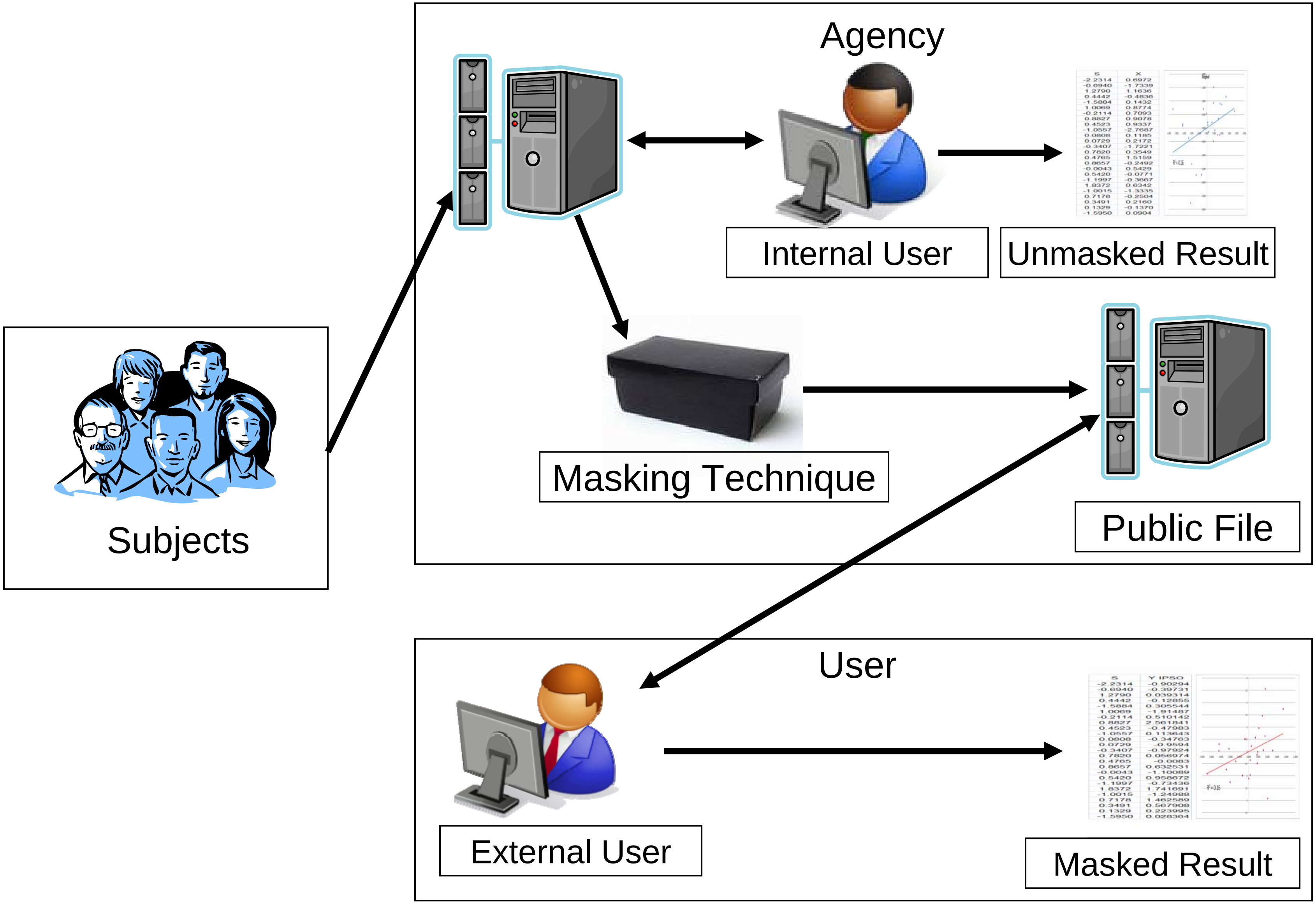
A THEORETICAL COMPARISON OF DATA MASKING TECHNIQUES FOR NUMERICAL MICRODATA

Krish Muralidhar
University of Kentucky

Rathindra Sarathy
Oklahoma State University



Subjects

Agency

Internal User

Unmasked Result

Masking Technique

Public File

User

External User

Masked Result

The Process

- Data is gathered from subjects
  - Subjects could be individuals or organizations
- The data is then organized by the Agency
  - Used for internal analysis by authorized users. Reports released by Agency
  - Data is masked
    - Masked data is then made available as a public file
- The external users analyze the public data file to produce their own reports

The Stakeholders

- Data Subjects
- Data Provider
  - The Agency
  - Also data gatherer
  - We use “Provider” because this is the function that we are interested in
    - The Agency may (and usually) has many other functions that it needs to perform
- External Data User

Types of Data

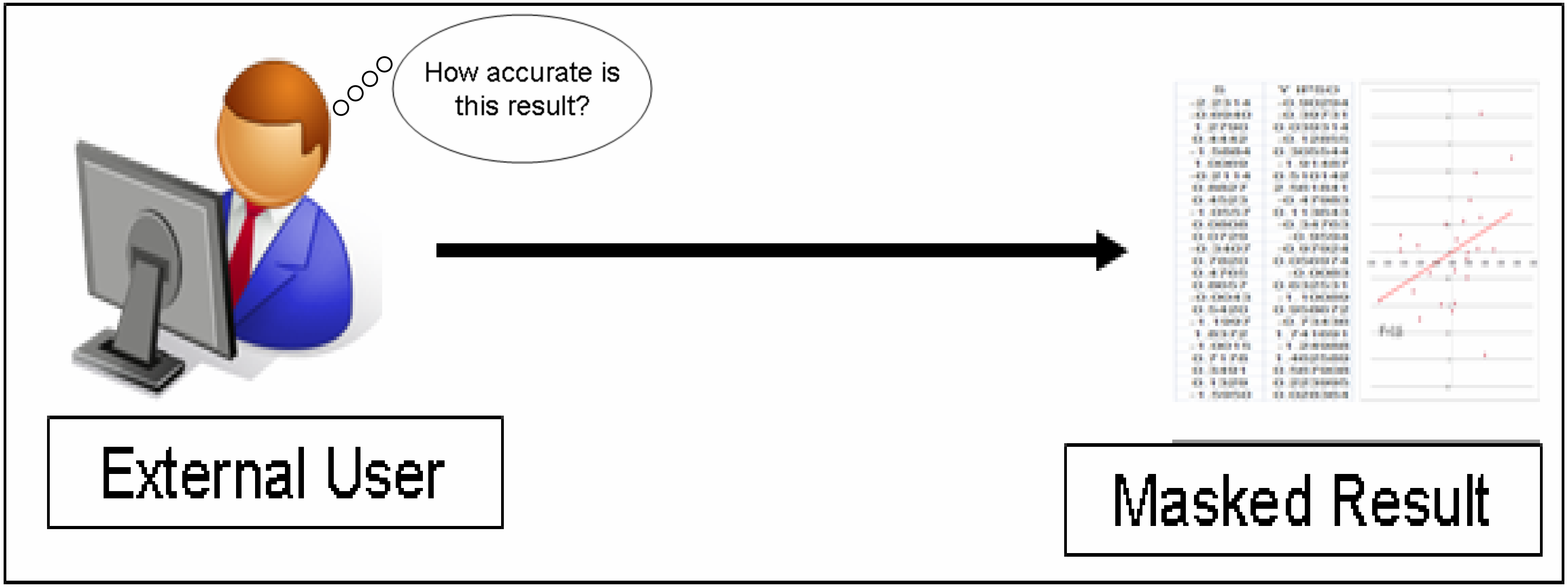
- Categorical data
- Numerical data
  - Discrete
  - Continuous
- The framework described in this study can be used for any type of data
  - The stakeholders will be the same, but the criteria used to evaluate the procedures may be different
- For the purposes of this presentation, we will focus on continuous numerical data

Subjects

- Primary interest is in protecting privacy and/or confidentiality of the Subjects. The Provider, in most cases, gives an explicit or implicit guarantee that the privacy and confidentiality of the Subjects will be protected.

User

- Primary interest is in the accuracy of the results obtained from the masked data. Currently receives very little information about the accuracy of the data



External User

Masked Result

Provider

- The provider has a dual responsibility
  - Responsible to ensure that the privacy and confidentiality of the data subjects is not violated
  - Responsible to ensure that the masked data provided to the users is accurate within some limits

Provider

- The provider attempts to balance the needs of the Subjects and the User
  - The needs of the Subjects are typically considered primary since explicit or implicit guarantees have been provided at the time of data collection

Some Basic Tenets

- The Provider will never intentionally provide incorrect information
- The Provider will always inform both the Subjects and the Users of the exact procedure that was used in data masking and all necessary information (including the masking parameters) that will allow the Subjects and Users to evaluate the efficacy of the procedure
- The provider will never choose a dominated technique

Dominated Technique

- If masking Method A
  - provides the same or lower information loss than Method B and **also** provides lower disclosure risk than Method B, then Method A dominates Method B
  - provides the same or lower disclosure risk than Method B and **also** provides lower information loss than Method B, then Method A dominates Method B
- In such cases, the Provider will always choose Method A

The Trade-off

- Disclosure Risk
  - Attempt to minimize (or reduce) disclosure risk

- Information Loss
  - Attempt to minimize (or reduce) information loss

- Selection of Masking Technique
  - Select the masking technique that best achieves both requirements

Other Stakeholder Interests

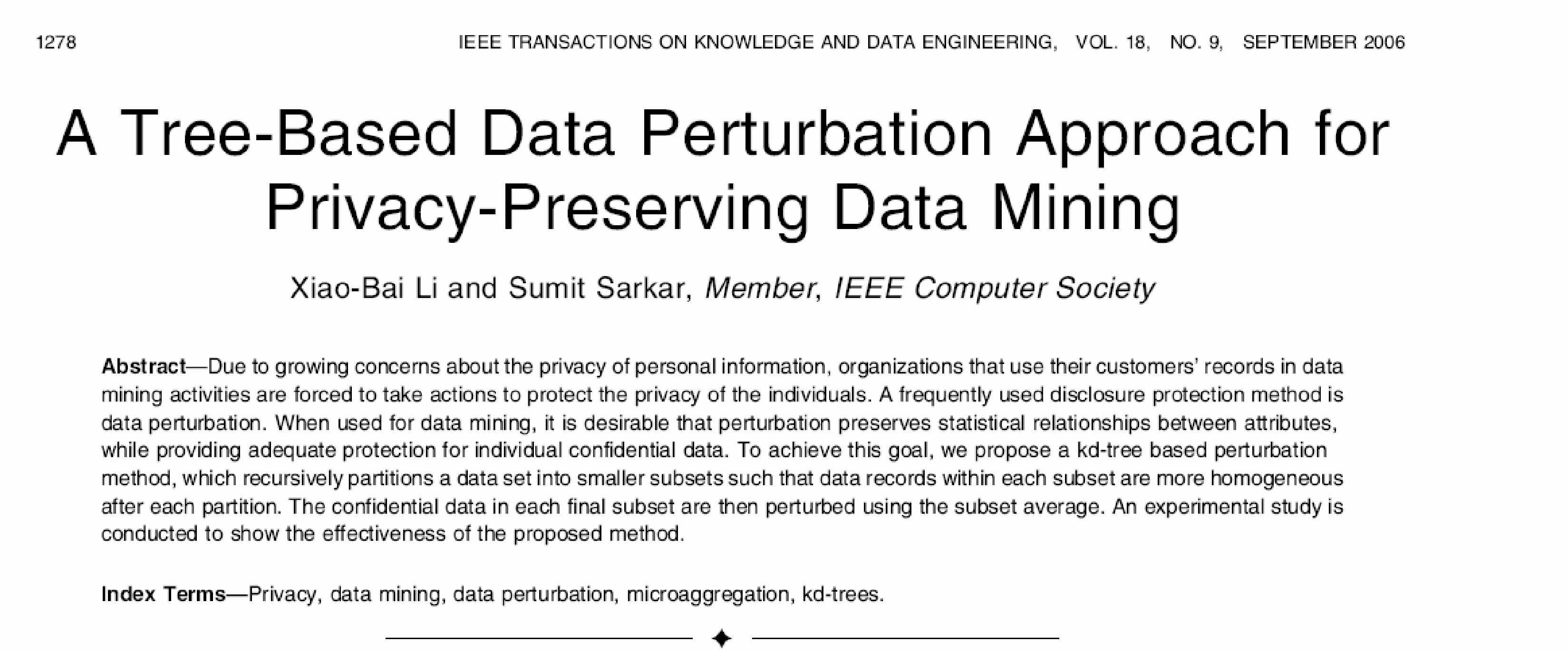
- While disclosure risk and information loss are two measures of interests, there are others
  - Ease of use
    - Ceteris paribus, the User would like to analyze the masked data exactly as they would analyze the original data. No additional effort should be required to analyze the masked data
  - Ease of implementation
    - Ceteris paribus, the Provider would like to use a masking technique that is easy to implement
      - It is true that once software becomes available, the implementation of any procedure is “easy”, we consider the complexity of the procedure in evaluating ease of implementation

Why Theoretical Evaluation?

- The evaluation of masking techniques is often performed empirically.
- It is not possible to generalize the results of such empirical evaluations.
  - Even for a given data set, with an empirical evaluation, it is not possible to provide specific assurances to the Subjects and User
- A theoretical evaluation will allow the Provider to provide specific assurances regarding both disclosure risk and information loss to the Subjects and the User

A Quick Example

Evaluating Tree Based Perturbation Method



1278 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 9, SEPTEMBER 2006

A Tree-Based Data Perturbation Approach for Privacy-Preserving Data Mining

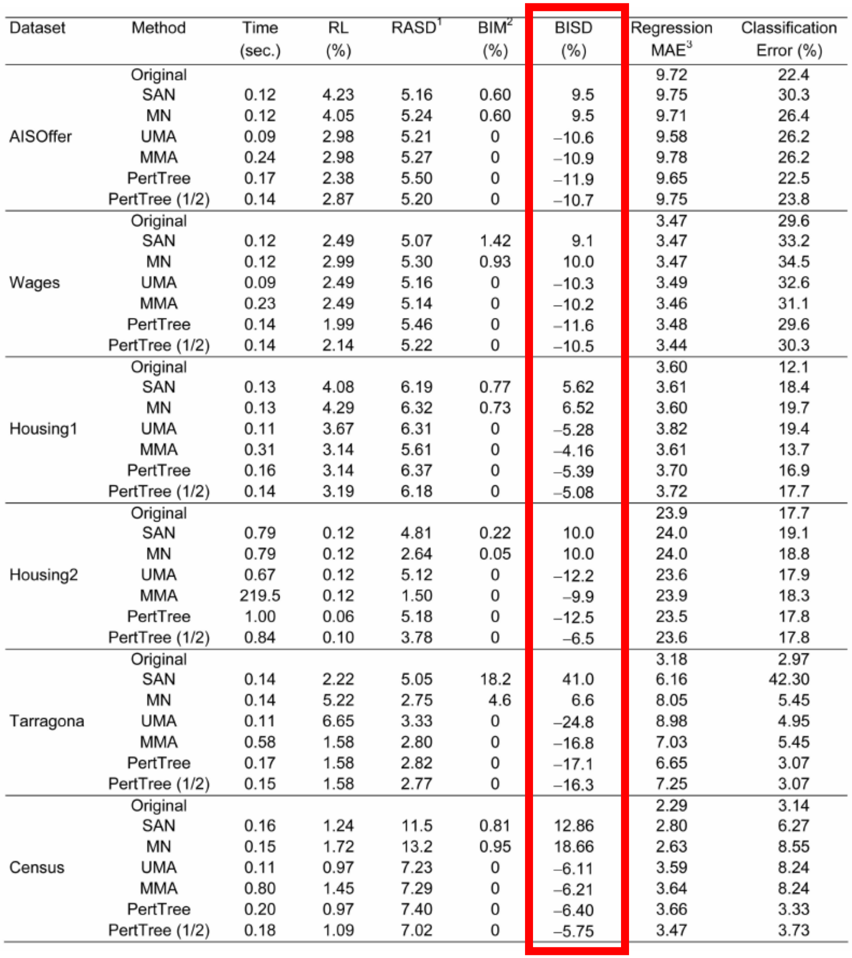
Xiao-Bai Li and Sumit Sarkar, *Member, IEEE Computer Society*

Abstract—Due to growing concerns about the privacy of personal information, organizations that use their customers' records in data mining activities are forced to take actions to protect the privacy of the individuals. A frequently used disclosure protection method is data perturbation. When used for data mining, it is desirable that perturbation preserves statistical relationships between attributes, while providing adequate protection for individual confidential data. 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. An experimental study is conducted to show the effectiveness of the proposed method.

Index Terms—Privacy, data mining, data perturbation, microaggregation, kd-trees.

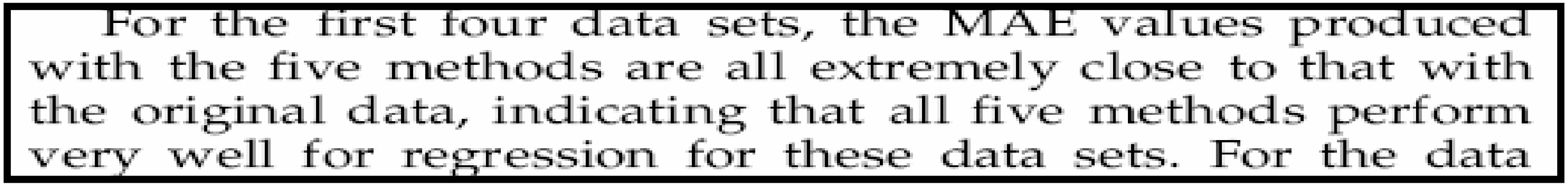
- The method was proposed by Li and Sarkar and involves splitting the data into “branches” and then micro-aggregating
- Compared to noise addition and micro-aggregation using 5 data sets

Empirical Evaluation



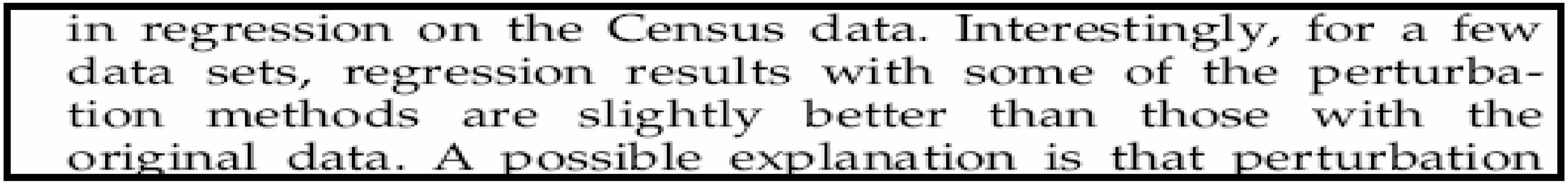
Dataset	Method	Time (sec.)	RL (%)	RASD[1]	BIM[2] (%)	BISD (%)	Regression MAE[3]	Classification Error (%)
AISOffer	Original						9.72	22.4
	SAN	0.12	4.23	5.16	0.60	9.5	9.75	30.3
	MN	0.12	4.05	5.24	0.60	9.5	9.71	26.4
	UMA	0.09	2.98	5.21	0	−10.6	9.58	26.2
	MMA	0.24	2.98	5.27	0	−10.9	9.78	26.2
	PertTree	0.17	2.38	5.50	0	−11.9	9.65	22.5
	PertTree (1/2)	0.14	2.87	5.20	0	−10.7	9.75	23.8
Wages	Original						3.47	29.6
	SAN	0.12	2.49	5.07	1.42	9.1	3.47	33.2
	MN	0.12	2.99	5.30	0.93	10.0	3.47	34.5
	UMA	0.09	2.49	5.16	0	−10.3	3.49	32.6
	MMA	0.23	2.49	5.14	0	−10.2	3.46	31.1
	PertTree	0.14	1.99	5.46	0	−11.6	3.48	29.6
	PertTree (1/2)	0.14	2.14	5.22	0	−10.5	3.44	30.3
Housing1	Original						3.60	12.1
	SAN	0.13	4.08	6.19	0.77	5.62	3.61	18.4
	MN	0.13	4.29	6.32	0.73	6.52	3.60	19.7
	UMA	0.11	3.67	6.31	0	−5.28	3.82	19.4
	MMA	0.31	3.14	5.61	0	−4.16	3.61	13.7
	PertTree	0.16	3.14	6.37	0	−5.39	3.70	16.9
	PertTree (1/2)	0.14	3.19	6.18	0	−5.08	3.72	17.7
Housing2	Original						23.9	17.7
	SAN	0.79	0.12	4.81	0.22	10.0	24.0	19.1
	MN	0.79	0.12	2.64	0.05	10.0	24.0	18.8
	UMA	0.67	0.12	5.12	0	−12.2	23.6	17.9
	MMA	219.5	0.12	1.50	0	−9.9	23.9	18.3
	PertTree	1.00	0.06	5.18	0	−12.5	23.5	17.8
	PertTree (1/2)	0.84	0.10	3.78	0	−6.5	23.6	17.8
Tarragona	Original						3.18	2.97
	SAN	0.14	2.22	5.05	18.2	41.0	6.16	42.30
	MN	0.14	5.22	2.75	4.6	6.6	8.05	5.45
	UMA	0.11	6.65	3.33	0	−24.8	8.98	4.95
	MMA	0.58	1.58	2.80	0	−16.8	7.03	5.45
	PertTree	0.17	1.58	2.82	0	−17.1	6.65	3.07
	PertTree (1/2)	0.15	1.58	2.77	0	−16.3	7.25	3.07
Census	Original						2.29	3.14
	SAN	0.16	1.24	11.5	0.81	12.86	2.80	6.27
	MN	0.15	1.72	13.2	0.95	18.66	2.63	8.55
	UMA	0.11	0.97	7.23	0	−6.11	3.59	8.24
	MMA	0.80	1.45	7.29	0	−6.21	3.64	8.24
	PertTree	0.20	0.97	7.40	0	−6.40	3.66	3.33
	PertTree (1/2)	0.18	1.09	7.02	0	−5.75	3.47	3.73

Some “Results”



For the first four data sets, the MAE values produced with the five methods are all extremely close to that with the original data, indicating that all five methods perform very well for regression for these data sets. For the data

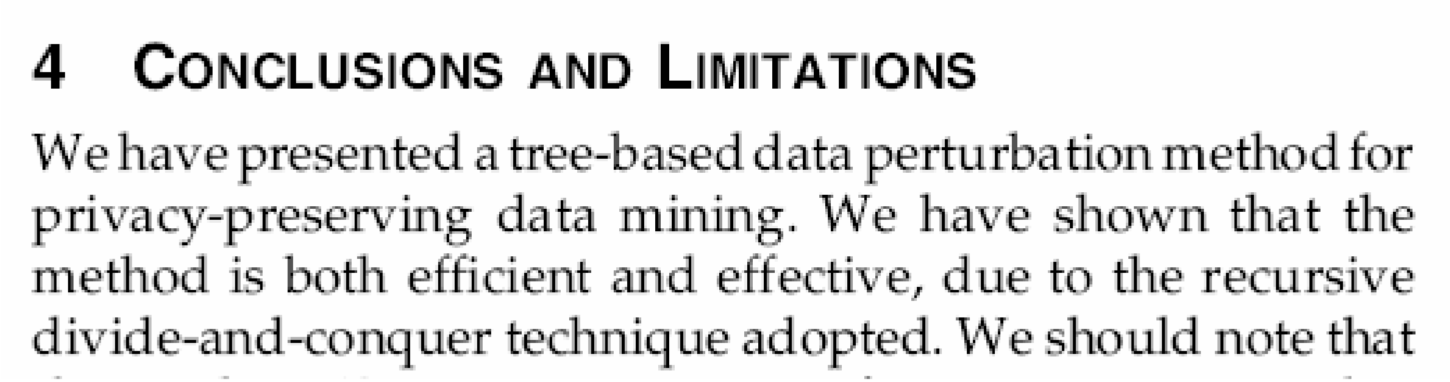
- The above statement explicitly states that all methods perform well for regression analysis in spite of the fact that we know that noise addition and micro-aggregation methods provide biased results for regression analysis (since the standard deviation of the masked data is biased for every method).



in regression on the Census data. Interestingly, for a few data sets, regression results with some of the perturbation methods are slightly better than those with the original data. A possible explanation is that perturbation

- The above statement actually makes the claim that the perturbed data provides “better results” than the original data! How is it possible for “perturbed data” to produce “better results” than the original data? And exactly what is “better results”?

Summary of the Comparison



4 CONCLUSIONS AND LIMITATIONS

We have presented a tree-based data perturbation method for privacy-preserving data mining. We have shown that the method is both efficient and effective, due to the recursive divide-and-conquer technique adopted. We should note that

- Based on this conclusion, we would assume that the tree based perturbation approach is an effective method for perturbing numerical data … after all it was published in the IEEE Transactions on Knowledge and Data Engineering
- Unfortunately, once the paper is published, it is difficult to refute the results
  - We tried …

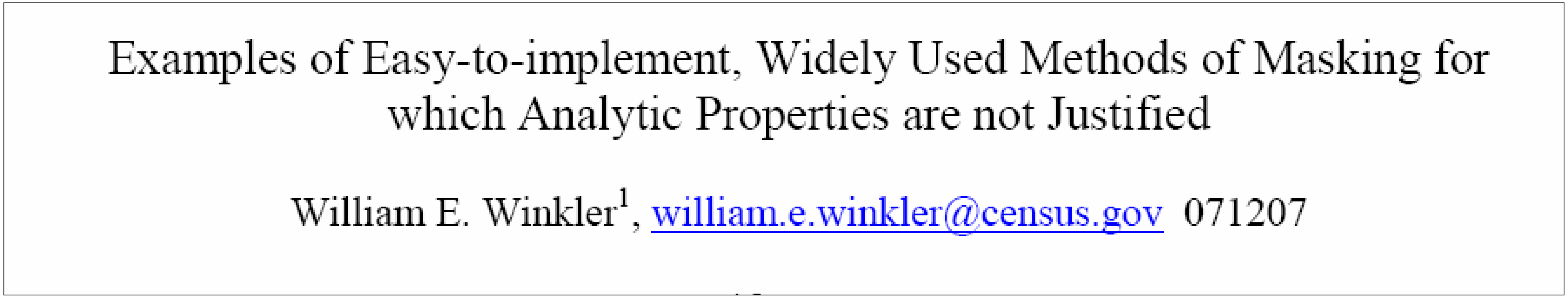
Theoretically

- We know the following about the Tree Based Perturbation Approach
  - The variance of the masked variable is lower than that of the original confidential variable
  - All relationships are altered
    - Unfortunately, unlike noise addition where it is possible to estimate the extent of the bias, it is not possible to evaluate the bias. We cannot even say the direction of the bias
      - Some relationships are attenuated
      - Some relationships are accentuated
  - Regression results are biased
  - Disclosure risk is not minimized
- The published paper makes no mention of theoretical characteristics of the procedure

Theoretical versus Empirical Evaluation

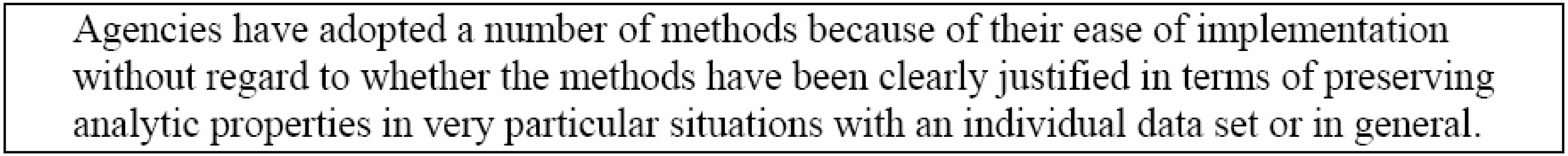
- A theoretical assessment assures the Provider that the data masking technique is capable of maintaining certain basic characteristics
- For a selected empirical data set, it may seem as if the procedure works effectively. The problem is that it may not work as well for other data sets
- Relying on the empirical “results” could have detrimental impact in other applications
- Even if you decide to use this approach, a theoretical evaluation will at least tell you exactly what the procedure is capable of so that you can let the users know of the capabilities of the procedure

Further Reasons

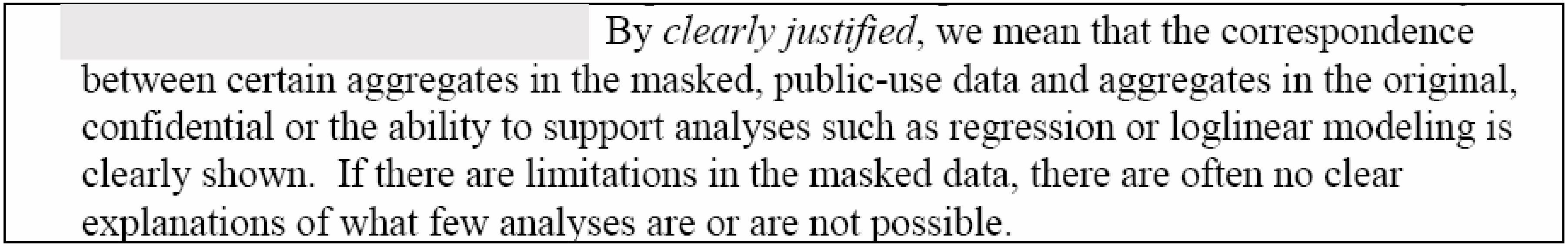


Examples of Easy-to-implement, Widely Used Methods of Masking for which Analytic Properties are not Justified

William E. Winkler[1], william.e.winkler@census.gov 071207

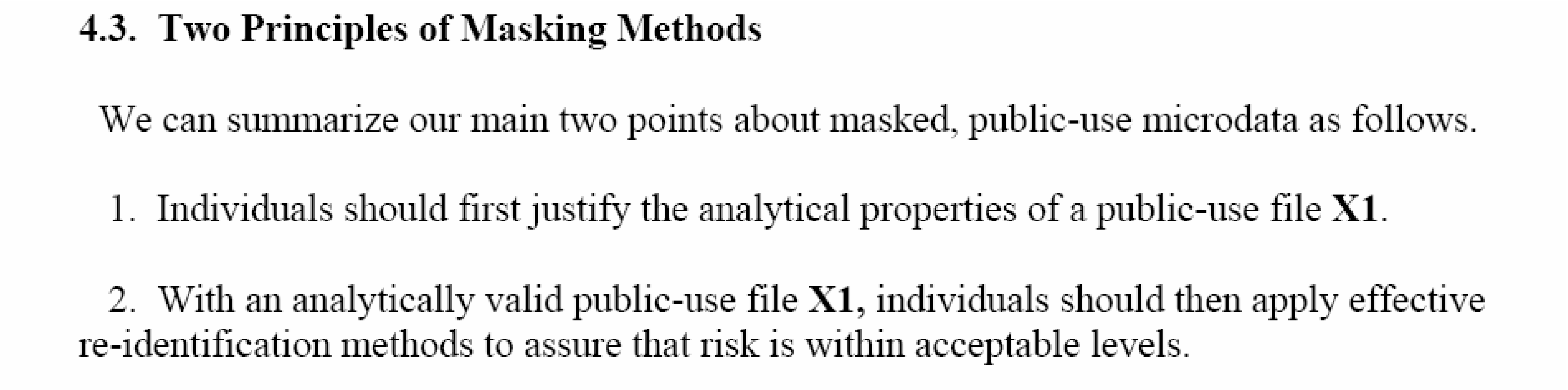


Agencies have adopted a number of methods because of their ease of implementation without regard to whether the methods have been clearly justified in terms of preserving analytic properties in very particular situations with an individual data set or in general.



By *clearly justified*, we mean that the correspondence between certain aggregates in the masked, public-use data and aggregates in the original, confidential or the ability to support analyses such as regression or loglinear modeling is clearly shown. If there are limitations in the masked data, there are often no clear explanations of what few analyses are or are not possible.

Winkler (2007) … Continued



4.3. Two Principles of Masking Methods

We can summarize our main two points about masked, public-use microdata as follows.

1. Individuals should first justify the analytical properties of a public-use file **X1**.

2. With an analytically valid public-use file **X1,** individuals should then apply effective re-identification methods to assure that risk is within acceptable levels.

- The objective of this presentation is to provide a theoretical framework for evaluating masking methods for numerical data
  - The general framework can be used for other masking methods as well

A Framework for a Theoretical Evaluation

- In the following, we provide a framework for evaluating the characteristics of data masking techniques for numerical data
  - We intend this to be the first step in a formal evaluation process
  - Our objective is to at least provide a basic set of characteristics that must be assessed
  - In specific situations, it may be necessary to add new pertinent characteristics

Assessing Disclosure Risk

- The total disclosure risk resulting from the releasing the data can be partitioned as
  - Disclosure risk resulting from releasing non-confidential microdata
  - Disclosure risk resulting from releasing aggregate information regarding confidential numerical variables (but no microdata)
  - Disclosure risk resulting from releasing masked microdata

Disclosure Risk of Masking Technique

- In assessing the disclosure risk resulting from the masking technique, it is necessary to **isolate** the risk that results from releasing the masked microdata
  - The other two aspects of disclosure risk cannot be attributed to the masking technique

The Data Release Process

- Release non-confidential data
- Release aggregate information regarding numerical confidential variables
  - Characteristics of the marginal distribution of individual variables
    - mean, median, mode, variance, skewness, kurtosis, and best fit distribution
  - Covariance matrix (between all variables)
  - Rank order correlation matrix (between all variables)
- Release masked numerical microdata

Stepwise Assessment of Disclosure Risk

- Assess disclosure risk at the first step (releasing non-confidential microdata)
- Assess disclosure risk at the second step (after release of aggregate information)
- Assess disclosure risk after release of masked microdata
- The disclosure risk due to the masking technique is the difference in the disclosure risk between the last two steps
  - At each step, it may be necessary to revise what is to be released based on the assessment of disclosure risk

Dalenius' definition of Disclosure Risk

- Dalenius defines disclosure to have occurred when the release of data allows the user to improve the estimate of an unknown confidential value
  - Microdata generated using some masking techniques prevent the intruder from improving the estimate of an unknown confidential value
  - Thereby, these technique minimize disclosure risk resulting from release of the microdata

Types of Disclosure Risk

- Identity disclosure
  - The ability to identify a particular record as belonging to a particular individual using the released data
- Value disclosure
  - The ability to predict the value of a confidential variable for a particular record using the released data

Evaluation of Disclosure Risk Measures

- In terms of disclosure risk, a masking technique
  - Minimizes the risk of disclosure according to Dalenius’ definition
  - Does not minimize the risk of disclosure
    - For comparing techniques that fall under this category, it would be necessary to use empirical measures of disclosure risk

Information Loss

- Ideally, there would be no information loss if, for any query, the response to the query using the masked data is exactly the same as that using the original data
  - Impossible to achieve in practice since a random query may involve the confidential value for a single record. If the response to this query is exactly the same as that original record, this results in completely disclosure

Types of Analysis

- In most cases, the released data is likely to be analyzed using statistical or data mining techniques
  - Analysis at the aggregate level rather than the individual record level
  - What is aggregate?
    - 2 records?
    - 5 records?

Information Loss … Modified

- The information loss represents the extent to which the results from the masked data are different from the results using the original data for aggregate analysis involving statistical or data mining techniques
  - Parametric statistical analysis
  - Non-parametric statistical analysis
  - Data mining

Assessing Information Loss

- At the aggregate level, we can assume that there is no information loss (at least asymptotically) if the joint distribution of the (entire) released data is the same as that of the original data

Asymptotically?

- Responses to individual queries using the masked data may be slightly different from the response using the original data but is unbiased.
- The difference between the two responses approaches zero as the size of the query set increases
  - When the response is biased, the difference in the responses between the masked and original **will not** approach zero as the size of the query set increases; it will approach the true bias

Measuring Information Loss

- We use the following characteristics for the purposes of this presentation
  - Marginal distribution
  - Relationships
    - Linear
    - Monotonic
    - Non-monotonic
  - Sub-group characteristics

Other Characteristics

- We strongly recommend adding other characteristics that are relevant in the general or special case
  - Please feel free to make recommendations in this regard
- It is important that we select only characteristics which provide some information about the underlying data set
  - Should include characteristics of the data … not specific measures of information loss

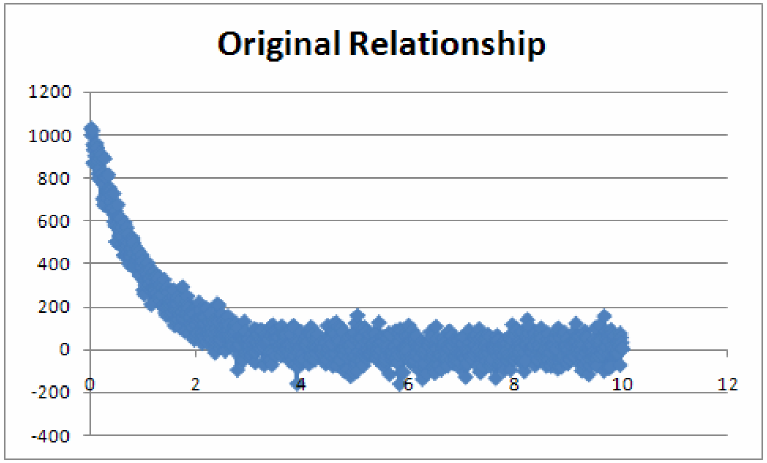
Evaluation

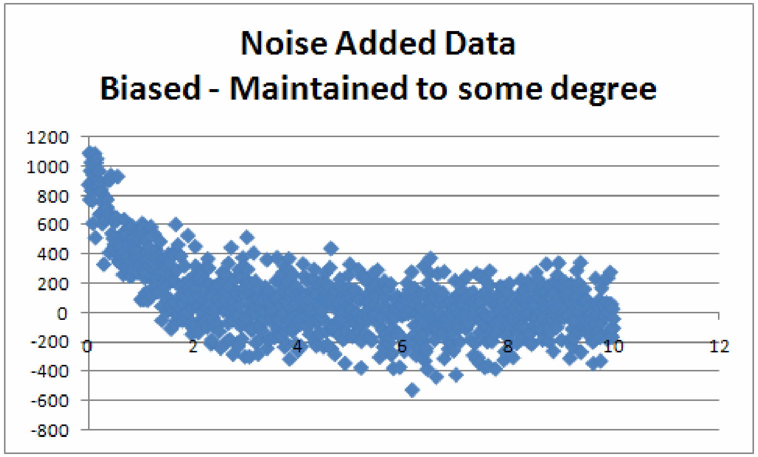
- For each information loss characteristic, the masking technique may
  - Maintain the characteristic exactly
  - Maintain the characteristic asymptotically
  - The characteristic is modified
    - Results in biased estimates of the characteristic
      - Empirical assessment would be necessary
    - Does not maintain the characteristic

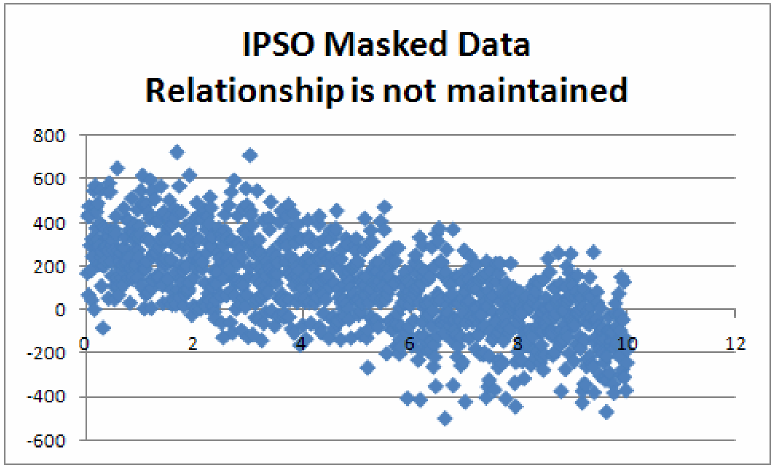
Difference between “Biased” and “Not Maintained”

- The term “Not maintained” indicates that a particular characteristic which existed in the original data WILL NOT exist in the masked data
  - Non-linear relationships in the original data will not be maintained by masked data generated using GADP, IPSO, or multiple imputation
- The term “Biased” indicates that a particular characteristic will be modified or changed, but not completely eliminated
  - Non-linear relationships in the original data will be attenuated, but not completely eliminated by data generated using noise addition

Example


Original Relationship


Noise Added Data
Biased - Maintained to some degree


IPSO Masked Data
Relationship is not maintained

- IPSO does not maintain the non-linear relationship
- Noise addition provides biased result. The extent of the bias varies by perturbation level
  - In some cases, we can estimate the direction (higher or lower) and extent of bias. But this is not always possible.

Conditional Distribution Approach

- Identify the joint distribution (f(S,X)) of the non-confidential (S) and confidential (X) variables
- Compute the conditional distribution f(X|S)
- For each record, generate the masked values y_i using $f(X|S = s_i)$
- Then the joint distribution of (S and Y) is the same as that of (S and X)
  - f(S,Y) = f(S,X)
  - Little or no information loss since the joint distribution of the original and masked data are the same

Disclosure Risk of CDA

- When the masked data is generated using CDA, it can be verified that $f(X|Y,S,A) = f(X|S,A)$
  - Releasing the masked microdata Y does not provide any new information to the intruder over and above the non-confidential variables S and A (the aggregate information regarding the joint distribution of S and X)

CDA is the answer … But

- The CDA approach results in very low information loss and minimizes disclosure risk and represents a complete solution to the data masking problem
- Unfortunately, in practice
  - Identifying f(S,X) may be very difficult
  - Deriving f(X|S) may be very difficult
  - Generating y_i using f(X|S) may be very difficult
- In practice, it is unlikely that we can use the conditional distribution approach

Relationship between masked and original data

- It is important to note that when the masked data is closely related to the original data
  - Disclosure risk is increased. Disclosure risk is inversely related to the correlation between the original and masked data
  - Information loss may not be affected. Just because the masked data is in closely related to the original data does not necessarily mean that the characteristics of the data set are maintained
    - We could add very little noise, but it would still result in biased estimates of the different characteristics
    - The masked data could be unrelated (independent) to the original data, but we can still maintain the characteristics of the data set

Masking Techniques for Numerical Data

- Broadly classified as
  - Model based techniques
    - Copula based data perturbation, General additive data perturbation, Information preserving statistical obfuscation, Multiple imputation, Data shuffling
  - Non-model based techniques
    - Simple noise addition, Sufficiency based noise addition, Micro-aggregation, Data swapping, other approaches

Model Based Approaches

- Model based approaches for data masking essentially attempt to model the data set by using an assumed $f^*(S,X)$ for the joint distribution of (S and X), derive $f^*(X|S)$, and generate the masked values from this distribution.
  - The masked data $f(S,Y)$ will have the joint distribution $f^*(S,X)$ rather than the true joint distribution $f(S,X)$
  - If the data is generated using $f^*(X|S)$ then the masking procedure minimizes disclosure risk since $f(X|Y,S,A) = f(X|S,A)$

Model Based Masking Methods

- General additive data perturbation
- Information preserving statistical obfuscation
- Multiple imputation
- Copula based perturbation
- Data shuffling
- PRAM

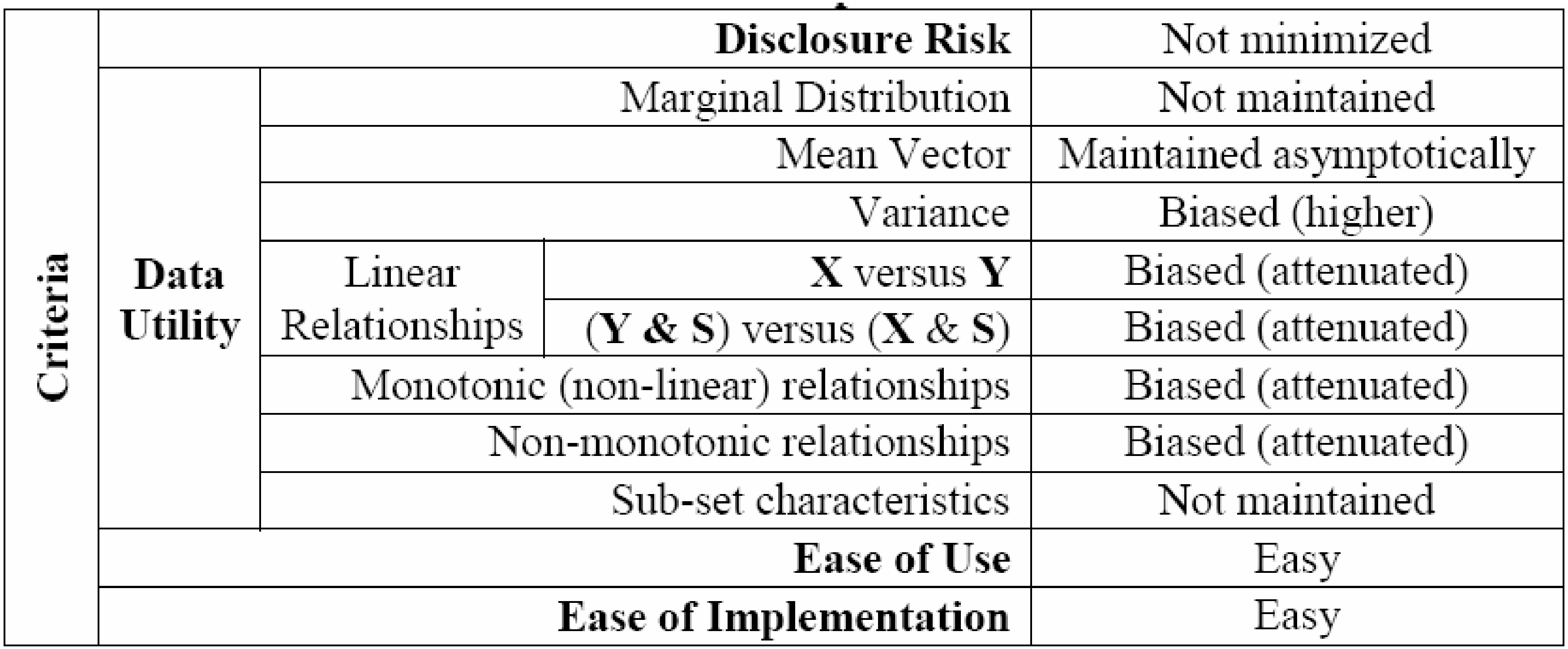
Non-Model Based Methods

- Noise addition
  - Simple, Kim’s method, Sufficiency based
- Micro-aggregation
  - Univariate
  - Multivariate
- Data swapping
- Other approaches
  - Rounding, truncation

Some simple illustrations

- In the following sections we illustrate the application of this framework to evaluate different data masking techniques

A Theoretical Evaluation of Simple Noise Addition

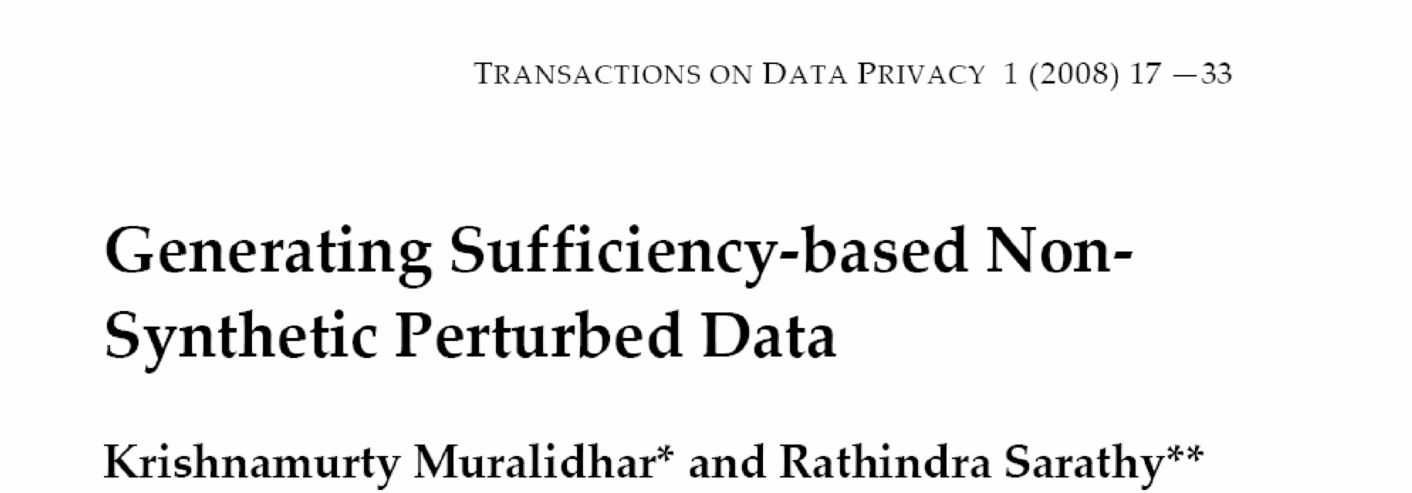


Criteria	**Disclosure Risk**			Not minimized
	Data Utility	Marginal Distribution		Not maintained
		Mean Vector		Maintained asymptotically
		Variance		Biased (higher)
		Linear Relationships	**X** versus **Y**	Biased (attenuated)
			(**Y & S**) versus (**X** & **S**)	Biased (attenuated)
		Monotonic (non-linear)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

No Data Necessary!

- We performed an evaluation of simple noise addition without using any data
- The results that we identified will hold for any and every data set
  - If we specify the level of noise, we can even identify the extent of bias that will result
- The only subjective aspect of this assessment are the two secondary criteria (ease of use and ease of implementation)
  - But we consider these only when two techniques have similar disclosure risk and information loss characteristics

Sufficiency Based Noise Addition



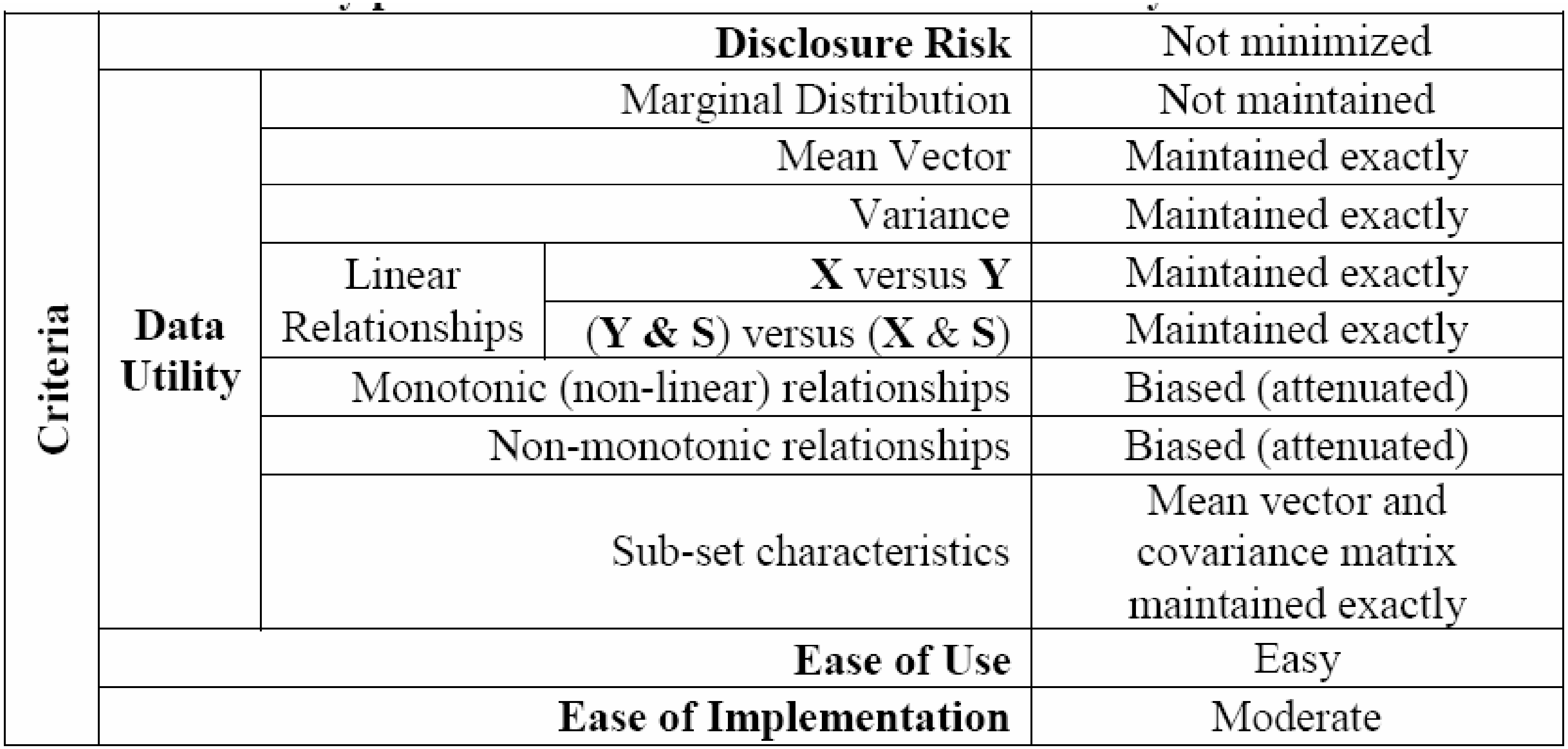
TRANSACTIONS ON DATA PRIVACY 1 (2008) 17 —33

Generating Sufficiency-based Non-Synthetic Perturbed Data

Krishnamurty Muralidhar* and Rathindra Sarathy**

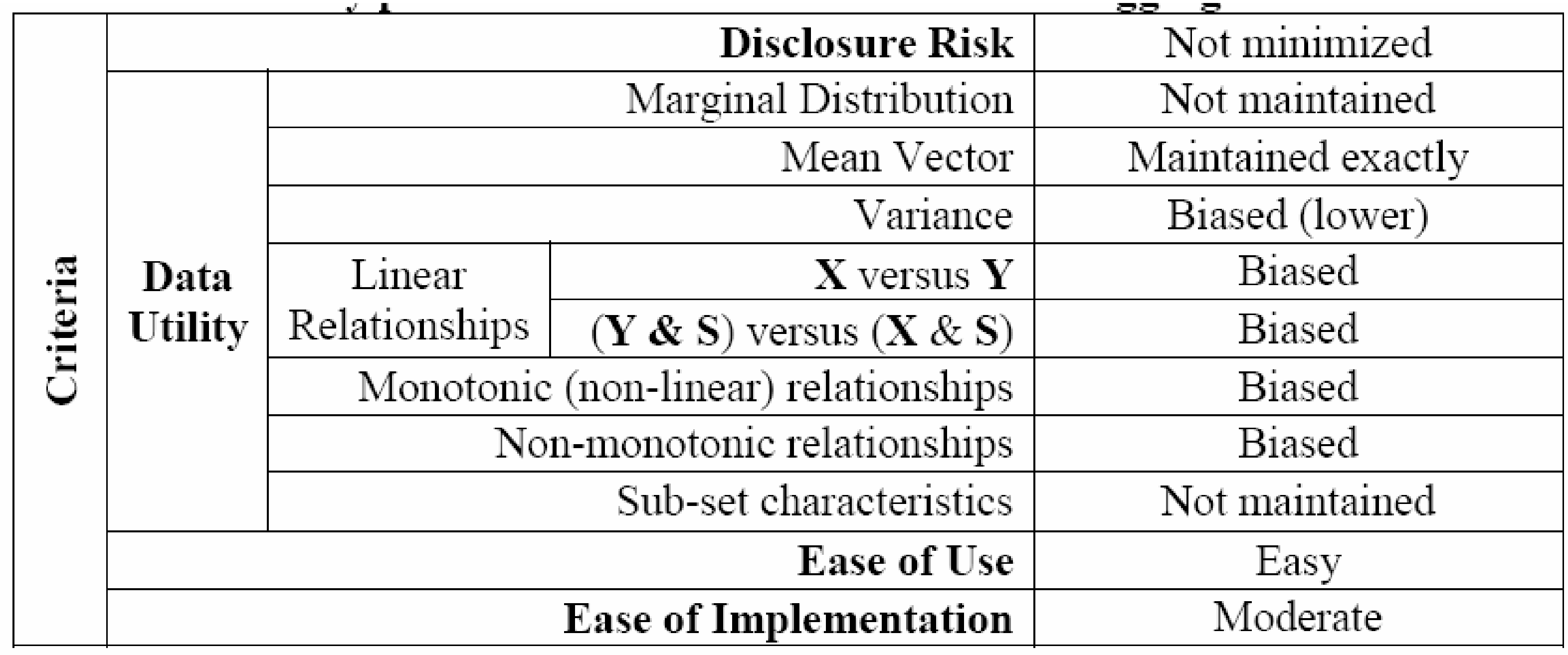
- A new noise addition method that is capable of providing exact results for many traditional statistical analyses
  - To be covered in detail later today

A Theoretical Evaluation 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

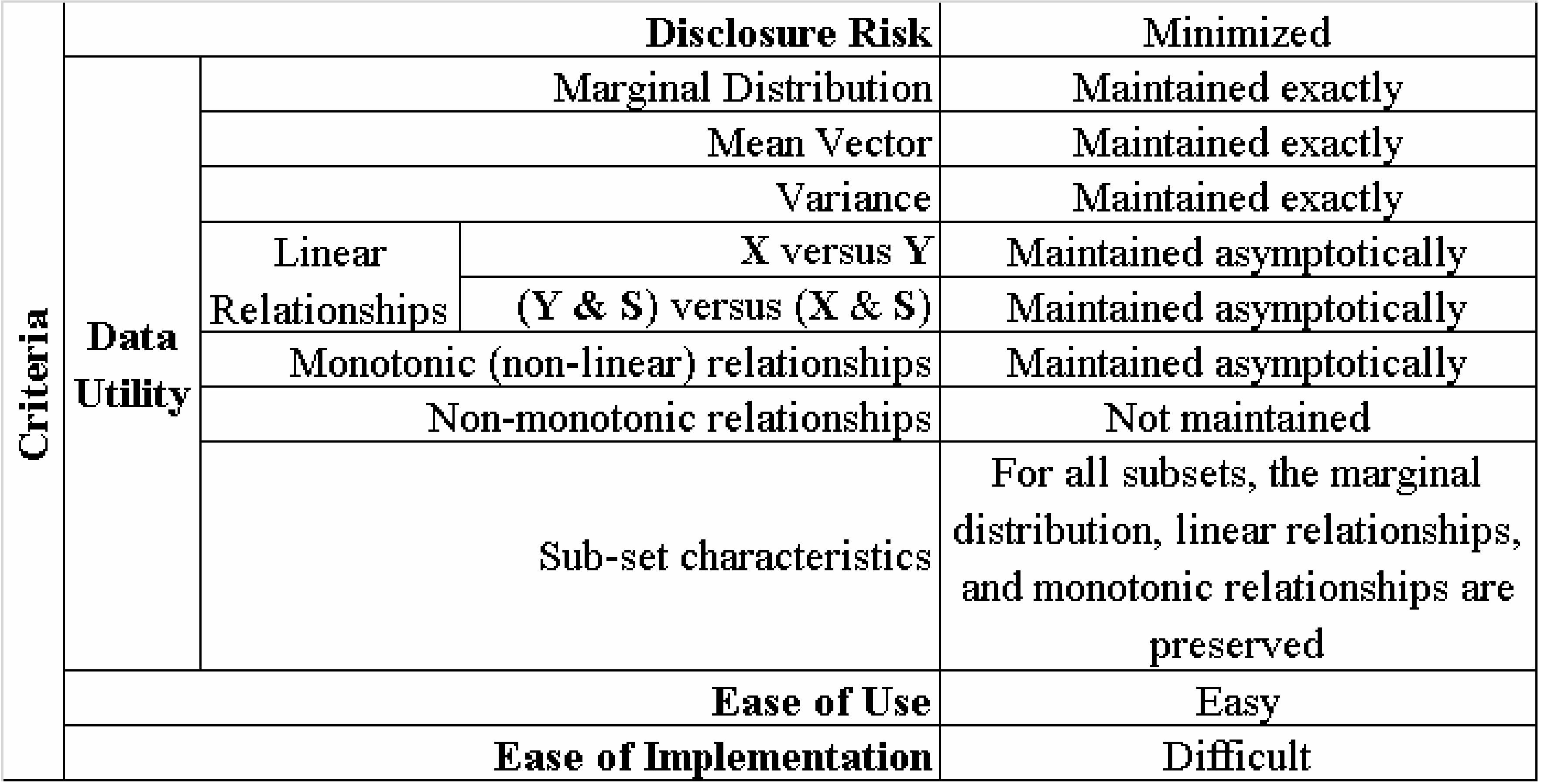
A Theoretical Evaluation of Micro-aggregation



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

- Unlike noise addition, with micro-aggregation it is difficult to determine the direction in which bias of relationship occurs. In some cases, it is attenuated and in other cases, it is accentuated
  - Implementing multivariate micro-aggregation requires the Provider to make several decisions regarding the techniques. Hence, we consider the ease of implementation for this technique to be moderate. Univariate microaggregation is easy to implement.

Theoretical Evaluation 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	For all subsets, the marginal distribution, linear relationships, and monotonic relationships are preserved
			Ease of Use	Easy
			Ease of Implementation	Difficult

Comparative evaluation of Multiple Methods

- The framework was developed with the intention of being used for comparing different methods of data masking
- If one method dominates another, then the decision is simple
- If one method does not dominate another, then additional (subjective or empirical) evaluation may be necessary

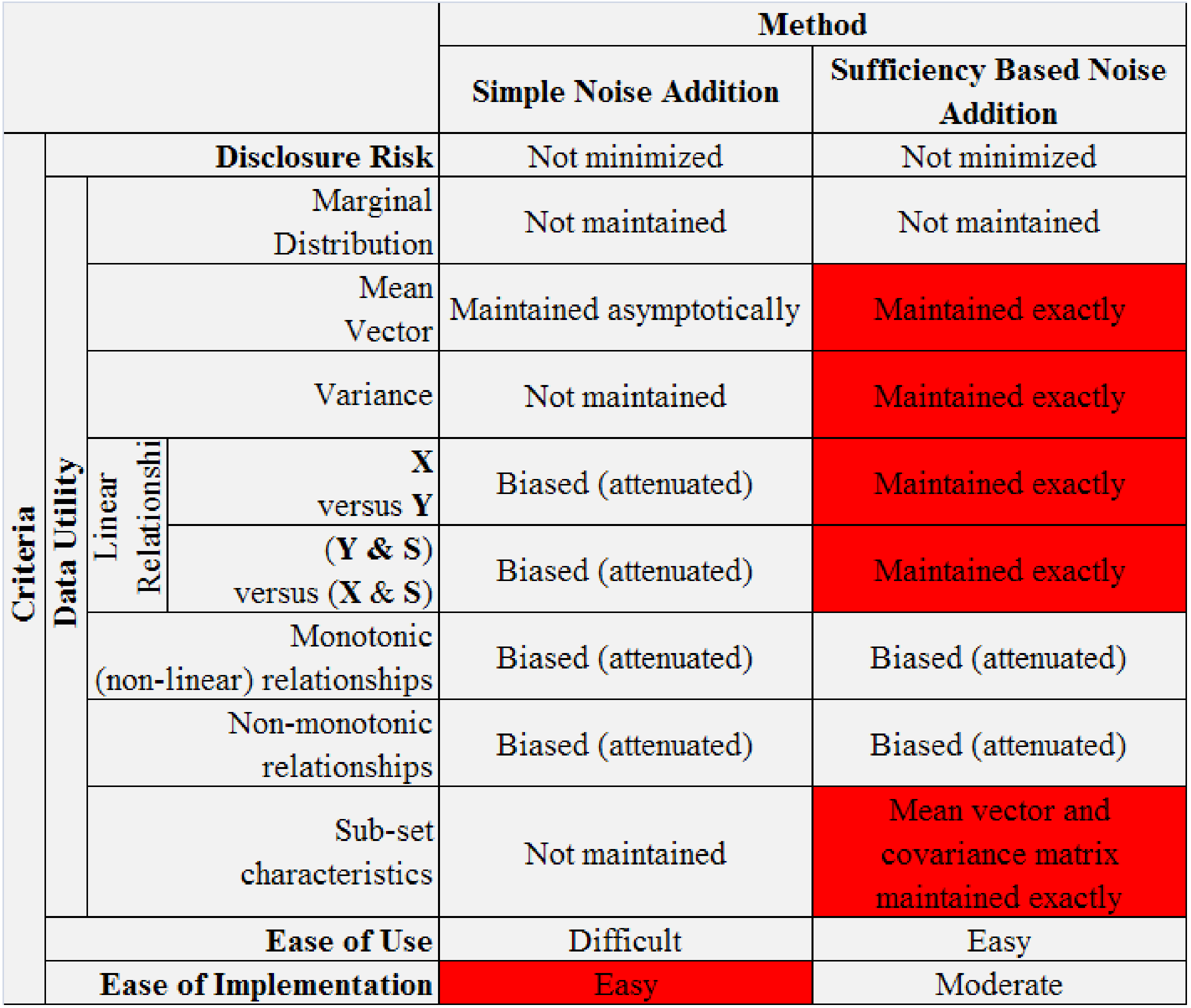
Comparing Information Loss

- Maintaining a characteristic exactly is better than maintaining a characteristic asymptotically
- Maintaining a characteristic asymptotically is better than modifying a characteristic
- Within modifying a characteristic
  - Are biased estimates of the characteristic necessarily better than not maintaining the characteristic especially if we cannot explicitly assess the extent of the bias?
    - At least when the characteristic is not maintained, we can clearly indicate this to the users
  - For our presentation, we assume that “biased estimates” are “slightly better” than “not maintaining” the characteristic, but not as strongly as the difference between maintaining or modifying the characteristic

Ease of Use and Implementation

- Note that the “Ease of Use” and “Ease of Implementation” should be considered only if two techniques offer the same level of disclosure risk and information loss performance
- In our opinion, and that of Winkler (2007), it is inappropriate for a Provider to choose a technique simple because it is easier to implement.

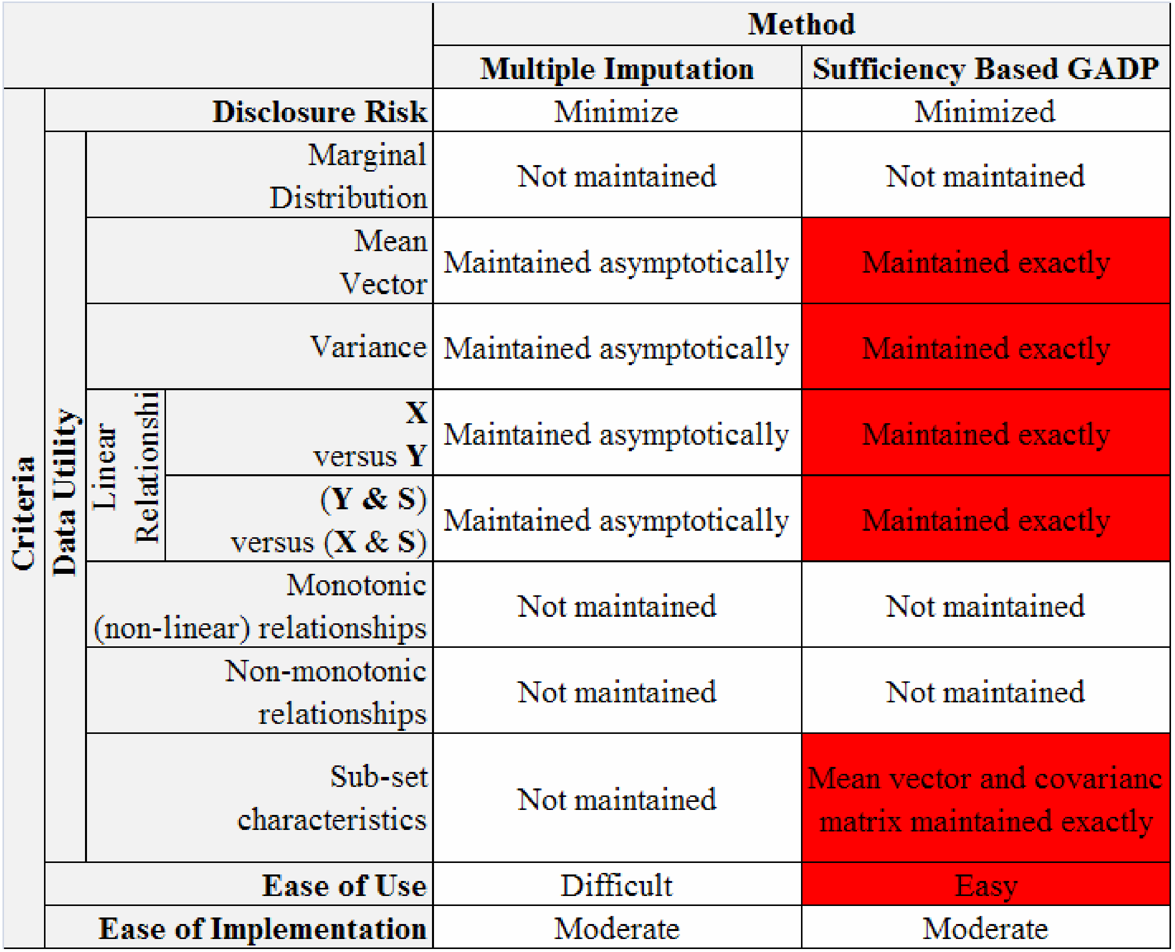
Comparison of Simple versus Sufficiency Based Noise Addition



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

- The sufficiency based method dominates the performance of simple noise for all Disclosure risk and Data utility criterion.
- If our objective is to provide secure data of high quality, then the sufficiency based approach should be used
  - The only reason to choose Simple noise addition is because it is easy to implement
  - This should always be a secondary criterion, not the primary criterion

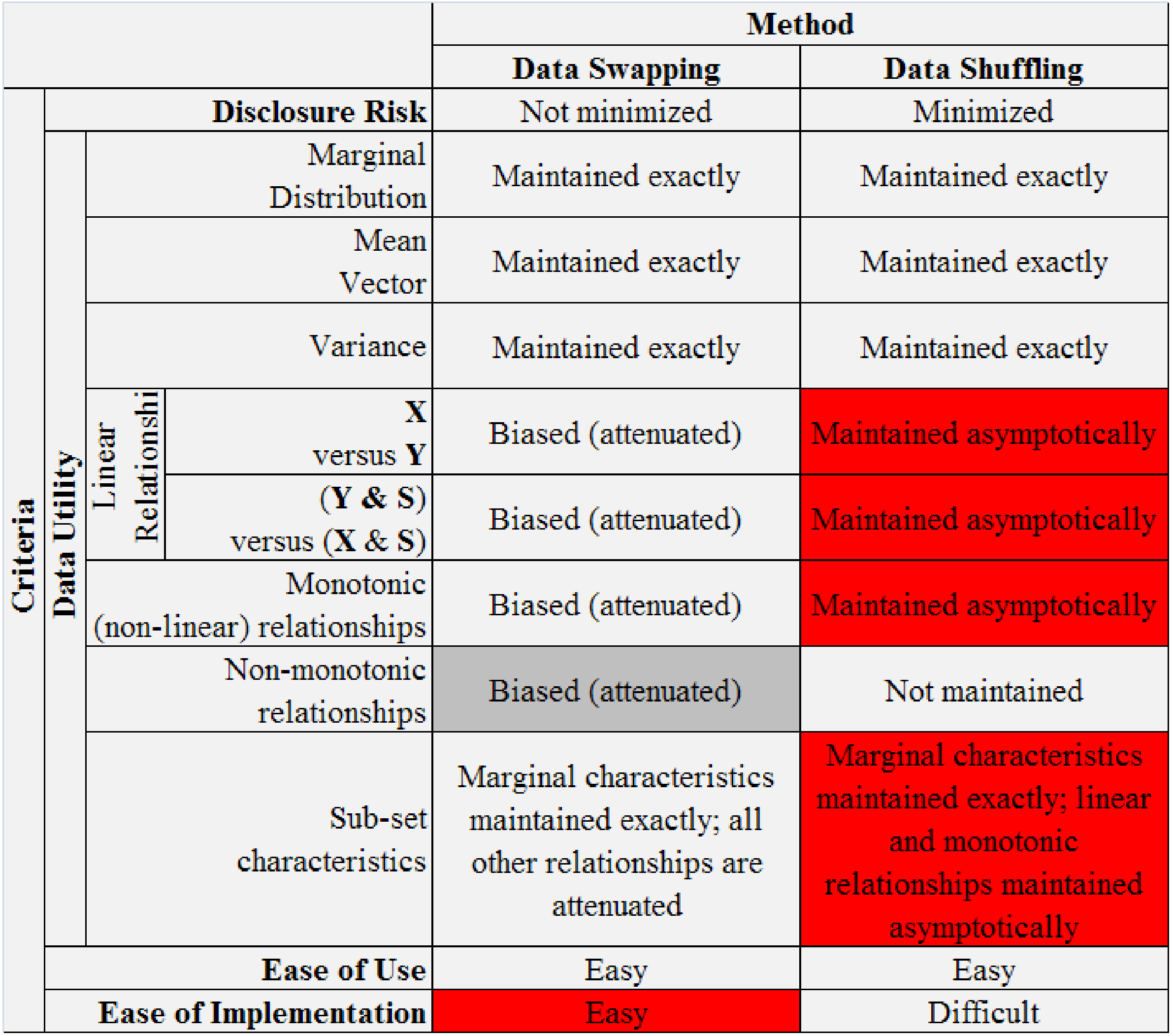
Comparison of Multiple Imputation and Sufficiency based GADP



				Method	
				Multiple Imputation	Sufficiency Based GADP
Criteria	Disclosure Risk			Minimize	Minimized
	Data Utility	Marginal Distribution		Not maintained	Not maintained
		Mean Vector		Maintained asymptotically	Maintained exactly
		Variance		Maintained asymptotically	Maintained exactly
		Linear Relationshi	**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	Mean vector and covarianc matrix maintained exactly
	Ease of Use			Difficult	Easy
	Ease of Implementation			Moderate	Moderate

- Sufficiency based GADP dominates the performance of multiple imputation
- We classify multiple imputation as "difficult to use" since it requires the user to analyze many data sets. Even if automated, for large data sets, this would be a drawback

Comparison of Data Shuffling and Data Swapping

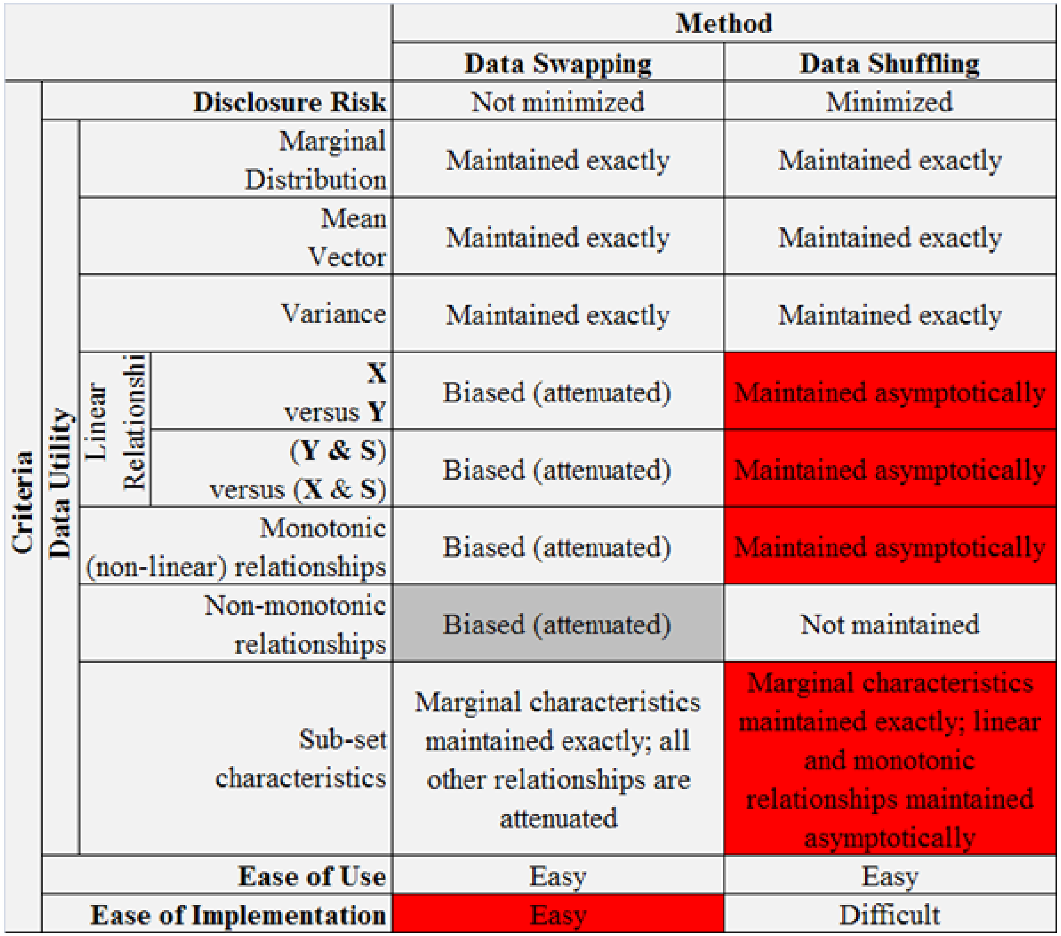


				Method	
				Data Swapping	**Data Shuffling**
Criteria			**Disclosure Risk**	Not minimized	Minimized
	Data Utility		Marginal Distribution	Maintained exactly	Maintained exactly
			Mean Vector	Maintained exactly	Maintained exactly
			Variance	Maintained exactly	Maintained exactly
		Linear Relationshi	**X** 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

Comments

- Data shuffling maintains more characteristics than rank based Data swapping and also provides lower disclosure risk than Data swapping
- The only information loss criterion where Data swapping could be considered “slightly better” Data shuffling is in maintaining non-monotonic relationships
  - Data shuffling would not maintain non-monotonic relationships while Data swapping would result in biased estimates of such relationships
  - What is the extent of the bias?
    - That would depend on proximity of the swapped values
- Data shuffling is more difficult to implement than Data swapping

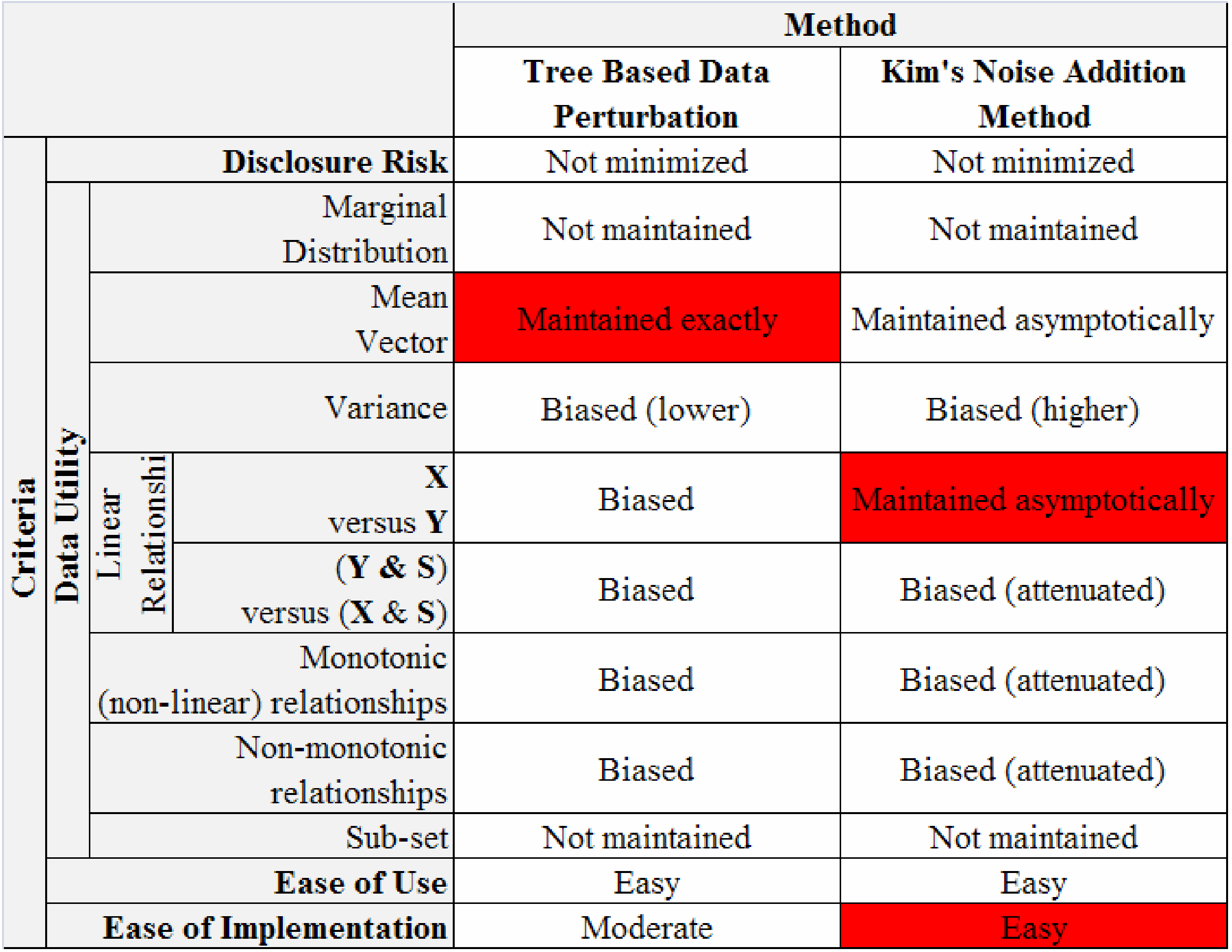
Shuffling versus Swapping …



				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 Relationshi	**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

- Which one do you think should be used?
  - Data shuffling minimizes disclosure risk
  - Data shuffling will not maintain non-monotonic relationships, but will provide unbiased results for all other analyses
  - Data swapping will result in biased estimates for **every** analysis performed on the masked data
- More importantly, which one do you think the **Subjects and Users** would prefer?

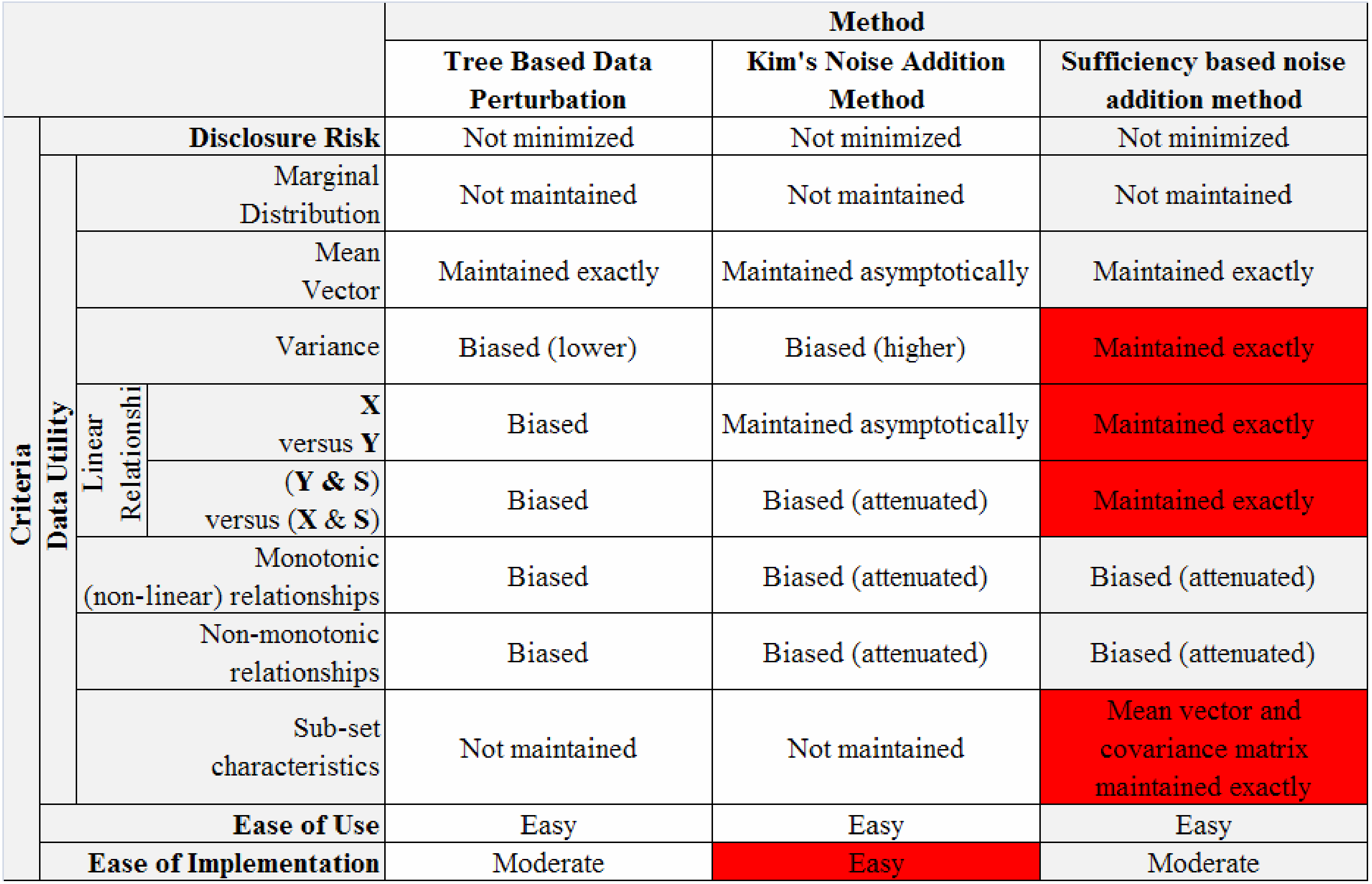
Comparison of Tree Based Data Perturbation & Kim's Method



				Method	
				Tree Based Data Perturbation	**Kim's Noise Addition Method**
Criteria			**Disclosure Risk**	Not minimized	Not minimized
	Data Utility		Marginal Distribution	Not maintained	Not maintained
			Mean Vector	Maintained exactly	Maintained asymptotically
			Variance	Biased (lower)	Biased (higher)
		Linear Relationshi	**X** versus **Y**	Biased	Maintained asymptotically
			(Y & S) versus **(X & S)**	Biased	Biased (attenuated)
			Monotonic (non-linear) relationships	Biased	Biased (attenuated)
			Non-monotonic relationships	Biased	Biased (attenuated)
			Sub-set	Not maintained	Not maintained
	Ease of Use			Easy	Easy
	Ease of Implementation			Moderate	Easy

- Theoretically, Kim's method is superior to Tree based perturbation
  - Unlike the other situations, one method does not dominate

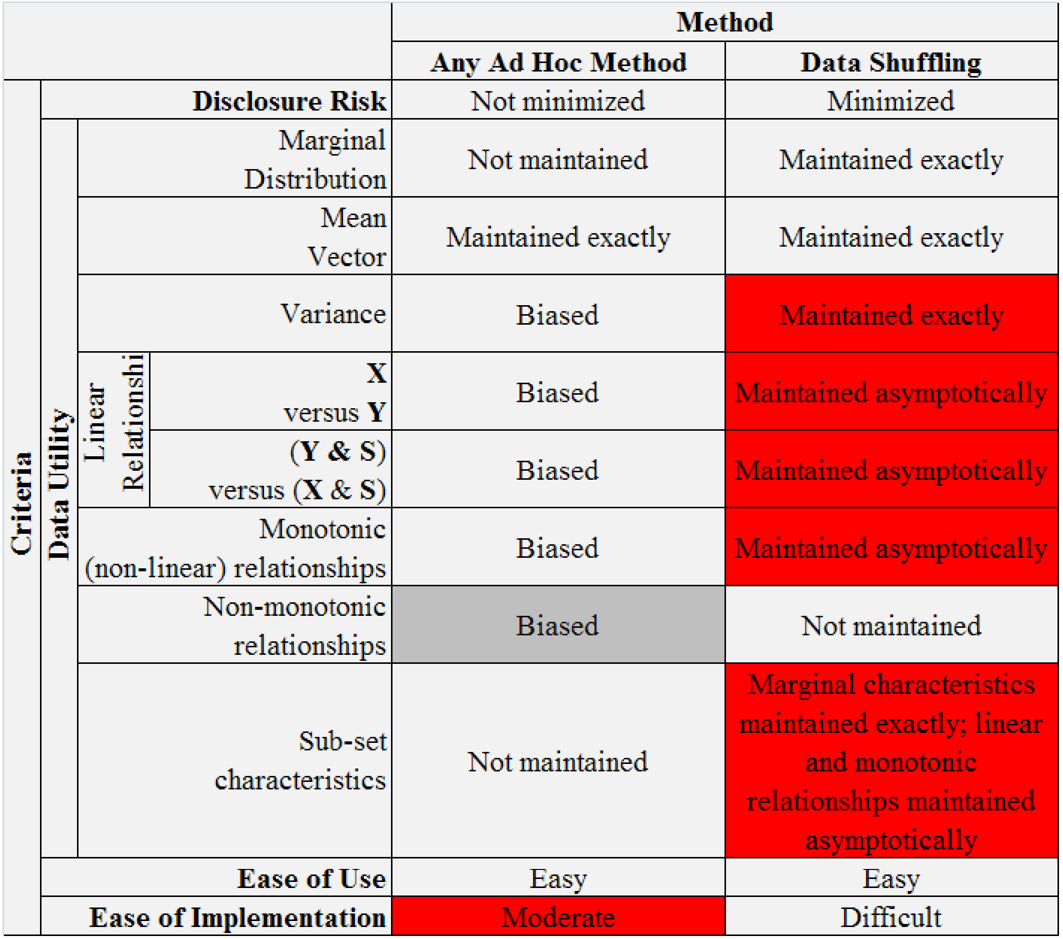
Comparison of Tree Based Data Perturbation and Other Methods



				Method		
				Tree Based Data Perturbation	**Kim's Noise Addition Method**	**Sufficiency based noise addition method**
Criteria			**Disclosure Risk**	Not minimized	Not minimized	Not minimized
	Data Utility		Marginal Distribution	Not maintained	Not maintained	Not maintained
			Mean Vector	Maintained exactly	Maintained asymptotically	Maintained exactly
			Variance	Biased (lower)	Biased (higher)	Maintained exactly
		Linear Relationshi	**X** versus **Y**	Biased	Maintained asymptotically	Maintained exactly
			(Y & S) versus **(X & S)**	Biased	Biased (attenuated)	Maintained exactly
			Monotonic (non-linear) relationships	Biased	Biased (attenuated)	Biased (attenuated)
			Non-monotonic relationships	Biased	Biased (attenuated)	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			Moderate	Easy	Moderate

- Theoretically, it is easy to see that the Tree based data perturbation performs very poorly even compared to noise addition methods.

Comparison of Data Shuffling and any Ad Hoc Method



				Method	
				Any Ad Hoc Method	Data Shuffling
Criteria			Disclosure Risk	Not minimized	Minimized
	Data Utility		Marginal Distribution	Not maintained	Maintained exactly
			Mean Vector	Maintained exactly	Maintained exactly
			Variance	Biased	Maintained exactly
		Linear Relationshi	X versus Y	Biased	Maintained asymptotically
			(Y & S) versus (X & S)	Biased	Maintained asymptotically
			Monotonic (non-linear) relationships	Biased	Maintained asymptotically
			Non-monotonic relationships	Biased	Not maintained
			Sub-set characteristics	Not maintained	Marginal characteristics maintained exactly; linear and monotonic relationships maintained asymptotically
			Ease of Use	Easy	Easy
			Ease of Implementation	Moderate	Difficult

- Given this information, which procedure do you think would be preferred
  - By the Subjects and Users?
  - By the Provider?

Summary

- The data provider must go through a rigorous process to evaluate the efficacy of a data masking technique
- The first step in this evaluation is a theoretical assessment of the different techniques
  - It may be necessary to perform additional evaluations (even empirical evaluations)
- The data provider must then release all information regarding the data masking technique that was used

Data Specific Issues

- One of the key aspects of implementing model based methods is to have a good understanding of the underlying characteristics of the data
- It may be necessary to make modifications based on data specific characteristics such as outliers
- One of the benefits of using model based methods is that it is possible to model these data specific characteristics (such as outliers and even non-monotonic relationships)

Aggregate Information

- The data provider must also release all possible aggregate information regarding the numerical confidential variables
  - True values of the mean, variance, correlation (product moment and rank order), covariance, and distributional characteristics
- The release of the microdata should be considered as additional information over and above aggregate information

Full Disclosure

- Releasing the aggregate information and then the details of the masking procedure provides full disclosure to all the stakeholders
  - The interested Subject can readily verify the disclosure risk claims
  - The interested User can readily assess the information loss claims
    - Releasing the aggregate information actually provides the user with the ability to compare actual versus masked results
    - Will let the User determine the “sensitivity” of the results

Assurance

- The greatest benefit from using the framework that was described here is that the data Provider can give
  - Explicit assurance to the data Subjects regarding disclosure risk
  - Explicit assurance to the data Users regarding information loss
    - The users now know the procedures for which the masked data will yield valid results and more importantly, procedures for which data will not yield valid results

An Example Assurance Statement for Data Shuffling

In order to protect the privacy and confidentiality of the data, the numerical data have been masked using Data Shuffling (Muralidhar and Sarathy 2006). In this procedure, the values of the confidential variables have been “shuffled” among the records. **Data Shuffling assures the lowest possible level of disclosure risk**. The original **data values remain unmodified** and hence all responses regarding an individual variable are maintained exactly. Data Shuffling also **preserves the linear and monotonic non-linear relationships** between all variables. Other types of relationships may not be preserved.

If you have any questions regarding the data, please contact our

An Example Assurance Statement for Sufficiency Based GADP

In order to protect the privacy and confidentiality of the data, the numerical data have been masked using a procedure called Sufficiency Based GADP (Muralidhar and Sarathy 2008). **The masking procedure assures the lowest possible level of disclosure risk**. 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, etc), **the estimates using the masked data will yield exactly the same estimates as the original data**. However, the marginal distribution of the individual variables and all non-linear relationships are not preserved.

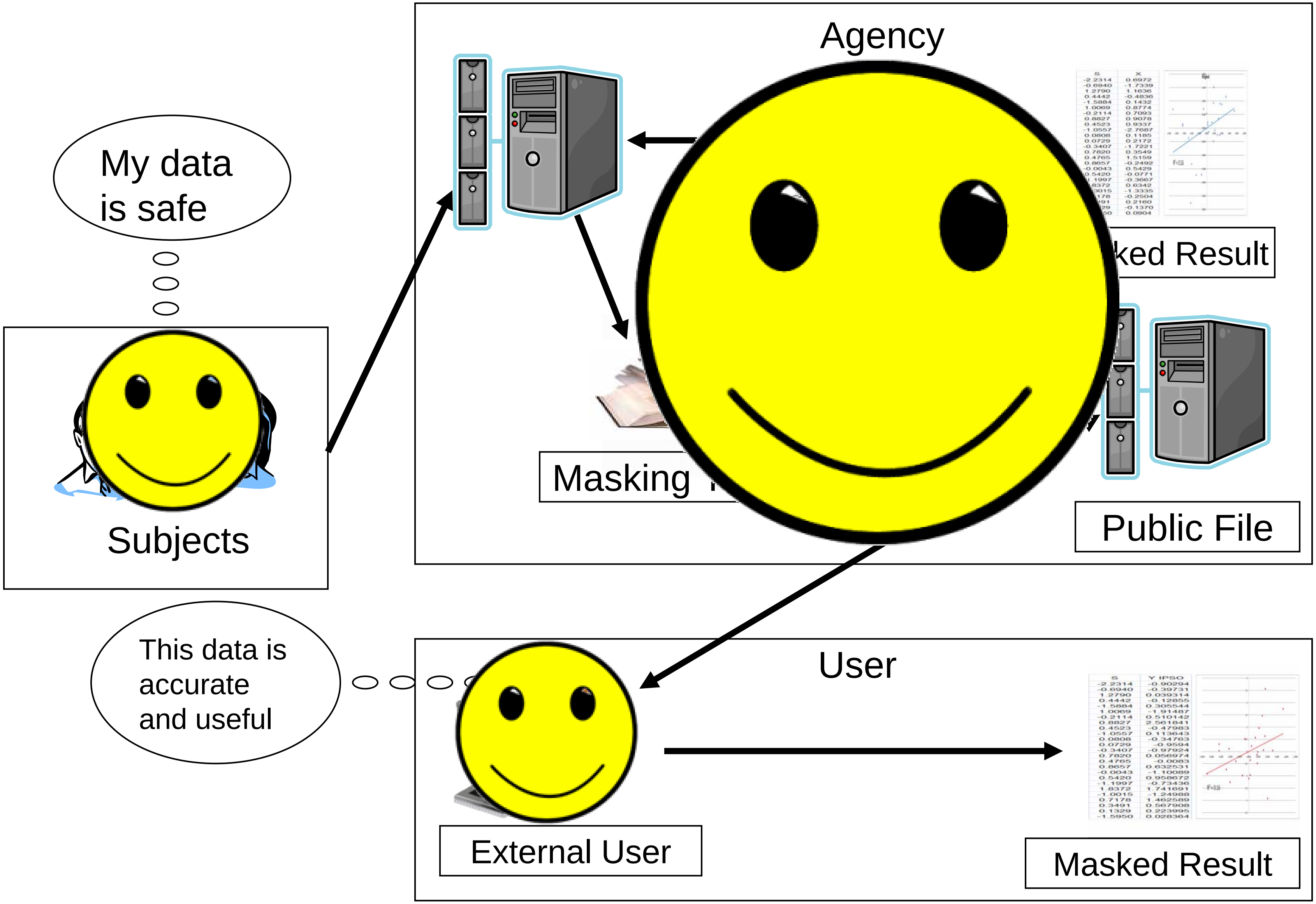
If you have any questions regarding the data, please contact our data administrator Jörg Dreschler

Conclusion – Agency Perspective

- We believe that the Agency that is entrusted with the task of gathering and disseminating data must provide maximum possible information to both the data subjects and data users regarding the efficacy of the data masking procedures. We believe that providing an assurance statement would alleviate the concerns of both the subjects and the users.

Conclusion – Research Perspective

- We also believe that when developing new masking techniques, it is the responsibility of the researcher to evaluate the theoretical performance characteristics of the masking technique being proposed. These theoretical performance characteristics must be clearly stated as a part of the paper.



My data is safe

Subjects

Agency

ked Result

Masking T

Public File

This data is accurate and useful

User

External User

Masked Result

Our Web Site

- You can find this and other papers at our web site:

http://gatton.uky.edu/faculty/muralidhar/maskingpapers

Questions, Comments, or Suggestions?

THANK YOU