships. Give these results, it would be reasonable to conclude that both micro-aggregation and data swapping are unlikely to maintain non-linear relationships. Yet, proponents of micro-aggregation and data swapping often defend the techniques under the claim that it *may* preserve relationships better. There is no theoretical or empirical evidence to suggest that these claims are true. Table 1 provides a summary of the performance characteristics.

Table 1. Comparison of Masking Techniques

Method	Disclosure Risk Minimized?	Information Loss				
		Are Marginal distributions maintained?	Are Sufficient statistics maintained?	Are Linear Relationships Maintained?	Are Monotonic relationships maintained?	Are Non-monotonic relationships maintained?
Simple Noise Addition	No	No	No	No	No	No
Simple Noise Addition (Kim's method)	No	No	No	No	No	No
Simple Noise Addition (Tendick and Matloff's method)	No	No	No	No	No	No
General linear model (equation 5)	No	No	No	Yes	No	No
General linear model (equation 6)	Yes	No	No	Yes	No	No
Sufficiency Based Linear Models	Yes	No	Yes	Yes	No	No
Multiple Imputation	Yes	No	No	Yes	No	No
Micro-aggregation	No	No	No	No	No	No
Data Swapping	No	Yes	No	No	No	No
Data Shuffling	Yes	Yes	No	Yes	Yes	No

From Table 1, it is easy to see that among methods where additive noise is employed (simple noise addition, linear models, multiple imputation), the sufficiency-based linear models approach that maintains sufficient statistics provides superior performance to the other techniques in this class. Among the other models (micro-aggregation, data swapping, and data shuffling), data shuffling provides better performance characteristics (lower disclosure risk and information loss) than the other techniques. For these reasons, we provide an empirical illustration of the application of these two techniques. In addition to evaluating their performance on the overall data set, we also address the issue of sub-group characteristics.

4. Performance for subsets of data

One key aspect of the characteristics of the masking techniques that we investigate in this study is their ability to maintain sub-group characteristics. For numerical variables, it is possible to generate an infinite number of possible sub-groups and it becomes difficult to evaluate all possible sub-groups. However, when there are categorical non-confidential variables, there are usually a finite number of sub-groups that are created by the intersection of these categorical variables. Furthermore, data consisting of both categorical and numerical variables are very common in practice. Hence, we evaluate the performance of the two selected techniques (sufficiency-based linear models and data shuffling) on sub-groups as well. For each sub-group, we evaluate the extent of information loss from the masked data. No such evaluation is necessary for disclosure risk since the masked values are independent of the original data set in each sub-group. However, the benchmark disclosure risk for each sub-group will be different and will be a function of the relationships between the confidential and non-confidential variables in the sub-group and the size of the sub-group.

5. Empirical Assessment

We performed an empirical assessment of the two masking techniques using two data sets. The first masking technique used was data shuffling that does not require any parameter specifications. The second masking technique was the SBLM procedure with the requirement that $\boldsymbol{\beta_2}$ be a diagonal matrix with the value d ($0 \leq d \leq 1$) in the diagonal and 0 in the off-diagonal terms. This simple specification implies that when d = 0, the resulting model is the one shown in equation (6) and when d = 1, the entire data set is released unmodified. Thus, the selection of d directly influences the extent to which the original values are used in the masking. Note that when d > 0, this method does not provide minimum security.

5.1. Simulated Data Set

We used two data sets in this empirical assessment. The first data set was simulated and consisted of 50000 observations. The data consisted of 3 categorical non-confidential variables Gender (male or female), Marital Status (married or other), and Age group (1 to 6). The 3 confidential numerical variables (Home value, Mortgage balance, Total net value of assets) were generated using the NORTA approach for generating related multivariate non-normal variables. Of the three confidential variables, two (Home value and Mortgage balance) had non-normal marginal distributions, while the third had a normal distribution. The relationship between the last two variables was linear while the other relationships were non-linear. Twenty four sub-groups were formed as a combination of the Gender × Martial status × Age group. Data shuffling was applied to the entire data set. In addition, 3 different levels of masking were applied for linear model approach (d = 0.00, 0.50, 0.90). As indicated earlier, when d = 0.00, given the non-confidential variables, the perturbed variables are independent of the original variables and are sometimes considered synthetic data.

5.1.1. Assessment of Disclosure Risk As indicated earlier, the first step in the assessment of the masking techniques was to compute the risk of identity disclosure. Table 2 provides the results of the identity disclosure assessment performed using the procedure suggested by Fuller (1993). There are many approaches for assessing identity disclosure and we could use any one of these procedures. However, the primary objective of this assessment is to compare the different methods rather than assess the extent of disclosure. While the specific results of using another procedure for assessing identity disclosure may be different, the relative performance of the different methods will be the same. Table 2 provides, for each sub-group defined by the categorical variables (a total of 24 sub-groups), the number of observations in each sub-group and the number of observations that were re-identified. As indicated earlier, when shuffling and perturbation with d = 0.00 are used to mask the variables, within a given sub-group, the original and masked variables are independent. Hence, the probability of re-identification within a sub-group is $(1/n_k)$ where n_k is the size of the sub-group. The results in Table 2 clearly show that this is indeed the case. The probability of re-identification is much higher for the other perturbed values, with the higher re-identification occurring when d = 0.90. Thus, in terms of disclosure risk, it is easy to see that the data shuffling and perturbation with d = 0.00 provide the best results, with re-identification occurring by chance alone.

It is also easy to assess the risk of value disclosure. As indicated earlier, for a given sub-group, the shuffled data and perturbed data with d = 0.00 are independent of the original data. This implies that the covariance between the original and masked data are close to zero for shuffled data and exactly 0.00 for the perturbed data with d = 0.00. Hence, the correlation between the original and masked data for these two methods will be 0.00, resulting in no predictive ability. By contrast, for the other two approaches, the correlation between the original and masked variables will be d and the intruder would be able to explain d^2 proportion of the variability in the values of the original variables using the masked variables.

Table 2. Risk of Identity Disclosure for Simulated Data

Gender	Marital	Age	Total number of observations	Number of Observations Identified			
				Shuffled	Perturbed (0.00)	Perturbed (0.50)	Perturbed (0.90)
0	0	1	1220	3	1	5	40
		2	1181	0	1	13	47
		3	1193	1	1	8	42
		4	1162	3	1	4	39
		5	1159	2	1	5	29
		6	1181	0	1	4	42
	1	1	4672	2	1	7	56
		2	4723	0	1	12	73
		3	4671	1	1	9	54
		4	4719	2	1	5	48
		5	4635	1	1	6	61
		6	4650	2	1	7	58
1	0	1	515	2	1	5	25
		2	468	2	1	6	33
		3	502	3	1	1	30
		4	511	0	1	3	24
		5	503	2	1	2	34
		6	464	0	1	2	21
	1	1	2019	2	1	7	59
		2	1968	0	1	6	43
		3	2044	0	1	3	49
		4	1940	0	1	4	35
		5	1960	1	1	5	52
		6	1940	0	1	4	50

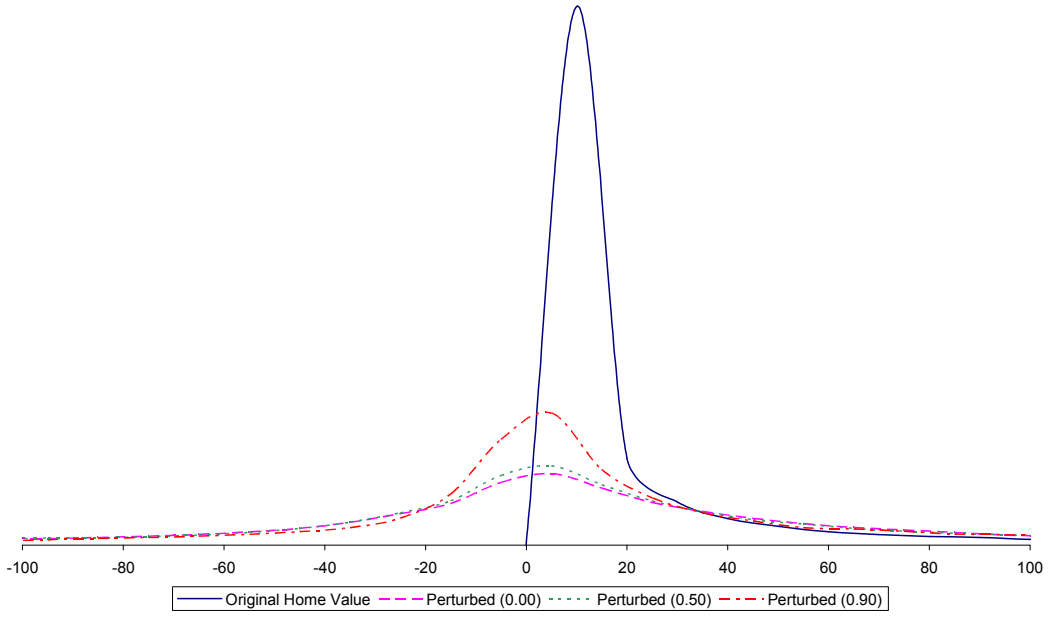
As an illustration, consider the sub-group Gender = 0, Marital = 0, and Age = 1. The mean and standard deviation of the Home value variable in this sub-group are 2.872 and 8.643, respectively. With only this information, for any observation in this sub-set, the best prediction of a 99% interval estimate of the true value of the Home value variable would have an interval of approximately (3 × 8.643). Now assume that the shuffled data is released. The correlation between the original and the shuffled home values is 0.03. Hence, if we perform regression analysis to predict the original value of the confidential variable using the shuffled values, the resulting R^2 would be 0.0009 resulting in a standard error of 8.642. Using this information, a simple 99% confidence interval would have an interval of approximately (3 × 8.642), which for all practical purposes is almost exactly the same as the interval constructed without access to the masked data. In other words, releasing the shuffled data does not allow the intruder to estimate the value of the confidential variable with any greater level of security. Similar results will be observed for the perturbed data when d = 0.00.

The above result does not hold for the other two perturbation parameters (d = 0.50, 0.90). When d = 0.50, if we perform a regression analysis to predict the original Home value variable using the perturbed values, the resulting standard error is 7.485. A 99% confidence interval estimate would have an interval of approximately (3 × 7.485). This implies that the intruder is able to gain a more accurate estimate compared to not having the perturbed values. When d = 0.90, the resulting standard error from the regression analysis is 3.767. If we construct a 99% confidence interval using this information, it results in an interval of approximately (3 × 3.767). Compared to the original interval, the width of this interval is less than 50% of the original width. This allows the intruder to gain far more accurate estimate of the value of the confidential variable.

Thus, an intruder would have a much better estimate of the original values when the data is masked using the perturbation approach with d = 0.50 and 0.90. In conclusion, when considering disclosure risk, because of their inherent property of conditional independence, data shuffling and perturbation with d = 0.00 perform better than perturbation with d = 0.50 and 0.90. If disclosure risk were the only criterion, data shuffling and perturbation with d = 0.00 would be the preferred methods.

5.1.2. Assessment of Information Loss In assessing information loss, the first step was to assess the extent to which the marginal distribution of the (entire) masked data set resembles the original data set. We know that, for the entire data set, *data shuffling maintains the marginal distribution of the masked variables to be exactly the same as that of the original variables*. By contrast, the SBLM approach is capable of maintaining the marginal distribution of the variables only when the variable has a normal distribution. In this example, Home value and Mortgage balance were non-normally distributed, while the Net assets variable was normally distributed. Figure 1 shows the marginal distribution of the original Home value variable and the variable masked using the 3 linear perturbation models. We did not include the shuffled data since the marginal distribution of the shuffled data will coincide exactly with the original data. As expected, the marginal distributions of the original and shuffled variables are identical. Also as expected, the marginal distribution of the variable masked using the linear model with d = 0 differs most from the original variable. It is easy to see that greater the influence of **X** (greater the value of d), the closer the marginal distribution of the masked data to the original data. One of the problems with the SBLM approach is that while the original variables are always positive, the masked variables may have negative values. Figure 2 provides the marginal distribution of the Mortgage balance variable. The results are similar to those observed for Home value variable. Finally, Figure 3 provides the marginal distribution of the original and 3 perturbed data sets. In this case, since the marginal distribution of Net Asset value was normal, all the masked variables maintain the marginal distribution to be the same as the original variable.

Figure 1. Marginal Distribution of Home Value



One of the attractive features of the data shuffling procedures is that *the marginal distribution of the shuffled data within any sub-group defined by the non-confidential categorical variables is exactly the same as that of the original variable*. This is not true for the other approaches. The marginal distribution of the perturbed data are different from that of the original data for sub-groups as well. To illustrate this, consider the case for the sub-group where Gender = 0, Marital Status = 0, and Age = 1. Figure 4 provides the marginal distribution of the original and the 3 perturbed data sets for the Home value variable. Again, we do not provide the shuffled data since it will coincide exactly with the original data. As can be seen from this example, the marginal distribution of the perturbed data differs considerably from the original data even when d = 0.90. Thus, the "addition of noise" results in a marginal distribution that is closer to normality than the original data. Note that, for the Net Assets variable, the marginal distribution of all the masked variables for all the sub-groups will be similar since the original variable was normally distributed.

Figure 2. Marginal Distribution of Mortgage Balance

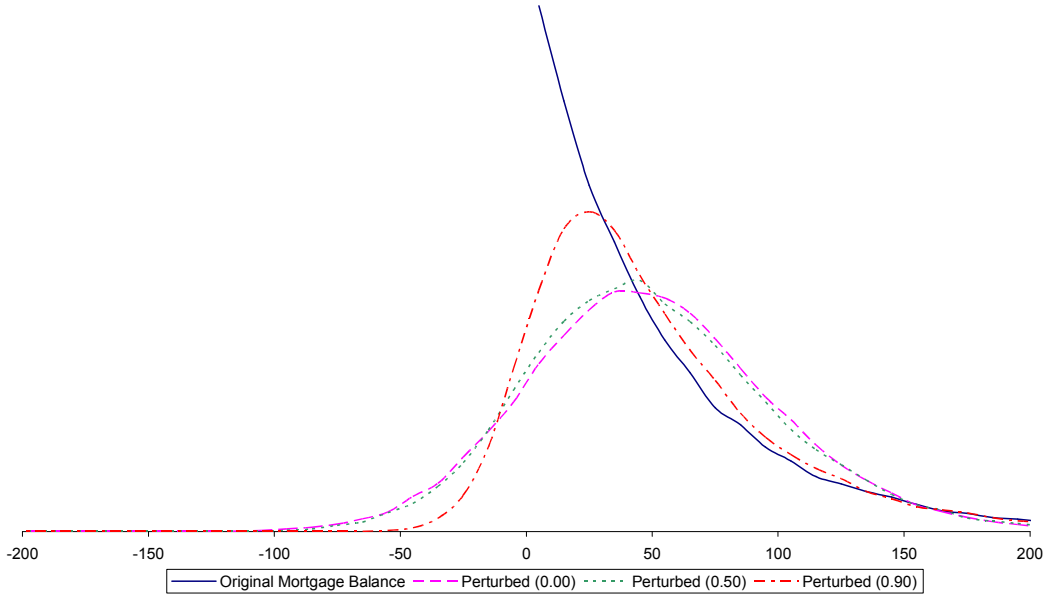
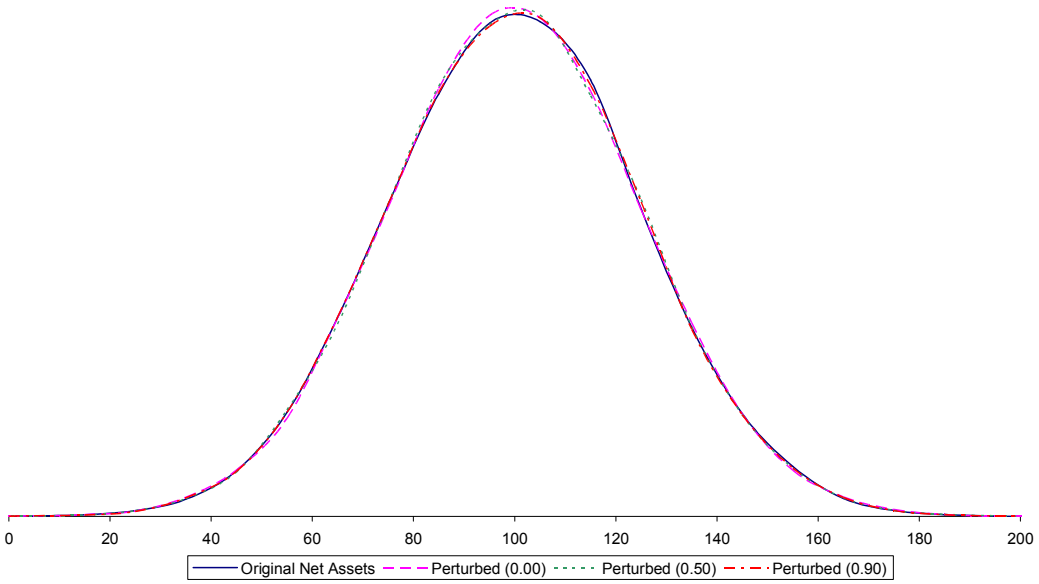
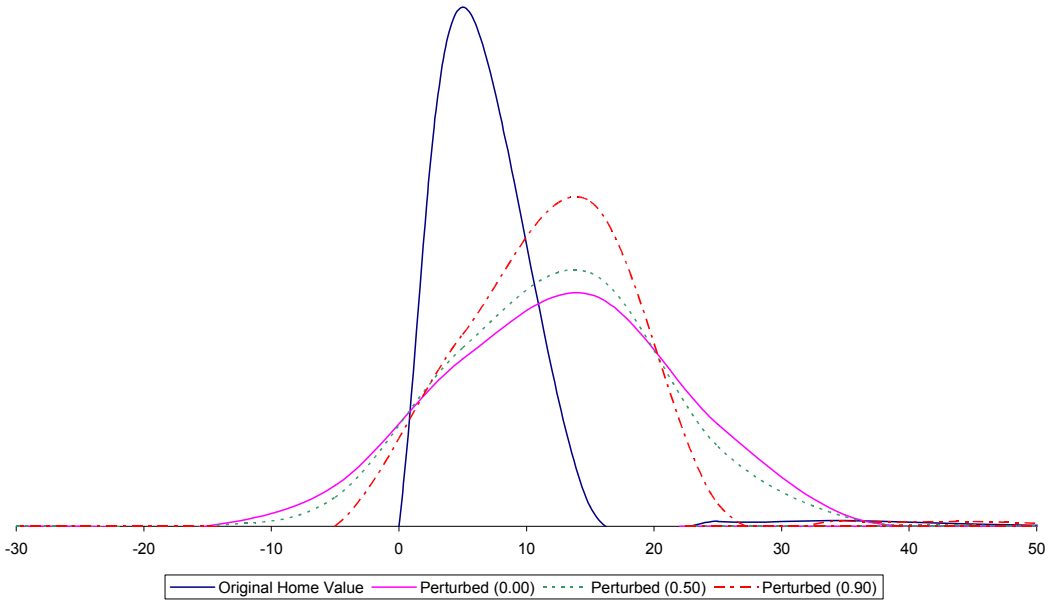


Figure 3. Marginal Distribution of Net Asset Value



As discussed earlier, one other attractive feature of all the methods considered in this study is that the mean and variance of all the variables in every sub-group defined by the non-confidential categorical variables will be exactly the same as that of the original data. Hence, we see no reason to provide this data. However, in addition to maintaining the mean and variance, the *shuffled data maintains all the univariate marginal characteristics of the masked data to be the same as that of the original data.*

Figure 4. Marginal Distribution for Original and Perturbed Home values in a Sub-Group (Gender = 0, Marital = 0, and Age = 1)



To assess the extent to which the methods maintain relationships among variables, we computed the product moment correlation between the variables in the entire data set as well as in each sub-group. The results of this analysis are provided in Table 3. As expected, *the product moment correlations of the original data and those of the perturbed data are exactly the same for the data set as a whole and for every sub-group*. The shuffled data does not provide exactly the same results, but *the product moment correlations of the shuffled data and those of the original data are very similar for the data set as a whole and for every sub-group*. Thus, in terms of maintaining product moment correlation, the SBLM approach seems to perform better than the shuffling approach. This is expected since the SBLM approach is intended to maintain first and second order moments (and consequently correlation) among the variables. However, this does not necessarily mean that it is superior to data shuffling as the following discussion shows.

Consider the relationship between the variables Mortgage balance and Net asset value. Figure 5 provides a scatter plot of the original values with Net asset values on the X-axis and Mortgage balance on Y-axis. It is clear from this figure that the relationship between the two variables is non-linear. In cases where the relationship is non-linear, product moment correlation which measures only the linear relationship is not an appropriate measure. The product moment correlation for these two variables in the data set is 0.719 and all three perturbed values maintain this correlation. By contrast, the correlation between the corresponding shuffled variables slightly different (0.718).

Now consider a plot of the perturbed values of Mortgage balance and Net asset values (with d = 0.00) overlaid on top of the original scatter plot (Figure 6). Figure 6 clearly indicates that the perturbation approach has considerably modified the relationship between the variables; the original relationship was non-linear while the perturbed data is almost linear. A plot of the shuffled values of Mortgage balance and Net asset values overlaid on the original data is shown in Figure 7. This figure clearly indicates that the shuffled data maintains the (monotonic) non-linear relationship between the two variables better than the perturbed data. Thus, although the SBLM approach maintains the product moment correlation exactly, it does not necessarily maintain non-linear relationships between the variables. By contrast, while the shuffled data does not maintain the product moment correlation exactly, it is capable of maintaining monotonic non-linear correlations much better than the perturbed data.

Table 3. Product Moment Correlation for Simulated Data

Sub-Group	Correlation between Home Value and Mortgage Balance					Correlation between Home Value and Net Assets					Correlation between Mortgage Balance and Net Assets				
	O	S	P1	P2	P3	O	S	P1	P2	P3	O	S	P1	P2	P3
1	0.338	0.363	0.338	0.338	0.338	0.373	0.328	0.373	0.373	0.373	0.693	0.701	0.693	0.693	0.693
2	0.216	0.245	0.216	0.216	0.216	0.214	0.274	0.214	0.214	0.214	0.700	0.707	0.700	0.700	0.700
3	0.402	0.306	0.402	0.402	0.402	0.375	0.319	0.375	0.375	0.375	0.707	0.702	0.707	0.707	0.707
4	0.294	0.338	0.294	0.294	0.294	0.282	0.277	0.282	0.282	0.282	0.705	0.698	0.705	0.705	0.705
5	0.201	0.251	0.201	0.201	0.201	0.250	0.267	0.250	0.250	0.250	0.707	0.716	0.707	0.707	0.707
6	0.320	0.355	0.320	0.320	0.320	0.337	0.344	0.337	0.337	0.337	0.746	0.755	0.746	0.746	0.746
7	0.266	0.229	0.266	0.266	0.266	0.230	0.222	0.230	0.230	0.230	0.698	0.695	0.698	0.698	0.698
8	0.313	0.318	0.313	0.313	0.313	0.281	0.292	0.281	0.281	0.281	0.695	0.705	0.695	0.695	0.695
9	0.276	0.253	0.276	0.276	0.276	0.264	0.264	0.264	0.264	0.264	0.694	0.697	0.694	0.694	0.694
10	0.195	0.210	0.195	0.195	0.195	0.179	0.196	0.179	0.179	0.179	0.708	0.708	0.708	0.708	0.708
11	0.284	0.285	0.284	0.284	0.284	0.274	0.261	0.274	0.274	0.274	0.710	0.700	0.710	0.710	0.710
12	0.259	0.250	0.259	0.259	0.259	0.262	0.243	0.262	0.262	0.262	0.728	0.723	0.728	0.728	0.728
13	0.288	0.243	0.288	0.288	0.288	0.244	0.256	0.244	0.244	0.244	0.698	0.712	0.698	0.698	0.698
14	0.321	0.294	0.321	0.321	0.321	0.310	0.351	0.310	0.310	0.310	0.718	0.730	0.718	0.718	0.718
15	0.356	0.371	0.356	0.356	0.356	0.364	0.376	0.364	0.364	0.364	0.705	0.751	0.705	0.705	0.705
16	0.386	0.354	0.386	0.386	0.386	0.329	0.325	0.329	0.329	0.329	0.694	0.664	0.694	0.694	0.694
17	0.387	0.352	0.387	0.387	0.387	0.393	0.418	0.393	0.393	0.393	0.707	0.714	0.707	0.707	0.707
18	0.195	0.208	0.195	0.195	0.195	0.193	0.228	0.193	0.193	0.193	0.737	0.692	0.737	0.737	0.737
19	0.320	0.389	0.320	0.320	0.320	0.338	0.356	0.338	0.338	0.338	0.676	0.662	0.676	0.676	0.676
20	0.349	0.264	0.349	0.349	0.349	0.288	0.259	0.288	0.288	0.288	0.692	0.663	0.692	0.692	0.692
21	0.357	0.303	0.357	0.357	0.357	0.348	0.318	0.348	0.348	0.348	0.703	0.714	0.703	0.703	0.703
22	0.239	0.236	0.239	0.239	0.239	0.216	0.227	0.216	0.216	0.216	0.701	0.710	0.701	0.701	0.701
23	0.289	0.328	0.289	0.289	0.289	0.272	0.292	0.272	0.272	0.272	0.707	0.717	0.707	0.707	0.707
24	0.307	0.263	0.307	0.307	0.307	0.275	0.253	0.275	0.275	0.275	0.721	0.727	0.721	0.721	0.721
Entire Data	0.223	0.220	0.223	0.223	0.223	0.201	0.197	0.201	0.201	0.201	0.719	0.718	0.719	0.719	0.719

Legend: O = Original Data; S = Shuffled Data; P1 = Perturbed (0.00); P2 = Perturbed (0.50); P3 = Perturbed (0.90)

Figure 5. Scatter plot of Net Asset Value and Mortgage Balance (Original)

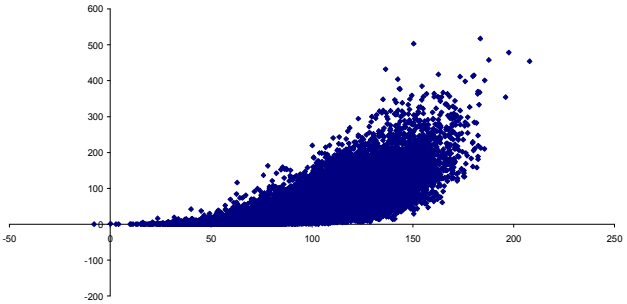


Figure 6. Scatter plot of Net Asset Value and Mortgage Balance (Original and Perturbed)

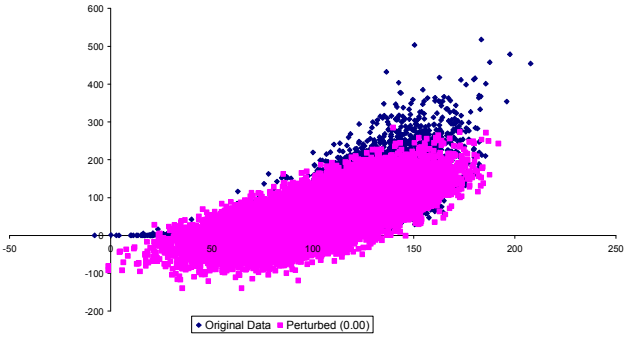
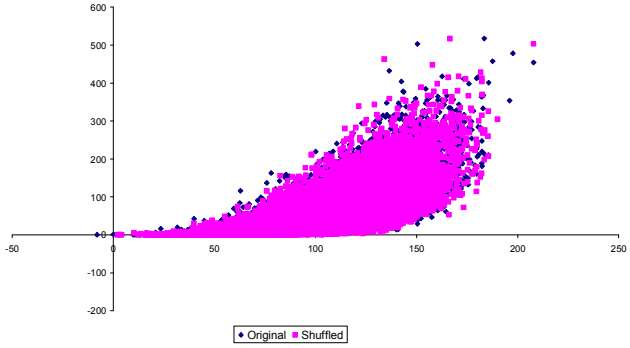


Figure 7. Scatter Plot of Net Asset Value and Mortgage Balance (Original and Shuffled)



In situations where the relationship is non-linear, in place of product moment correlation, rank order correlation is used to measure the strength of the relationship. The results provided in Table 4 indicate that, in general, the shuffled data maintain rank order correlation better than the perturbed data. This is to be expected since the shuffling procedure attempts to maintain all monotonic relationships, while the SBLM approach only deals with linear relationships. Note that the shuffling procedure is able to maintain the rank order correlation of the masked data to be very close to that of the original data both at the overall and sub-group level. This is a significant advantage of the shuffling approach over the SBLM approach. It is also important to note that *any approach based on linear models* (simple additive noise, Kim's method, multiple imputation, among others) are susceptible to the same "linearization" of non-linear relationships. Currently, only data shuffling offers the ability to maintain monotonic relationships among variables.

5.2. Census Data

In the previous example, we used a simulated data to highlight the strengths and weaknesses of the two procedures. In this section, we illustrate the applicability of the two procedures to any data set by considering the often used "Census Data". The original Census Data consists of 13 variables and 1080 observations. Of the 13 variables, the variable called PEARNVAL (Total personal earnings) equals PTOTVAL (Personal total income) – POTHVAL (Total other person's income). Hence, rather than using all 3 variables, we only used PEARNVAL in the analysis. Since all 13 of the variables were numerical, in order to illustrate the performance of these procedures for sub-groups, we converted 3 variables (AFLNWGT – Final weight, EMCOMTRB – Employer contribution, and PEARNVAL – Total personal earnings) to categorical variables. For each observation, if the value of each of these variables was less than the average for the entire data set, the value of the corresponding categorical variable was specified as 0 otherwise as 1. This resulted in a total of 8 possible combinations (sub-groups). We used shuffling and perturbation (d = 0.00, 0.50, and 0.90) to mask the data.

5.2.1. Assessment of Disclosure Risk As before, we assessed identity disclosure risk using the procedure described in Fuller (1993). The results of this assessment are provided in Table 5. As with the simulated data set, it is easy to see that the shuffled data and perturbed data (d = 0.00) provide the lowest risk of identity disclosure, with just one or two records being identified in each sub-group. The other two perturbed data sets do not fare quite as well. Using the perturbed data with d = 0.50, an intruder could identify a greater proportion of individuals in each sub-group. With the perturbed data with d = 0.90, the level of identity disclosure is extremely high.

In terms of value disclosure, the width of the confidence interval estimate for the perturbed data with d = 0.00 is exactly 100% of the original width. For the shuffled data, the width of the confidence interval is very close to 100% of the original data. For perturbed data with d = 0.50, the width of the confidence interval is 86.6% $[(1 - 0.5^2)^{0.5}]$ of the width of the original interval. The width of the confidence interval for the perturbed data with d = 0.90 is only 43.6% $[(1 - 0.9^2)^{0.5}]$ of the original width. Thus, the shuffled data and perturbed data with d = 0.00 minimize the risk of value disclosure. The perturbed data with d = 0.50 results in value disclosure which may be considered acceptable. The value disclosure risk resulting from the perturbed data with d = 0.90 is very high and allows the intruder to estimate the values of the confidential variables with much greater accuracy than without access to the data.

Table 4. Rank Order Correlation for Simulated Data

Sub-Group	Correlation between Home Value and Mortgage Balance					Correlation between Home Value and Net Assets					Correlation between Mortgage Balance and Net Assets				
	O	S	P1	P2	P3	O	S	P1	P2	P3	O	S	P1	P2	P3
1	0.553	0.575	0.329	0.340	0.365	0.624	0.648	0.368	0.368	0.376	0.762	0.786	0.680	0.679	0.708
2	0.527	0.535	0.211	0.216	0.234	0.608	0.625	0.214	0.224	0.259	0.772	0.768	0.680	0.682	0.722
3	0.536	0.539	0.371	0.379	0.388	0.617	0.605	0.357	0.360	0.398	0.764	0.743	0.692	0.696	0.731
4	0.516	0.535	0.275	0.282	0.290	0.603	0.615	0.255	0.255	0.291	0.741	0.744	0.688	0.689	0.718
5	0.540	0.554	0.201	0.205	0.255	0.622	0.647	0.237	0.240	0.285	0.763	0.757	0.692	0.691	0.723
6	0.566	0.566	0.301	0.305	0.315	0.646	0.663	0.322	0.316	0.336	0.792	0.792	0.735	0.737	0.770
7	0.529	0.531	0.252	0.258	0.266	0.623	0.613	0.216	0.224	0.247	0.761	0.767	0.674	0.682	0.707
8	0.523	0.503	0.296	0.300	0.307	0.604	0.591	0.266	0.272	0.297	0.768	0.770	0.677	0.684	0.718
9	0.535	0.532	0.264	0.268	0.282	0.608	0.605	0.253	0.257	0.292	0.759	0.761	0.676	0.679	0.714
10	0.538	0.525	0.183	0.195	0.220	0.613	0.612	0.172	0.183	0.224	0.761	0.759	0.693	0.693	0.724
11	0.539	0.540	0.279	0.285	0.301	0.619	0.621	0.263	0.267	0.297	0.752	0.756	0.692	0.699	0.723
12	0.538	0.532	0.242	0.247	0.273	0.623	0.632	0.245	0.256	0.291	0.768	0.764	0.719	0.718	0.741
13	0.590	0.606	0.287	0.302	0.330	0.669	0.672	0.257	0.273	0.315	0.798	0.816	0.685	0.683	0.705
14	0.510	0.550	0.304	0.310	0.310	0.640	0.652	0.295	0.305	0.321	0.764	0.788	0.704	0.698	0.723
15	0.539	0.552	0.351	0.339	0.342	0.633	0.617	0.349	0.336	0.357	0.752	0.783	0.706	0.689	0.712
16	0.560	0.564	0.378	0.370	0.366	0.598	0.631	0.337	0.317	0.322	0.732	0.742	0.670	0.682	0.692
17	0.523	0.530	0.374	0.365	0.370	0.597	0.603	0.410	0.399	0.412	0.754	0.782	0.678	0.677	0.707
18	0.569	0.542	0.161	0.168	0.217	0.650	0.663	0.170	0.175	0.244	0.763	0.732	0.710	0.713	0.747
19	0.536	0.562	0.314	0.328	0.353	0.638	0.648	0.316	0.332	0.376	0.762	0.758	0.656	0.665	0.697
20	0.538	0.530	0.332	0.330	0.323	0.622	0.607	0.273	0.273	0.284	0.755	0.751	0.676	0.680	0.698
21	0.521	0.525	0.351	0.358	0.364	0.601	0.609	0.328	0.341	0.365	0.759	0.761	0.680	0.685	0.719
22	0.553	0.573	0.228	0.233	0.251	0.638	0.641	0.223	0.226	0.261	0.760	0.766	0.688	0.696	0.727
23	0.521	0.541	0.279	0.291	0.314	0.598	0.610	0.262	0.274	0.316	0.762	0.773	0.696	0.697	0.715
24	0.554	0.566	0.297	0.300	0.294	0.628	0.637	0.261	0.272	0.292	0.767	0.769	0.708	0.715	0.740
Entire Data	0.582	0.583	0.262	0.265	0.283	0.681	0.682	0.255	0.258	0.292	0.782	0.783	0.707	0.711	0.740

Legend: O = Original Data; S = Shuffled Data; P1 = Perturbed (0.00); P2 = Perturbed (0.50); P3 = Perturbed (0.90)

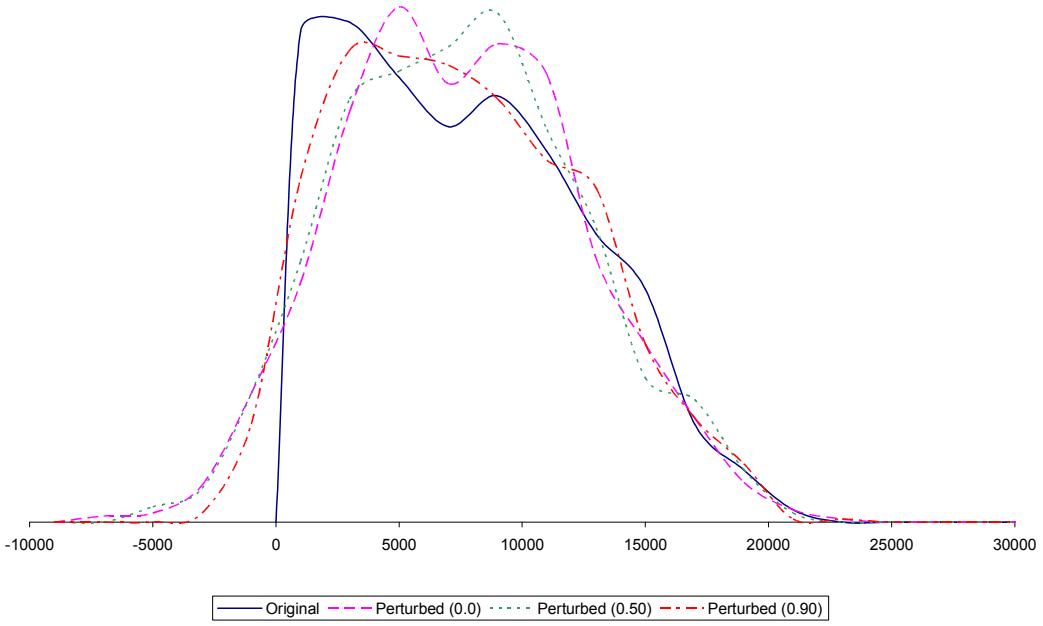
Table 5. Risk of Identity Disclosure for Census Data

Categorical Variable 1	Categorical Variable 2	Categorical Variable 3	Total number of observations	Number of Observations Identified			
				Shuffled	Perturbed (0.00)	Perturbed (0.50)	Perturbed (0.90)
0	0	0	156	4	1	8	83
		1	89	2	1	9	57
	1	0	57	4	1	5	35
		1	156	2	1	7	68
1	0	0	203	4	1	8	90
		1	103	4	1	13	47
	1	0	96	2	1	10	52
		1	220	3	1	10	82

5.2.2. Assessment of Information Loss As in the case of the simulated data, all masked data maintain the following important characteristics. *The mean and variance of the masked data are exactly the same as that of the original data for the entire data as well as for every sub-group*. In addition, *for the shuffled data, the marginal distributions of all the variables are exactly the same for the entire data set as well as for every sub-group*. This is not true for the perturbed data. As before,

the marginal distribution of the perturbed data are considerably different from the original data. Figure 8 shows the marginal distribution of the original FEDTAX variable along with the three perturbed values.

Figure 8. Marginal Distribution of FEDTAX Variable



The figure clearly shows that the marginal distribution of all three perturbed variables differs from the original variable. One problem with the SBLM approach is that while all the values of the original variable are zero or higher, the perturbed values are negative. Given the nature of the data set, it would be practically impossible not to have negative values when the data is perturbed.

Figure 9 shows the marginal distribution of the original and perturbed data for the INTVAL variable for the first sub-group (when the value of all the categorical variables is zero). The marginal distribution of the perturbed values are very different from the original values. As with the previous example, there are many negative values while the original variable does not consist of any negative values. While we have limited our discussion to this particular variable for one sub-group, this behavior is observed for practically all variables in all sub-groups. In our opinion, this is a significant problem with the perturbation approach. We also experimented with using alternative distributions for the noise term. The results however are similar to those observed in these cases.

One major advantage of the shuffling approach is that for all variables and all sub-groups, the shuffled data have exactly the same marginal distribution as the original variables. When the data is shuffled, users will be able to analyze individual variables within sub-groups without any information loss. SBLM at least maintains the mean and variance of the variables within the sub-groups. The other procedures (simple additive noise, Kim's method, multiple imputation, micro-aggregation, and swapping) do not typically maintain even mean and variance. Thus, from the perspective of univariate analysis of the masked data for the complete data set and sub-groups, data shuffling provides the best alternative among existing procedures.

As in the previous example, we analyzed both product moment and rank order correlation among the variables. In this case, with as many as 8 variables, there are a total of 21 different correlations to be considered for each of the 8 sub-groups and 4 methods. For the sake of brevity, we did not reproduce the entire set of results. Instead, Table 6 provides the product moment correlation of FICA (Social security deduction) and WSALVAL (Annual total wage and salary). We selected this particular example because of the fact that in one of the sub-groups, the correlation among the two variables is exactly 1.0. The results in Table 6 are similar to those observed for the simulated data. The perturbed data maintains the product moment correlation to be exactly the same for the overall data set and for each sub-group. The product moment correlation of the shuffled data, while very close to the original data, is not exactly the same.

Figure 9. Marginal Distribution of INTVAL Variable for Sub-Group 1

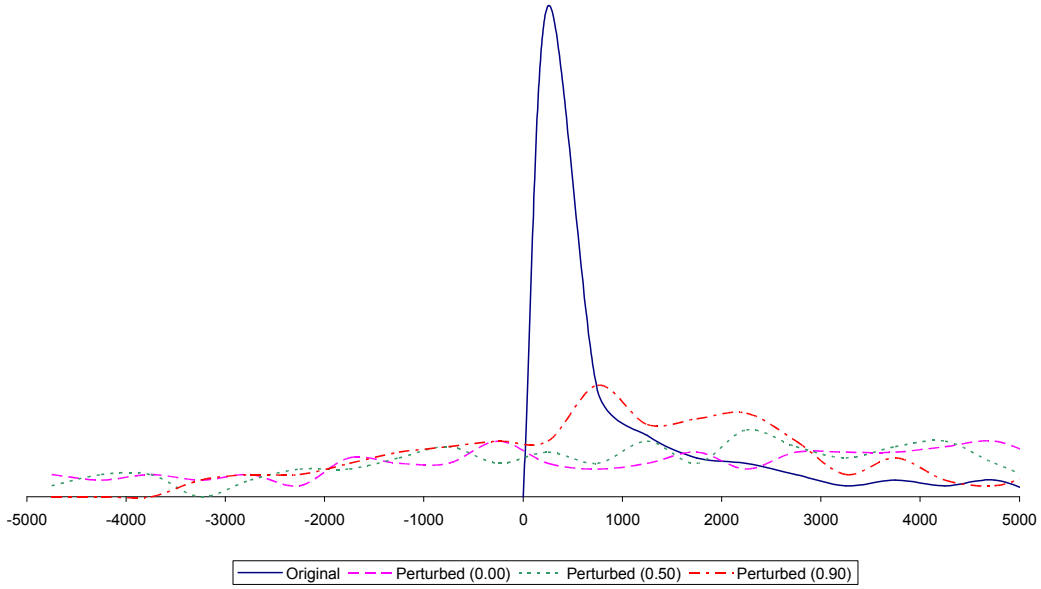


Table 6. Product Moment Correlation between FICA and WSALVAL

Group	Original	Shuffled	Perturbed (0.00)	Perturbed (0.50)	Perturbed (0.90)
1	0.642	0.800	0.642	0.642	0.642
2	1.000	1.000	1.000	1.000	1.000
3	0.817	0.899	0.817	0.817	0.817
4	0.863	0.915	0.863	0.863	0.863
5	0.529	0.734	0.529	0.529	0.529
6	0.988	0.971	0.988	0.988	0.988
7	0.766	0.885	0.766	0.766	0.766
8	0.929	0.943	0.929	0.929	0.929
All Observations	0.910	0.946	0.910	0.910	0.910

As observed earlier however, we do not believe that the product moment correlation is the best method for assessing the relationship among these variables. Figure 10 provides a scatter plot of the FICA and WSALVAL variables for the entire data set. The plot shows a clear non-linear relationship among the variables. Similar results were observed for practically all the variables. Hence, we computed the rank order correlation among the variables as an additional measure of information loss. The results are presented in Table 7.

The results in Table 7 clearly indicate that the shuffled data maintains the rank order correlation among these two variables better than the perturbed data for the overall data set as well as for practically every sub-group. Note that the data shuffling procedure is able to maintain the perfect correlation among the variables in sub-group 2 as does the perturbed data with d = 0.00. Thus, as expected, in addition to maintaining linear correlation, data shuffling performs better in maintaining non-linear relationships among variables while the perturbed data do not.

Figure 10. Scatter Plot of WSALVAL and FICA (Original Data)

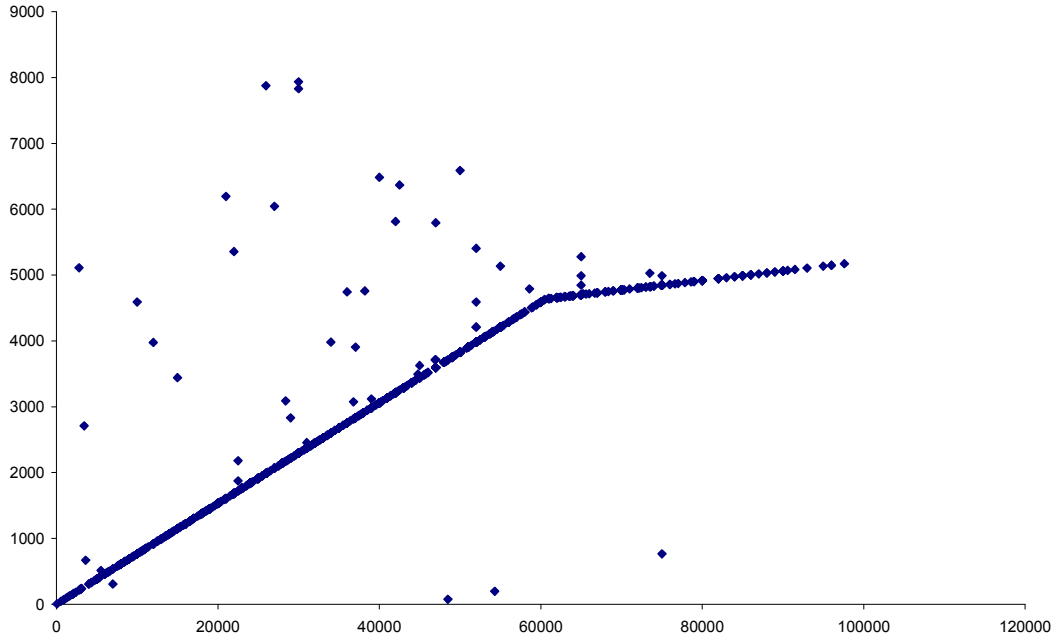


Table 7. Rank Order Correlation between FICA and WSALVAL

Group	Original	Shuffled	Perturbed (0.00)	Perturbed (0.50)	Perturbed (0.90)
1	0.857	0.858	0.597	0.642	0.770
2	1.000	1.000	1.000	0.955	1.000
3	0.876	0.930	0.821	0.779	0.857
4	0.938	0.930	0.834	0.914	0.913
5	0.807	0.787	0.502	0.472	0.653
6	0.975	0.977	0.986	0.879	0.985
7	0.923	0.945	0.737	0.654	0.846
8	0.965	0.948	0.932	0.922	0.954
All Observations	0.953	0.968	0.930	0.919	0.943

5.3. Summary of Results of the Empirical investigation

The results of the empirical investigation can be summarized as follows.

Data Shuffling:

(1) Disclosure risk is minimized,
(2) The marginal distribution of the shuffled data is exactly the same as that of the original data for the complete data set as well as for every sub-group, and
(3) The rank order correlation of the shuffled data is very similar to that of the original data for the complete data set as well as for every sub-group.

SBLM:

(1) Disclosure risk is minimized for the perturbed data set when d = 0, but not in the other cases.
(2) The mean vector and covariance matrix of the perturbed data is exactly the same as that of the original data for the complete data set as well as for every sub-group. However, the marginal distribution of the perturbed data is different from that of the original data,

(3) The product moment correlation of the perturbed data is exactly the same as that of the original data for the complete data set as well as for every sub-group. However, the rank order correlation of the perturbed data is very different from the original rank order correlation.

The selection of the specific approach would depend on the characteristics of the data. If the numerical data does not deviate significantly from normality and/or we are only interested in estimating linear relationships among variables, then the SBLM perturbation approach may be preferred since it offers the advantage that the results of traditional statistical analyses conducted on the masked data would yield *exactly* the same results as those using the original data. However, if the data is known to be non-normal and/or we are interested in estimating non-linear monotonic relationships, then shuffling would be preferred since it maintains the marginal distribution *exactly* and is also capable of maintaining monotonic non-linear relationships among variables. In practice, since data sets that exhibit multivariate normality are not very common, data shuffling would generally be the preferred approach.

6. Conclusions

In recent years there has been considerable research in the development of techniques for masking numerical data. The performance of these techniques both in terms of disclosure risk and information loss is better than those of most existing techniques. In this study, using both theoretical and empirical analyses, we evaluate the performance of two sets of techniques. In the empirical analysis, we use both simulated and real data. The results of the analyses clearly indicate that the shuffling and perturbation techniques put forth by this study offer several advantages in terms of minimizing information loss that none of the other techniques can match. In terms of disclosure risk, data shuffling and perturbation with d = 0.00 minimize the risk of both value and identity disclosure.

These techniques have however, received a lukewarm reception (if that) from most government agencies. The following other explanations are also often offered:

(1) "This is too complicated to work in practice" – In our opinion, this is an indefensible argument since in almost every case where we have heard this statement the techniques in question have never been tried. Our response to this explanation would be "How would you know if you never tried?"
(2) "Your technique is based on assumptions that will not be met in real data sets." – It is true that many of the newer techniques are model based. However, this does not mean that they will perform poorly when the assumptions are not met. More importantly, just because the simpler techniques are not model based does not automatically make them perform better. To paraphrase Fienberg et al. (1998) it is better to make certain assumptions regarding the data set since "we believe that this would be far preferable to throwing our hands up in despair or resorting to total ad hockery."
(3) "The procedures you have described will not maintain certain types of relationships. We think that micro-aggregation and data swapping will." – This is a very interesting statement which is made possible by the very fact that the newer techniques are theoretically based and allow us to make *a priori* statements regarding their performance. This very advantage is often used as a key to disallow their performance. Consider the following facts. Data shuffling maintains product moment and rank order correlation while data swapping results does not. What possible reason could there be to believe that data swapping will perform better than data shuffling when it comes to other types of relationships?
(4) "Users will not accept this procedure." – Another one of those arguments that cannot be refuted by individuals outside the agency because they do not have access to the data available with agencies. However, we believe that if the users are offered a choice between a complicated technique that offers better performance and a simpler one that offers poor performance, they will choose the complicated technique.

We agree that, compared to some procedures such as micro-aggregation and data swapping, these techniques may be perceived as substantially more difficult to implement. However, the degree of complexity is not much more than that of say multiple imputation. We would also like to believe that some of the resistance to accepting these procedures can be attributed to a lack of familiarity. We are hoping that this presentation will alleviate some of the familiarity problems.

When a government agency implements a less than optimal technique for masking data, the eventual losers are the individuals or organizations about whom data has been gathered and the users of the data. Settling for an “easy to implement” technique like micro-aggregation would result in a level of information loss that makes the data practically useless. At the same, the released data is susceptible to very high disclosure risk. By contrast, while the techniques recommended in this study are not perfect, they provide the users with certain assurance regarding the usefulness of the data. Simultaneously, they also provide the government agency with strong assurances against identity and value disclosure. It is our hope that we have provided a convincing argument for the use of these advanced techniques for masking numerical data.

References

1. Burridge, J. 2003. Information preserving statistical obfuscation. *Statistics and Computing*, 13 321-327.
2. Dandekar, R.A., M. Cohen, and N. Kirkendall 2002. Sensitive microdata protection using Latin Hypercube Sampling technique. In *Inference Control in Statistical Databases* (J. Domingo-Ferrer, Editor), Springer-Verlag, New York.
3. Domingo-Ferrer, J. and J. Mateo-Sanz. 2002. Practical data oriented micro-aggregation for disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14 189-201.
4. Domingo-Ferrer, J. and V. Torra, 2001a. Disclosure control methods and information loss for microdata," in Doyle, P., J.I. Lane, J.M. Theeuwes, and L.V. Zayatz (Editors), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier, Amsterdam, 91-110.
5. Fienberg, S. E., U. E. Makov, A. P. Steele. 1998b. Rejoinder, *Journal of Official Statistics*, 14 509-511.
6. Fienberg, S.E. and J. McIntyre. 2005. Data swapping: Variations on a theme by Dalenius and Reiss. *Journal of Official Statistics*, 21 309-323.
7. Franconi, L. and J. Stander (2002). A model based for disclosure limitation of business microdata. *Journal of the Royal Statistical Society Series D*, 51, 1-11.
8. Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. Journal of Official Statistics. 9, 383-406.
9. Kim, J. 1986. A Method for limiting disclosure in microdata based on random noise and transformation. *Proceedings of the American Statistical Association*, Survey Research Methods Section, ASA, Washington D.C. 370-374.
10. Moore, R.A. 1996. Controlled data swapping for masking public use microdata sets. *U.S. Census Bureau Research Report 96/04*, http://www.census.gov/srd/papers/pdf/rr96-4.pdf.
11. Muralidhar K., R. Parsa, R. Sarathy. 1999. A general additive data perturbation method for database security. *Management Science*, 45 1399-1415.
12. Muralidhar K., R. Sarathy R., R. Parsa. 2001. An improved security requirement for data perturbation with implications for e-commerce. *Decision Sciences*, 32 683-698.
13. Muralidhar, K. and R. Sarathy. 2003a. A theoretical basis for perturbation methods. *Statistics and Computing*, 13 329-335.
14. Muralidhar, K. and R. Sarathy. 2006a. A comparison of multiple imputation and data perturbation for masking numerical variables. *Journal of Official Statistics* 22 507-524.
15. Muralidhar, K. and R. Sarathy. 2006b. The Myth of Micro-Aggregation. INFORMS Joint Meeting, October 5-7, Pittsburgh, PA.
16. Muralidhar, K. and R. Sarathy. 2006c. "Why swap when you can shuffle? A comparison of the proximity swap and the data shuffle for numeric data," in Domingo-Ferrer and Franconi, Eds.: *Privacy in Statistical Databases (PSD 2006)*, 164-176, Springer Verlag, Berlin.
17. Muralidhar, K. and R. Sarathy. 2006d. Data shuffling - A new masking approach for numerical data. *Management Science*, 52(5), 658-670.
18. Muralidhar, K. and R. Sarathy. 2007. Generating Sufficiency Based Non-Synthetic Perturbed Data. Working paper.
19. Raghunathan, T.E., J.P. Reiter, and D.B. Rubin. 2003. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19 1-6.
20. Rubin, D.B. 1987. *Multiple imputation for nonresponse in surveys*. Wiley, New York.
21. Rubin, D.B. 1993. Discussion of statistical disclosure limitation. *Journal of Official Statistics*, 9 461-468.
22. Sarathy, R., K. Muralidhar, and R. Parsa. 2002. Perturbing non-normal confidential attributes: The copula approach, *Management Science*, 48(12), 1613-1627.
23. Tendick, P. and N. Matloff 1994. A modified random perturbation method for database security. *ACM Transactions on Database Systems*, 19 47-63.
24. Ting, D., Fienberg, S., and Trottini, M. 2005. ROMM methodology for microdata release. *Monographs of Official Statistics – Work Session on Statistical Data Confidentiality*, Geneva, November 9-11, 89-98.
25. Winkler, W. 2002. Single-ranking micro-aggregation and re-identification. *Research Report Series (Statistics 2002-08)*, US Census Bureau, http://www.census.gov/srd/papers/pdf/rrs2002-08.pdf.
26. Winkler, W. 2006. "Modeling and quality of masked microdata," *Research Report Series (Statistics 2006-01)*, US Census Bureau, http://www.census.gov/srd/papers/pdf/rrs2006-01.pdf.