A Theoretical Comparison of Data Masking Techniques for Numerical Microdata

Krish Muralidhar
Gatton College of Business & Economics
University of Kentucky
Lexington KY 40506
krishm@uky.edu

Rathindra Sarathy
Spears School of Business
Oklahoma State University
Stillwater OK 74078
sarathy@okstate.edu

Abstract

In this study, we perform a comprehensive theoretical evaluation of masking techniques for numerical microdata. The objective of this comparison is to establish the extent to which existing techniques can satisfy disclosure risk, data utility, ease of implementation, and ease of use requirements. This evaluation allows data providers to select from these techniques to account for the demands of the data subjects and data users.

Introduction

In recent years interest in tools and techniques used to mask data has grown considerably. Until recently, government agencies have been the primary providers of confidential data regarding subjects (individuals and other entities) and the development of data masking techniques had been driven by the specific requirements of such agencies. The recent explosion in the ability to gather, store, analyze, and disseminate data by non-governmental organizations has given rise to the demand for new and innovative techniques for data masking. Recent legislative requirements have also increased the demand for such techniques.

Not surprisingly, there has been a corresponding increase in the development of new methods and techniques for data masking. From a research perspective, this is an encouraging trend and has led to the development of several new techniques. However, this also creates the issue of evaluating the performance of the different techniques that have been developed. Often, a particular technique is developed and its performance is evaluated based on specific criteria selected by the researcher making it impossible to compare different techniques.

The problem is further complicated by the fact that we now have researchers from different areas working on this problem. These include researchers at government agencies, statisticians, computer scientists, business, and social sciences. From a research perspective, this is an encouraging trend since this brings the nuances of these different areas to bear on the data masking problem. Unfortunately, the involvement of researchers

from these different areas has also meant that new techniques are evaluated in different ways across these areas.

In addition, there has been a recent trend to evaluate the performance of different techniques using *only empirical data*. We do not disagree with the use of empirical data to illustrate the application of a new technique. However, we vigorously disagree when a few data sets are used to "validate" or "prove" the superior performance of any technique. The superiority of a technique can only be established *from a comprehensive theoretical comparison which can then be backed up by confirmation using simulated or empirical data*.

Thus, there is a need to develop a general framework that can be used to assess the theoretical performance of data masking techniques. In this paper, we attempt to develop such a framework for techniques used to mask numerical microdata from the traditional perspective of the literature in statistical disclosure limitation. In developing this framework, we attempt to take into account the interests of all the parties involved in the data. While we do not contend that this framework is appropriate for all situations, we believe that this framework represents the most general assessment of data masking techniques.

The Stakeholders

Generally, there are three stakeholders who are involved in this problem. The first are those individuals or organizations about whom the data has been collected ("The subjects"). The second is the organization or agency that collects the data and then analyzes, shares, or disseminates the data ("The provider"). Finally, there are the consumers of the data who actually use the data that is being made available ("The user"). Obviously, the interests of each of these three stakeholders differ and in some cases conflict. The user would in fact prefer to get the original data without any modification while the data subjects may prefer that this data not be analyzed, shared or disseminated by anyone. In these cases, the data provider becomes the arbiter of the (utility) demands of the user versus the (confidentiality) demands of the source. One key assumption that we are making in this analysis is that the data is intended for aggregate analysis and not analysis involving individual records or observations.

Data Subjects

Data subjects are individuals (or organizations) about whom data has been gathered. This data could have been collected in a variety of ways. The data could have been collected in the ordinary course of business (such as when data is reported to the Internal Revenue Service) or for the specific purpose of gathering data regarding individuals (such as when data is gathered by the Census Bureau). The data subjects are aware that some analysis will be conducted on such data by a *select group of individuals* (by IRS officials or Census officials). In most cases, they are aware that their privacy and confidentiality is not assured when such analysis is conducted.

The data subjects may also be aware that such data may be shared, or disseminated to other individuals. However, when the data is provided to *individuals for whom the original data was not intended*, the data subjects demand privacy and confidentiality guarantees. This is a completely reasonable requirement since in most situations they cannot "opt out" from their data being shared. Ideally the data subjects would prefer that the data is not analyzed, shared or disseminated beyond the restricted group of individuals for whom the data was intended. However, if the decision is made that the data will be available to outside individuals, then obviously they would like the data provider to take the utmost care to prevent disclosure of their identity and/or value of a confidential attribute.

The Data User

This group represents those individuals who actually use the confidential data. The data user could be an analyst within the organization in which the data resides. More commonly however, the user is outside the organization and is receiving the data because it is being shared or disseminated. Ideally, the user would like the data to be released unmodified since the user is not concerned with confidentiality of the data (unless of course the user also happens to also be a data subject). However, the user expects that some type of masking will be performed on the data. The primary interest of the user is the analytical validity of the data that is provided. A secondary interest of the user is the ease of performing analysis on the data. Ultimately, the data user expects the results of analysis of the masked data to yield the same results as the original data.

The Data Provider

The provider acts as the trusted intermediary between the data subjects and the data user. The provider has obligations both to the data subjects and the data user., The data provider is obligated to the data subjects to maintain both their privacy and confidentiality. To the data user, the data provider is obligated to provide assurance regarding the analytical validity of the data. In addition, the data provider is also interested in selecting the method that is "easy to implement". However, when choosing between methods, if one method dominates the performance of another method (provides better performance characteristics on at least one criterion and the same benefits on all others), the data provider is *obligated* to choose the dominant method regardless of the ease of implementation.

Criteria for Evaluating Performance of Data Masking Techniques

Data Security

The primary responsibility of the data provider is to assure the data subjects that the data will be protected to the fullest extent possible. The focus of this study is on numerical data. However, the framework described here should be applicable for non-numerical data as well. In protecting the data, it would be necessary to consider both identity and value disclosure. Identity disclosure refers to the situation where, using the released data, an intruder is able to identify a particular released record as belonging to a particular data subject. Attribute or value disclosure refers to the situation where, using the released data,

an intruder is able to estimate the value of a confidential variable with a high degree of precision (Lambert 1993).

Prior to releasing microdata, we assume that the data provider will release aggregate summary information regarding the variables (including but not limited to) the mean vector (for numeric variables), percentages (for non-numeric variables) and covariance matrix of the entire data set. Since the release of this information alone does not result in any disclosure regarding individual records, we believe that this is an appropriate first step for data release. Note that releasing this data could also serve as a benchmark for evaluating the analytical validity of the masked data.

The data provider now has to assess whether any microdata should be released and if so, which variables are to be released. The data provider may decide that, because of high disclosure risk, no microdata will be released. On the other hand, if the data provider has decided to release microdata, the variables whose microdata will be released should be selected carefully. In this study, we assume that after careful assessment, the data provider has selected a set of non-confidential or categorical variables **S** and a set of numerical confidential variables **X** for n records are to be released. Let **Y** represent the masked values of the numerical confidential variables **X**.

The set of variables **S** could comprise both of categorical variables and, in situations where identity disclosure is not an issue, of numerical confidential variables. However, in situations where identity disclosure is an issue (which is probably the general case), numerical confidential variables cannot be released unmodified since releasing just a few such variables will inevitably result in complete identity disclosure. Further, we assume that the appropriate masking techniques have been applied to categorical variables in order to prevent disclosure.

Our specific objective is to assess the disclosure risk performance of a masking technique for numerical data. Hence, any disclosure that occurs as a result of the release of the non-confidential information combined with the summary data should not be considered as a part of the disclosure risk resulting from the masking technique. With this in mind, we propose a two step process for evaluating the risk of disclosure. The first step involves the assessment of disclosure risk in the confidential variables **X** using the summary information and the non-confidential variables **S**. Recall that we always assume that at least summary information of the original data has to be released; otherwise, the data protection problem does not exist since no meaningful use of the confidential data can be presumed. We will define the disclosure risk due to the release of summary data as DR(**X**|**S**). We then compute the disclosure risk from releasing the entire released (potentially, masked) data set DR(**X**|S,**Y**). The disclosure risk attributable to the masking technique is the *incremental risk of disclosure that occurs because of the release of the masked data **Y** in place of **X*** which can be computed as DR(**X**|**S**,**Y**) – DR(**X**|**S**).

There exists a class of data masking techniques for which the expected incremental disclosure risk can theoretically be shown to be zero. For these techniques, the masked variable **Y** is generated independent of the values of the confidential variable **X** and as a

function only of the non-confidential variables **S**. Several authors including Fienberg (1998), Fuller (1993), and Muralidhar and Sarathy (2003) have shown that the methods in this class have an expected incremental disclosure of zero, thereby guaranteeing *minimum disclosure risk* when numerical microdata is released. The specific techniques which satisfy these requirements will be discussed in a later section.

When masking techniques do not satisfy the minimum disclosure risk criterion, it would be necessary to assess the risk of disclosure using empirical techniques. A comprehensive discussion of these techniques is beyond the purview of this manuscript.

Data Utility

We use data utility to describe the extent to which results of analysis performed on the masked data is similar to the results using the original data. This criterion is also sometimes referred to as analytical validity, or conversely, information loss. A masking technique offers the highest level of data utility if *for any ad hoc analysis, the results using the masked data are identical to that using the original data*. Since we are considering any ad hoc analysis, it is possible that the analysis involves a single variable for a single record. Hence, the only way to satisfy this ideal requirement is to release the unmasked data. However, because the data consists of confidential variables, the data will be masked and consequently, no masking technique will satisfy the ideal data utility objective.

In practice, it would be necessary to evaluate the masking techniques using specific criteria that describe different aspects of data utility. In this study, we assess both the marginal and the joint characteristics of the variables. The marginal assessment is relatively straightforward and can be answered by the question: "Is the marginal distribution of the masked variable(s) the same as that of the original confidential variable(s)?"

The assessment of the joint distribution of the variables is far more difficult, with no single means of assessing the joint distribution. Hence, it is necessary to evaluate specific characteristics which when combined provide an assessment of the joint distribution. The specific characteristics include:

(1) Is the mean vector of the masked variables the same as the mean vector of the original confidential variables?
(2) Does the masking technique preserve linear relationships?
    a. Is the covariance matrix of the masked confidential variables the same as the covariance matrix of the original confidential variables?
    b. Is the covariance matrix between the masked confidential and non-confidential variables the same as the covariance matrix between the original confidential and non-confidential variables?
(3) Does the masking technique preserve monotonic non-linear relationships?
(4) Does the masking technique preserve non-monotonic relationships between all the variables?

(5) Does the masking technique preserve the above characteristics in subsets of the data defined by the categorical variables?

There are four possible responses to the above questions. The first response is that the values of the estimates using the masked data are *exactly the same as the original* data. The second response is that the estimates from the masked data are *unbiased* and, asymptotically, the estimates from the masked data converge to the estimates using the original data. The third response is that the estimates from the masked data are *biased* compared to those derived from the original data. The final response is that the estimates are *not maintained* by the masked data at all.

We do not contend that this is an exhaustive list of all possible characteristics that may be of interest in every situation. For example, if it is known that the joint distribution of the entire data set is multivariate normal, then the only relevant estimates are the mean vector and covariance matrix. However, the criteria listed allow us to evaluate the performance of any masking technique for any data set.

Ease of Use

Typically, data users would like to analyze the data in the same manner as the original data set without having to use special procedures to account for any differences between the original and masked data set. From the data user's perspective, the extent to which the masked data facilitates this process is an important consideration. For some masking techniques, in order to achieve unbiased estimates, it would be necessary to modify the results obtained from the analyses performed on the data. Regardless of the nature of the modifications necessary, this imposes additional requirements on the data users which, if possible, should be eliminated. In addition, we cannot always assume that every user is capable of implementing and more importantly understanding the suggested modifications.

Ease of Implementation

From the perspective of a data provider, a masking technique that is easier to implement would be preferred over those that are more difficult to implement. While all techniques require some level of cleansing and preparation, some advanced procedures require more effort to implement than others. Given that all other criteria are equal, it is reasonable for the data provider to select a masking technique that is easier to implement than another. Winkler (2007) argues that this criterion seems to dictate the selection of the appropriate technique with simpler techniques being preferred over more complex techniques. This is unfortunate since Winkler (2007) clearly shows that simple to implement techniques usually have low data utility.

An Evaluation of Common Masking Techniques for Numerical Data

In this section, we attempt to evaluate the performance of most commonly used data masking techniques for numerical data, using the criteria described above. In our

evaluation, it must be understood that when the expected incremental disclosure risk is not zero, we say that disclosure risk is not minimized.

Simple Noise Addition

In simple noise addition, noise is added to each confidential variable independent of all other variables. From a multivariate perspective, this type of noise addition can be seen as using noise terms that are not only independent of the original values, but also independent of one another and can be described as follows:

$$\mathbf{Y} = \mathbf{X} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ is specified to have a mean vector of zero with a diagonal covariance matrix. In terms of disclosure risk, since the masked values are not generated independent of the original confidential values, the incremental disclosure is non-zero. The level of security provided is a function of the level of noise added (the variance of the noise term $\boldsymbol{\varepsilon}$) and must be evaluated empirically. In terms of data utility, since the noise is being added to the original confidential variable the marginal distribution of **Y** is different from that of **X**. In addition, since noise is being added to each variable independently, the covariance matrix of **Y** is different from that of **X**. In addition, the covariance between **Y** and **S** is also different from that between **X** and **S**. The same is true for rank order correlation as well. No relationships are preserved either exactly or asymptotically and all relationships result in biased estimates. The extent of the bias is a function of the level of noise added. The procedure is easy to use and implement.

Table 1. Performance characteristics of simple noise addition

Criteria			**Disclosure Risk**	Not minimized
	Data Utility		Marginal Distribution	Not maintained
			Mean Vector	Maintained asymptotically
			Variance	Biased (higher)
		Linear Relationships	**X** versus **Y**	Biased (attenuated)
			(**Y** & **S**) versus (**X** & **S**)	Biased (attenuated)
			Monotonic (non-linear) relationships	Biased (attenuated)
			Non-monotonic relationships	Biased (attenuated)
			Sub-set characteristics	Not maintained
			Ease of Use	Easy
			Ease of Implementation	Easy

Kim's Method

Kim (1986) proposed an important enhancement to the simple noise addition procedure. In Kim's method, the covariance matrix of the noise terms $\boldsymbol{\varepsilon}$ is specified to be of the form $(d\boldsymbol{\Sigma}_{XX})$ where d is a constant chosen by the data provider and $\boldsymbol{\Sigma}_{XX}$ is the covariance matrix of **X**. A further enhancement was discussed by Kim (1986) and illustrated in Tendick and Matloff (1994). The advantage of this procedure is that the covariance matrix of the masked variables **Y** is the same as that of the original variables **X**. Apart from this enhancement, this procedure has similar characteristics to simple noise addition

as summarized below. In terms of implementation, the only difference between this procedure and simple noise addition is the generation of **ε**.

Table 2. Performance characteristics of Kim's method

Criteria			**Disclosure Risk**	Not minimized
	Data Utility		Marginal Distribution	Not maintained
			Mean Vector	Maintained asymptotically
			Variance	Maintained asymptotically
		Linear Relationships	**X** versus **Y**	Maintained asymptotically
			(**Y & S**) versus (**X** & **S**)	Biased (attenuated)
			Monotonic (non-linear) relationships	Biased (attenuated)
			Non-monotonic relationships	Biased (attenuated)
			Sub-set characteristics	Not maintained
			Ease of Use	Easy
			Ease of Implementation	Easy

Sufficiency Based Noise Addition Method

Recently, Muralidhar and Sarathy (forthcoming) proposed a new noise addition procedure whereby the mean vector and covariance matrix of the masked data are *exactly* the same as that of the original data. In this procedure, the masked values are generated as follows:

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\gamma}\mathbf{S} + \boldsymbol{\varepsilon}.$$

The data provider specifies the structure of **β** from which the values of **α**, **γ**, and the covariance matrix of **ε** are derived. Using these derivations, the values of **ε** are generated in such a manner that the mean vector and covariance matrix of **Y** are exactly the same as**X**. Hence, for any statistical analysis for which the mean vector and covariance matrix of **X** are sufficient statistics, the estimates derived using the masked data are identical to that using the original data. In addition, this procedure can be easily implemented in such a manner that characteristics of subsets of **X** defined by **S** are maintained. However, like all noise addition procedures, this procedure does not minimize disclosure risk, and results in bias in other estimates. In terms of implementation, this procedure is computationally more difficult than the previous approaches. The summary of the performance characteristics are provided in Table 3.

Table 3. Summary performance characteristics of sufficiency based noise addition

Criteria			**Disclosure Risk**	Not minimized
	Data Utility		Marginal Distribution	Not maintained
			Mean Vector	Maintained exactly
			Variance	Maintained exactly
		Linear Relationships	**X** versus **Y**	Maintained exactly
			(**Y** & **S**) versus (**X** & **S**)	Maintained exactly
			Monotonic (non-linear) relationships	Biased (attenuated)
			Non-monotonic relationships	Biased (attenuated)
			Sub-set characteristics	Mean vector and covariance matrix maintained exactly
			Ease of Use	Easy
			Ease of Implementation	Moderate

Sufficiency Based GADP Method

This procedure is a combination of the GADP method (Muralidhar et al, 1999, 2001) and the IPSO method (Burridge 2003). In the GADP method, using the available data, the intercept ($\boldsymbol{\alpha}$) and slope coefficients ($\boldsymbol{\beta}$) to predict the value of **X** using **S** is estimated, as is the error variance. The masked values of Y are then generated using these estimates as follows:

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\gamma}\mathbf{S} + \boldsymbol{\varepsilon}$$

The mean vector and covariance matrix of (**Y** and **S**) are *asymptotically* the same as that of (**X** and **S**). The values of **Y** are considered *synthetic* since they are generated independent of the values of **X**. In this IPSO method, Burridge (2003) proposed a simple but very attractive modification where he showed that by generating the values of $\boldsymbol{\varepsilon}$ appropriately, the mean vector and covariance matrix of (**Y** and **S**) are *exactly* the same as that of (**X** and **S**), thereby maintaining the two sufficient statistics mean vector and covariance matrix. Note that this procedure is a special case of the sufficiency based noise addition method with the requirement that $\boldsymbol{\beta} = \mathbf{0}$. The advantage of this method is that in addition to maintaining sufficient statistics, this procedure also minimizes disclosure risk since the masked values are generated independent of the original values. As with the previous method, it is also easy to adapt this method to provide sufficient statistics not just for the entire data set, but sub-sets as well. One important aspect of this procedure is that, since the values of **Y**are generated using a linear model, any non-linear relationship that exists in the original data set is not preserved in the masked data set. The performance characteristics summary is provided in Table 4.

Table 4. Summary performance characteristics of the sufficiency based GADP

Criteria			**Disclosure Risk**	Minimized
	Data Utility		Marginal Distribution	Not maintained
			Mean Vector	Maintained exactly
			Variance	Maintained exactly
		Linear Relationships	**X** versus **Y**	Maintained exactly
			(**Y & S**) versus (**X** & **S**)	Maintained exactly
			Monotonic (non-linear) relationships	Not maintained
			Non-monotonic relationships	Not maintained
			Sub-set characteristics	Mean vector and covariance matrix maintained exactly
			Ease of Use	Easy
			Ease of Implementation	Moderate

Multiple Imputation

Originally proposed for missing data, multiple imputation was suggested as a possible mechanism 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 generates the perturbed values as in the synthetic data approach. Using the available data, the intercept, slope coefficients, and error variance are estimated. 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.

In terms of disclosure risk, in the original form of multiple imputation the values of **Y** are generated independent of the values of **X**. This synthetic nature of the approach assures that the disclosure risk is minimized. In terms of the data utility, the performance of multiple imputation is very similar to that of the sufficiency based synthetic data method with one important exception. The sufficiency based synthetic data method guarantees that the mean vector and covariance matrix of the masked data set will be identical to the original data. Multiple imputation does not offer this guarantee. In addition, multiple imputation requires the user to analyze multiple data sets (perhaps as many as 100) and then aggregate the results. This imposes additional computational requirements on the user. By contrast, the sufficiency based GADP method provides the same results as the multiple imputation approach without the additional computational requirements (Muralidhar and Sarathy 2006a). In its original form, multiple imputation cannot maintain subset characteristics (since it may not even have the same subsets). Variations

of the multiple imputation approach have been suggested and can be evaluated similarly. A summary of the performance of multiple imputation is provided in Table 5.

Table 5. Summary Performance characteristics of multiple imputation

Criteria			**Disclosure Risk**	Minimized
	Data Utility		Marginal Distribution	Not maintained
			Mean Vector	Maintained asymptotically
			Variance	Maintained asymptotically
		Linear Relationships	**X** versus **Y**	Maintained asymptotically
			(**Y & S**) versus (**X** & **S**)	Maintained asymptotically
			Monotonic (non-linear) relationships	Not maintained
			Non-monotonic relationships	Not maintained
			Sub-set characteristics	Not maintained
			Ease of Use	Difficult
			Ease of Implementation	Moderate

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). A summary of the characteristics of micro-aggregation is provided in Table 6.

Table 6. Summary performance characteristics of micro-aggregation

Criteria			**Disclosure Risk**	Not minimized
	Data Utility		Marginal Distribution	Not maintained
			Mean Vector	Maintained exactly
			Variance	Biased (lower)
		Linear Relationships	**X** versus **Y**	Biased
			(**Y & S**) versus (**X** & **S**)	Biased
			Monotonic (non-linear) relationships	Biased
			Non-monotonic relationships	Biased
			Sub-set characteristics	Not maintained
			Ease of Use	Easy
			Ease of Implementation	Moderate

In terms of disclosure risk, the masked values **Y** are not independent of the original values and hence, disclosure risk is not minimized. In terms of data utility, micro-aggregation results in reduced variance for most data sets. The lower variance of the confidential variables results in an accentuation in the correlation between variables. However, micro-aggregation also results in attenuating relationships. In some situations

the attenuation in correlation is higher than the accentuation in correlation resulting from the reduction in variance and the overall result is correlation attenuation. Thus, micro-aggregation always results in biased estimates of relationships, but the direction of the bias cannot be derived theoretically. In terms of implementation, univariate micro-aggregation is relatively easy to implement while some forms of multivariate micro-aggregation are difficult. This method does not require the user to make any modification in their analyses and hence is easy to use.

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 (univariate) marginal distribution as the original confidential variables. However, unless the swapping is performed randomly, data swapping does not minimize disclosure risk. In addition, data swapping also results in attenuation in the relationship between the variables (Moore 1996; Fienberg and McIntyre 2005). As indicated earlier, it is easy to implement and easy to use. Table 7 provides the summary characteristics for the data swapping method.

Table 7. Summary performance characteristics of data swapping

Criteria			**Disclosure Risk**	Not minimized
	Data Utility		Marginal Distribution	Maintained exactly
			Mean Vector	Maintained exactly
			Variance	Maintained exactly
		Linear Relationships	**X** versus **Y**	Biased (attenuated)
			(**Y & S**) versus (**X** & **S**)	Biased (attenuated)
			Monotonic (non-linear) relationships	Biased (attenuated)
			Non-monotonic relationships	Biased (attenuated)
			Sub-set characteristics	Not maintained
			Ease of Use	Easy
			Ease of Implementation	Easy

Data Shuffling

Data Shuffling is a new patented procedure (US Patent # 7200757) developed by Muralidhar and Sarathy (2006b). 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. Data shuffling is also a more general version of the LHS procedure suggested by Dandekar et al. (2002).

Data shuffling does not use the original values of **X** in generating the values of **Y** and hence offers minimum disclosure risk. Since the original values of **X** are used in the masked values of **Y**, the marginal distribution is maintained exactly. In addition, data shuffling maintains both linear and (non-linear) monotonic relationships. But all non-monotonic relationships in the original data are not maintained in the masked data. Table 8 provides a summary of these characteristics.

Table 8. Summary performance characteristics of data shuffling

Criteria			**Disclosure Risk**	Minimized
	Data Utility		Marginal Distribution	Maintained exactly
			Mean Vector	Maintained exactly
			Variance	Maintained exactly
		Linear Relationships	**X** versus **Y**	Maintained asymptotically
			(**Y & S**) versus (**X** & **S**)	Maintained asymptotically
			Monotonic (non-linear) relationships	Maintained asymptotically
			Non-monotonic relationships	Not maintained
			Sub-set characteristics	Not maintained
			Ease of Use	Easy
			Ease of Implementation	Difficult

Tree-Based Data Perturbation Method

Li and Sarkar (2006) recently proposed a new approach for data perturbation that is described by the authors is as follows:

> "To achieve this goal, we propose a kd-tree based perturbation method, which recursively partitions a data set into smaller subsets such that data records within each subset are more homogeneous after each partition. The confidential data in each final subset are then perturbed using the subset average." (page 1278)

The authors then proceed to illustrate the application of this method using 4 data sets. The authors also compare the performance of the tree based data perturbation method to noise addition and micro-aggregation and conclude, *based on 4 data sets* that the tree-based perturbation method performs better. Unfortunately, the authors provide no assessment of the general characteristics of the procedure.

From the description of the procedure, it is easy to see that this procedure essentially is micro-aggregation with the specific records selected differently from traditional micro-aggregation. Hence, the essential characteristics of this procedure should be similar to that of micro-aggregation. Using the framework we have suggested in this study, Table 9 provides a summary of the performance characteristics of this method.

Table 9. Summary performance of the tree-based perturbation method

Criteria			**Disclosure Risk**	Not minimized
	Data Utility		Marginal Distribution	Not maintained
			Mean Vector	Maintained exactly
			Variance	Biased (lower)
		Linear Relationships	**X** versus **Y**	Biased
			(**Y & S**) versus (**X** & **S**)	Biased
			Monotonic (non-linear) relationships	Biased
			Non-monotonic relationships	Biased
			Sub-set characteristics	Not maintained
			Ease of Use	Easy
			Ease of Implementation	Easy

Comparative Evaluation of Different Data Masking Methods

Comparison of Tree-based perturbation with other methods

Li and Sarkar (2006) claim that tree-based perturbation method is "generally superior" to the other methods of data masking used in their study. The authors make no effort to theoretically evaluate the performance characteristics of their new approach. While an empirical evaluation can be a useful tool for illustrating the performance of a particular method, it must be preceded by a comprehensive theoretical evaluation of the performance of a particular technique. If Li and Sarkar (2006) had adopted the approach suggested in this study to evaluate performance, they would have come up with very different conclusions. Based on the framework suggested above, we compare the performance of the tree-based perturbation method, micro-aggregation, Kim's method, and sufficiency based noise addition. The results are provided in Table 10.

Based on the results in Table 10, if the data provider is to select the procedure based on theoretical characteristics, the sufficiency based noise addition method would be the preferred method since it provides better data utility characteristics than the other methods. Since all four methods fail to minimize disclosure risk, it would be appropriate to perform some empirical evaluation of the disclosure risk performance of the methods prior to making the final selection. However, given that all four methods allow for the specification of the security parameter, we should be able to achieve the desired level of security.

What is interesting about the results in Table 10 is that, in terms of data utility, the performance of Kim's method is actually superior to that of the tree-based perturbation method. In fact, in terms of data utility, the tree-based perturbation method does not provide better performance than any of the other techniques. Yet, in an empirical comparison involving four data sets, the authors claim that the tree-based perturbation method is "generally superior". A theoretical evaluation of the performance of the techniques shows otherwise.

Table 10. Comparison of the Tree-based perturbation method with other methods

				Method			
				Tree Based Perturbation	**Micro-Aggregation**	**Kim's Method**	**Sufficiency Based noise addition**
Criteria	**Disclosure Risk**			Not minimized	Not minimized	Not minimized	Not minimized
	Data Utility		Marginal Distribution	Not maintained	Not maintained	Not maintained	Not maintained
			Mean Vector	Maintained exactly	Maintained asymptotically	Maintained asymptotically	Maintained asymptotically
			Variance	Biased (lower)	Biased (lower)	Maintained asymptotically	Maintained exactly
		Linear Relationships	**X** versus **Y**	Biased	Biased	Maintained asymptotically	Maintained exactly
			(**Y & S**) versus (**X & S**)	Biased	Biased	Biased (attenuated)	Maintained exactly
			Monotonic (non-linear) relationships	Biased	Biased	Biased (attenuated)	Biased (attenuated)
			Non-monotonic relationships	Biased	Biased	Biased (attenuated)	Biased (attenuated)
			Sub-set characteristics	Not maintained	Not maintained	Not maintained	Mean vector and covariance matrix maintained exactly
	Ease of Use			Easy	Easy	Easy	Easy
	Ease of Implementation			Easy	Easy	Easy	Moderate

Comparison of noise addition methods

As the first illustration, we choose to compare the different noise addition methods. This is a rather straight-forward assessment that can be done just from the summary provided in Tables 1, 2, and 3. The summary of all 3 methods is provided in Table 11.

Table 11. Comparison of noise addition methods

				Method		
				Simple Noise Addition	**Kim's Method**	**Sufficiency Based noise addition**
Criteria	**Disclosure Risk**			Not minimized	Not minimized	Not minimized
	Data Utility		Marginal Distribution	Not maintained	Not maintained	Not maintained
			Mean Vector	Maintained exactly	Maintained asymptoticall y	Maintained asymptotically
			Variance	Biased (lower)	Biased (lower)	Maintained exactly
		Linear Relationships	**X** versus **Y**	Biased	Biased	Maintained exactly
			(Y & S) versus **(X & S)**	Biased	Biased	Maintained exactly
			Monotonic (non-linear) relationships	Biased	Biased	Biased (attenuated)
			Non-monotonic relationships	Biased	Biased	Biased (attenuated)
			Sub-set characteristics	Not maintained	Not maintained	Mean vector and covariance matrix maintained exactly
	Ease of Use			Easy	Easy	Easy
	Ease of Implementation			Easy	Easy	Moderate

The comparison of the three approaches clearly shows that none of the methods minimize disclosure risk. However, it is possible to select the specification for each of the methods such that the resulting disclosure risk for all three methods is comparable. Hence, the methods can be evaluated based on the extent to which the methods satisfy the data utility criteria. From the table, it is evident that the performance of the sufficiency based method dominates the performance of the other two methods since it provides the same or better

performance on every criterion. From the perspective of the data user, all three methods are equally easy to use. However, the sufficiency based approach is, in relative terms, more difficult to implement than the other two methods.

Thus, if the objective of data masking was to protect the privacy and confidentiality of the respondents and to provide data that is of high analytical value, the sufficiency based noise addition should be preferred over the other noise addition approaches.

Comparison of synthetic data perturbation approaches

In this case, we compare the two synthetic data approaches of multiple imputation and sufficiency based GADP. For a comprehensive (theoretical and empirical) evaluation of the two methods, please see Muralidhar and Sarathy (2006a). The summary of the comparison is provided in Table 12. The summary results easily show that the performance of the sufficiency based GADP is superior to that of or equal to the performance of multiple imputation. Hence, sufficiency based GADP should be preferred to multiple imputation.

Table 12. Comparison of multiple imputation and sufficiency based GADP

				Method	
				Multiple Imputation	**Sufficiency based GADP**
Criteria	**Disclosure Risk**			Minimized	Minimized
	Data Utility	Marginal Distribution		Not maintained	Not maintained
		Mean Vector		Maintained asymptotically	Maintained exactly
		Variance		Maintained asymptotically	Maintained exactly
		Linear Relationships	**X** versus **Y**	Maintained asymptotically	Maintained exactly
			(Y & S) versus **(X & S)**	Maintained asymptotically	Maintained exactly
		Monotonic (non-linear) relationships		Not maintained	Not maintained
		Non-monotonic relationships		Not maintained	Not maintained
		Sub-set characteristics		Not maintained	Subset mean vector and covariance matrix maintained exactly
	Ease of Use			Difficult	Easy
	Ease of Implementation			Moderate	Moderate

Comparison of data swapping and data shuffling

As a final illustration, we compare the performance of data swapping and data shuffling. The unique advantage of both these approaches is that they are data masking techniques that do not require modifying the original confidential values. Hence, the marginal distribution of the masked variable is exactly the same as the original confidential variable. While both methods have this performance characteristic in common, they differ considerably in other characteristics. The comparative evaluation of the two approaches is provided in Table 13. For a comprehensive theoretical and empirical comparison of the two methods, please refer to Muralidhar et al. (2006c).

Table 13. Comparison of data swapping and data shuffling

				Method	
				Data Swapping	**Data Shuffling**
Criteria	**Disclosure Risk**			Not minimized	Minimized
	Data Utility	Marginal Distribution		Maintained exactly	Maintained exactly
		Mean Vector		Maintained exactly	Maintained exactly
		Variance		Maintained exactly	Maintained exactly
		Linear Relationships	**X** versus **Y**	Biased (attenuated)	Maintained asymptotically
			(**Y** & **S**) versus (**X** & **S**)	Biased (attenuated)	Maintained asymptotically
		Monotonic (non-linear) relationships		Biased (attenuated)	Maintained asymptotically
		Non-monotonic relationships		Biased (attenuated)	Not maintained
		Sub-set characteristics		Marginal characteristics maintained exactly; all other relationships are attenuated	Marginal characteristics maintained exactly; linear and monotonic relationships maintained asymptotically
	Ease of Use			Easy	Easy
	Ease of Implementation			Easy	Difficult

As is evident from the above table, data shuffling dominates the performance of data shuffling in every category except “ease of implementation”. As observed earlier if the objective of the data masking technique is to provide lowest level of disclosure risk and highest level of data utility data shuffling would be the preferred method. The only reason that data swapping would be preferred over data shuffling would be if the data

provider makes the selection based exclusively on ease of implementation. Unfortunately, such a selection would have an adverse impact on both the respondents who provided the data and the users.

Comparison of sufficiency based noise addition and data shuffling
In this example, we selected two methods where the performance of one method does not necessary dominate that of the other. The performance summary for the two methods is provided in Table 14.

Table 14. Comparison of sufficiency based noise addition and data shuffling

				Method	
				Sufficiency based noise addition	**Data Shuffling**
Criteria	**Disclosure Risk**			Not minimized	Minimized
	Data Utility		Marginal Distribution	Not maintained	Maintained exactly
			Mean Vector	Maintained exactly	Maintained exactly
			Variance	Maintained exactly	Maintained exactly
		Linear Relationships	**X** versus **Y**	Maintained exactly	Maintained asymptotically
			(**Y & S**) versus (**X** & **S**)	Maintained exactly	Maintained asymptotically
			Monotonic (non-linear) relationships	Biased (attenuated)	Maintained asymptotically
			Non-monotonic relationships	Biased (attenuated)	Not maintained
			Sub-set characteristics	Mean vector and covariance matrix are maintained exactly	Marginal characteristics maintained exactly; linear and monotonic relationships maintained asymptotically
	Ease of Use			Easy	Easy
	Ease of Implementation			Moderate	Difficult

From the above table, it is difficult to determine which of the two methods would be preferred. From the perspective of disclosure risk, data shuffling would be preferred. Data shuffling also provides the ability to maintain the marginal distribution as well as preserving (asymptotically) linear and monotonic relationships. However, using data shuffling will not maintain non-monotonic relationships. By contrast, the sufficiency

based noise addition method maintains the mean vector and covariance matrix exactly which provides some significant advantages to the user. Furthermore, unlike data shuffling which will not maintain non-monotonic relationships, sufficiency based noise addition will attenuate the relationship but will not completely eliminate these relationships. The selection of the appropriate procedure in this case should be based on the specific context and perhaps an empirical evaluation of both methods.

Conclusion

Several new data masking techniques have been proposed in recent years allowing data providers many alternatives to choose from. However, in selecting the most appropriate technique, the data provider must carefully consider the theoretical performance of the procedure. Recently, there has been a disturbing trend towards relying simply on empirical assessments to evaluate the performance of a procedure.

Consider the case of the Tree-based perturbation method. It is easy to see that this approach is essentially a variation of the traditional micro-aggregation method. For micro-aggregation, it is well known that the relationship estimates using the masked data are biased estimates of the original relationship. The authors evaluate the performance of the tree-based perturbation method along with two types of micro-aggregation, simple noise addition, and multiplicative noise addition. We can *theoretically show that all of these approaches result in bias in measuring the correlation between the masked variable and all other variables*. Hence, if the primary purpose for which the data will be used is for traditional regression analysis, these approaches would not be considered as good alternatives. However, based on applying the techniques to 4 data sets, the authors conclude that "indicating that all five methods perform very well for regression for these data sets." Such a conclusion is meaningless and misleading.

The empirical data approach for evaluating the performance of data masking techniques also does not allow us to generalize the results observed for a limited number of data sets to all data sets. Such *a general conclusion can come only from evaluating the performance characteristics theoretically*. This study provides a general framework for such a theoretical evaluation. Using a structured framework like the one described in this study allows us to evaluate the performance of different techniques from the perspective of all the stakeholders, namely, the data subjects, data users, and data providers.

Finally, such an evaluation serves another important purpose. As a part of the data release, it allows the data provider to state explicitly the performance characteristics of the data masking method, which provides specific assurances to the data subjects regarding disclosure risk and specific assurances to the data user regarding data utility. For instance, when the sufficiency based GADP is used as the data masking tool, the data provider to make the following very specific statement:

> "In order to protect the privacy and confidentiality of the data, the original data has been masked. The masking approach assures the highest possible level of security. In addition, the masking has been performed in such a

manner that, for any traditional statistical analyses for which the mean vector and covariance matrix are sufficient statistics (such as simple hypothesis testing, ANOVA, regression, MANOVA, basic principal components, canonical correlation analysis), *the estimates using the masked data will yield exactly the same estimates as the original data.* However, the marginal distribution of the individual variables has been modified. The procedure also does not maintain non-linear relationships."

We believe that every masked data should be accompanied by such a statement and a comprehensive description of the data masking procedure and, where appropriate, the parameters of the data masking procedure.

References

Burridge, J. (2003). Information preserving statistical obfuscation. *Statistics and Computing*, 13 321-327.

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.

Domingo-Ferrer, J. and J.M. Mateo-Sanz (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*. 14, 189-201.

Fienberg, S.E. and J. McIntyre (2005). Data swapping: Variations on a theme by Dalenius and Reiss. Journal of Official Statistics, 21 309-323.

Fienberg, S.E., Makov, U.E., and Steele, R.J. (1998). Disclosure Limitation Using Perturbation and Related Methods for Categorical Data, *Journal of Official Statistics*. 14, 485-502.

Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*. 9, 383-406.

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.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics*, 9(2), 313-331.

Li, X.B. and S. Sarkar, "A Tree-Based Data Perturbation Approach for Privacy-Preserving Data Mining," IEEE Transactions on Knowledge and Data Engineering, vol. 18, pp. 1278-1283, 2006.

Moore, R.A. 1996. Controlled data swapping for masking public use microdata sets. *U.S. Census Bureau Research Report RR96/04*, Washington DC.

Muralidhar K., R. Parsa, R. Sarathy (1999). A general additive data perturbation method for database security. *Management Science*. 45, 1399-1415.

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.

Muralidhar, K. and R. Sarathy (2003). A theoretical basis for perturbation methods. *Statistics and Computing*. 13, 329-335.

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.

Muralidhar, K. and R. Sarathy (2006b). Data shuffling - A new masking approach for numerical data. Management Science, 52(5), 658-670.

Muralidhar, K. and R. Sarathy (forthcoming). Generating sufficiency-based non-synthetic perturbed data. *Transactions on Data Privacy*.

Muralidhar, K., R. Sarathy, and R. Dandekar (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.

Raghunathan, T.E., J.P. Reiter, and D.B. Rubin (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*. 19, 1-6.

Rubin, D.B. (1993). Discussion on 'Statistical Disclosure Limitation'. *Journal of Official Statistics*. 9, 461-468.

Sarathy R., K. Muralidhar, R. Parsa. 2002. Perturbing non-normal confidential variables: The copula approach. *Management Science*. 48, 1613-1627.

Tendick, P. and N. Matloff (1994). A modified random perturbation method for database security. *ACM Transactions on Database Systems*. 19, 47-63.

Winkler, W.E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. *Census Bureau Research Report* RRS2007/21 (http://www.census.gov/srd/papers/pdf/rrs2007-21.pdf).