Journal of Privacy and Confidentiality (2016-2017) **7**, Number 3, 99–129

A New Data Collection Technique for Preserving Privacy

Samuel S. Wu*, Shigang Chen†, Deborah Burr‡ and Long Zhang§

Abstract. A major obstacle that hinders medical and social research is the lack of reliable data due to people's reluctance to reveal private information to strangers. Fortunately, statistical inference always targets a well-defined population rather than a particular individual subject and, in many current applications, data can be collected using a web-based system or other mobile devices. These two characteristics enable us to develop a data collection method, called *triple matrix-masking (TM^2)*, which offers strong privacy protection with an immediate matrix transformation so that even the researchers cannot see the data, and then further uses matrix transformations to guarantee that the data will still be analyzable by standard statistical methods. The entities involved in the proposed process are a masking service provider who receives the initially masked data and then applies another mask, and the data collectors who partially decrypt the now doubly masked data and then apply a third mask before releasing the data to the public. A critical feature of the method is that the keys to generate the matrices are held separately. This ensures that nobody sees the actual data, but because of the specially designed transformations, statistical inference on parameters of interest can be conducted with the same results as if the original data were used. Hence the TM^2 method hides sensitive data with no efficiency loss for statistical inference of binary and normal data, which improves over Warner's randomized response technique. In addition, we add several features to the proposed procedure: an error checking mechanism is built into the data collection process in order to make sure that the masked data used for analysis are an appropriate transformation of the original data; and a partial masking technique is introduced to grant data users access to non-sensitive personal information while sensitive information remains hidden.

Keywords: Orthogonal transformation; Privacy-preserving data collection; General linear model; Contingency table analysis; Logistic regression.

1 Introduction

There is opportunity and need in medical and social research today to collect more and better data, while at the same time there is increasing pressure to safeguard the privacy of study subjects whose data are collected and analyzed. This sounds much like the "growing tension between confidentiality and data access" (Duncan and Pearson, 1991) in use of government databases. The medical community has recognized the need for systematic development of methods for data privacy (American Association of Medical

*Department of Biostatistics, University of Florida, Gainesville, Florida mailto:sw45@ufl.edu

†Computer & Information Science & Engineering, University of Florida, Gainesville, Florida mailto:sgchen@cise.ful.edu

‡Statistics, University of Florida, Gainesville, Florida mailto:burr@stat.ufl.edu

§Statistics, University of Florida, Gainesville, Florida mailto:????

© 2016-2017 by the authors

Colleges, 2010); however, statistical methods for data privacy have not focused on the needs of medical research as much as on those of social science research.

A common scenario where data confidentiality is a problem in social science research involves four parties: a *statistical agency*, *data users*, *data providers*, and *intruders*. The statistical agency plans and carries out the data collection, and once the data have been collected, plans the release of a possibly masked version of the data. The data users, who may be the same as the statistical agency, wish to do research at a population level using the data; such research is intended to provide benefit to society. The intruders wish to get around the built-in security and privacy barriers, to identify sensitive data about particular data providers, and to use this information in harmful ways. In this scenario, the goal of data masking or other methods to guarantee privacy of the data is to protect each individual data provider from having his data exposed to intruders, while allowing legitimate use of the data for beneficial research. Various statistical disclosure limitation methods have been proposed to achieve this goal, such as addition of noise (Kim, 1986; Kim and Winkler, 1995; Chawla et al., 2005), multiple imputation (Rubin, 1993), information preserving statistical obfuscation (Burridge, 2003), the post-randomization method (Gouweleeuw et al., 1998), controlled tabular adjustment (Cox et al., 2004), data shuffling (Muralidhar and Sarathy, 2006), random projection based perturbation (Liu et al., 2006), random orthogonal matrix masking (Ting et al., 2008). In addition, there are many approaches that were particularly developed for privacy protection of contingency table data, especially for the release of high-dimensional contingency tables. They include generalized shuttle algorithm (Dobra and Fienberg, 2009), synthetic data (Fienberg and Slavkovic, 2008; Winkler, 2008; Slavkovic and Lee, 2010), algebraic statistics (Dobra et al, 2008; Slavkovic and Fienberg, 2009), and differential privacy (Blum et al., 2005; Dwork, 2006; Barak et al., 2007; Fienberg et al., 2010; Yang et al., 2012), among others.

On the other hand, in a typical clinical study (such as a multi-center medical trial), the privacy scenario involves the funding agency (such as the National Institutes of Health), the study investigators (data collectors), the study participants (data providers) and potential intruders. In this scenario, the *data users* include the study investigators, as well as external researchers if the investigators make the data available to them. The usual approach to privacy is regulated by the Health Insurance Portability and Accountability Act of 1996 and subsequent rulings. Among other things, the law requires all researchers in both the clinical and data branches to undergo regular training on ethics and methods of guaranteeing data privacy and safety. The methods are to restrict access to all personal identifiers (such as name and social security number) from research databases, and to follow standard computer security practices. Data masking or transformation methods have not been used much if at all. One negative impact of the privacy regulations is that it often takes many months to get approval from the Institutional Review Board (IRB) before a clinical study can start, and even then the use of the data is subject to stringent restrictions. General linear regression, contingency table analysis, and logistic regression are commonly used in a typical multi-center medical trial. Furthermore the statistical analysis plan is often prespecified in the study protocol before recruitment and data collection. Once the data are analyzed and main

results are published by the research team, researchers on government-funded grants are required to release the data for academic and public use, and the only privacy protocol is that all personal identifiers are removed from the data.

Our overall aim in the present work is development of a system for privacy-preserving data collection and analysis which will be useful in both medical and social research. We propose a new method called *triple matrix-masking (TM^2)* that is performed *at the time of data collection*. There are three key ideas behind the approach we take in this paper. We use specially designed matrix transformations that preserve data features needed for standard statistical analyses, an idea developed by Ting et. al. (2008) for the purpose of microdata release for social science research. A new twist in our approach is the application of a transformation at the moment the data is collected, so that not even the study investigators know the actual values of sensitive variables. And in addition, we have incorporated ideas from computer science work on data security, including a protocol for handling of keys which involves an additional entity in the scenario, termed a *masking service provider.* Keller-McNulty (1991) made the valid point that statisticians working on data privacy need to incorporate ideas that have been developed by computer scientists working on private sector data security.

The TM^2 method works as follows. A masking service provider only receives masked data from data providers and then applies another mask. The data collectors who hold the key to the first mask partially decrypt the doubly masked data and apply a third mask before releasing the data to the public. The critical feature of the method is that the keys used to generate the masking matrices are held separately by the masking service provider and the data collectors. This ensures that nobody sees the actual data, but statistical inference on parameters of interest can be conducted with the same results as if the original data were used.

One motive for this work is to contribute to security of sensitive data, beyond the simple removal of personal identifiers from databases. In the medical area, this additional security may lead to a less cumbersome IRB approval process, and it may encourage more sharing of data when research is completed. In addition, there is a need to persuade potential study participants up front that any sensitive data that will be gathered will be secure from intruders. In studies about sensitive topics such as illegal activities, medical history and personal finance, research could be hindered by the potential subjects' concern about privacy. People often refuse to participate in research altogether. Or, they may consent to participate, but then purposely provide wrong information because they do not have enough trust in confidentiality protection or simply are reluctant to release private information.

The method we present here is an improvement of Warner's (1965) *randomized response* technique, which is well summarized in a monograph by Chaudhuri and Mukerjee (1988) and has been used in many applications (Ostapczuk et al., 2009; Quercia et al., 2011). This technique requests an interviewee to report whether or not his true binary answer to a sensitive question is the same as a randomly generated response, which only the interviewee sees. That is, the algorithm randomly flips an interviewee's true binary response with probability $(1 - c)$, where c is the chance of "yes" answer from

the random device. The investigator's ability to guess the response may be calibrated by adjusting the distribution of the randomly generated response, but the investigator cannot determine absolutely the interviewee's response. Therefore this technique meets the dual objectives of generating enough reliable data to yield fruitful inference and protecting respondents' privacy despite their truthful replies. However, Warner's randomized response technique can apply only to binary data and it is inefficient (see Section 4 for more details), while the TM^2 method loses no efficiency for statistical inference of binary and normal data because sufficient statistics are preserved.

The rest of the paper is organized as follows. In Section 2, we summarize the known facts that orthogonally record-transformed data preserve sufficient statistics for the general linear model and contingency table analysis; and under logistic regression the same inference results on parameters of interest can be obtained from certain attribute-transformed data as they would have obtained with the original data. In Section 3, we apply these results to matrix masking at the time of data collection. We show that, by distributing the keys of the random transformations, we can ensure that nobody sees the actual data, yet the masked data provides the same statistical inference results. We also add several features to the proposed procedure: an error checking mechanism is built into the data collection process in order to make sure that the masked data used for analysis are an appropriate transformation of the original data; and a partial masking technique is introduced to grant data users access to non-sensitive personal information while sensitive information remains hidden. In addition, we illustrate the new method through a subset of 20 observations from a recently completed clinical trial. In Section 4, we compare the TM^2 method with related work on privacy-preserving data collection, including Warner's randomized response technique, various cryptographic solutions, and anonymous communications. We summarize our contributions and further research in Section 4, while Appendix 1 provides a Matlab program for generating a random orthogonal matrix.

2 Properties of Matrix Masked Data

We use two types of matrix transformation in order to change data values yet preserve that information in the data which is essential for statistical analysis. In this section we summarize the properties of matrix masked data.

2.1 Orthogonally Record-Transformed Data Preserve Sufficient Statistics

First, we review the known fact that orthogonally record-transformed data preserve sufficient statistics for parameters of interest with the use of general linear model and contingency table analysis. Consequently, the exact same analytical results can be obtained with orthogonally-transformed data as with the original data. This fact has been used by Ting et al. (2008), who proposed a method they called random orthogonal matrix masking (ROMM) that preserves sufficient statistics under a linear model.

In ROMM and earlier work (Duncan & Pearson, 1991), the data collectors have the raw data matrix, which is multiplied by an orthogonal masking matrix before sending the resulting matrix to data analysts or others who request the data. This procedure assumes that the data collectors know the raw data before performing their masking operation. We propose a new method that improves privacy protection by preventing anyone other than data providers (participants themselves) from knowing the raw data; the procedure is performed distributively, allowing the data to be incrementally masked for each participant. Before presenting our procedure, we show that orthogonal transformation of data preserves sufficient statistics. For clarity, we decompose the data matrix $X_{n\times(p+1)}$ into two parts, $X = [Y, Z]$, where $Y_{n\times 1}$ is the vector for the outcome variable and $Z_{n\times p}$ denotes the model matrix. First, consider the general linear model,

$$Y = Z\beta + \epsilon,$$

where $\beta_{p\times 1}$ is the vector of unknown parameters, and $\epsilon_{n\times 1}$ is the vector of zero-mean random error terms (usually assumed to be normally distributed). The usual least-squares estimate $\hat{\beta}$ is the vector which minimizes the sum of squared errors $\|Y - Z\beta\|_2^2$; it is also the maximum likelihood estimate when ϵ is normal. Recall that when matrix Z is of full rank, the minimizer of the sum of squared errors is unique and the estimate $\hat{\beta}$ can be expressed as $\hat{\beta} = (Z'Z)^{-1}Z'Y$, where apostrophe ($'$) denotes transpose.

We consider applying an orthogonal transformation to the outcome vector $Y_{n\times 1}$, and the same transformation to the model matrix Z. An orthogonal transformation is a mapping from R^n to R^n which preserves lengths of vectors and angles between vectors. It may be represented by a square matrix $A_{n\times n}$ such that $A'A = I$, where I is the identity matrix. Now we fit the model based on AY and AZ rather than the original model based on Y and Z. That is, $AY = AZ\beta_{\text{new}} + A\epsilon$, where A is a *row* operator that transforms data *records* (each row represents one case). Denote the original least-squares estimate by $\hat{\beta}_{\text{orig}}$, and the new least-squares estimate on orthogonally-transformed data by $\hat{\beta}_{\text{new}}$. We have $\hat{\beta}_{\text{new}} = \big((AZ)'(AZ)\big)^{-1}(AZ)'(AY) = (Z'Z)^{-1}(Z'Y) = \hat{\beta}_{\text{orig}}$.

In other words, the least-squares estimates from the original and transformed data are the same when left-multiplying the data by an orthogonal matrix. This result can be confirmed by considering the usual geometric representation of the least-squares estimate. Stated in terms of the original estimate, the geometric interpretation is that $\hat{\beta}_{\text{orig}}$ provides a linear combination of the column vectors in Z such that the distance between the vector Y and the vector of predicted values $Z\hat{\beta}$ is the shortest, among all vectors in the subspace spanned by the column vectors of Z. Using the facts that orthogonal transformations preserve distances and angles between vectors, it is a short argument to show that $\hat{\beta}_{\text{new}} = \hat{\beta}_{\text{orig}}$. From this perspective, it is also a short argument to show that the regression parameter estimates are identical for the two models even if only a subset of variables from Z (and the corresponding subset from AZ) is used.

The residual vector for the original data is defined to be $e = Y - Z\hat{\beta}$. For the new data, the residual vector is $AY - AZ\hat{\beta} = A(Y - Z\hat{\beta}) = Ae$, which is the original residuals transformed by A. Since length is preserved by orthogonal transformation, the residual sum of squares will be the same for the two models. Furthermore, because

the covariance of $\hat{\beta}$ depends on only $Z'Z = (AZ)'(AZ)$ and the variance of ϵ, the estimate of the covariance matrix as well as the usual inference procedures will be identical. However, the individual residuals will be transformed so that residual plots and diagnostic methods will no longer be valid.

When an intercept term is included in a regression analysis, 1_n is a column of Z, where 1_n denotes the vector of n 1's. In this case, $A1_n$ is a column of AZ, therefore the first and second sample moments of Z can be derived from AZ. On the other hand, if we restrict A to be an orthogonal matrix that keeps 1_n invariant, i.e., $A1_n = 1_n$, then the sample means and sample covariance matrix for X and AX are the same (see Theorem 1 of Ting et al., 2008). In Remark 2, we describe a simple algorithm to generate such an orthogonal matrix.

Next we consider analysis of data in 2×2 tables. The raw data are two binary (0-1) vectors, Z_1 and Z_2, containing n observations. The data are commonly summarized as counts in a 2×2 table shown in Table 1, with rows labeled by the values of variable Z_1 and columns labeled by the values of variable Z_2. More specifically, the four cell values are: a is the number of observations that are 0's in both vectors Z_1 and Z_2, b the number of observations with 0 in Z_1 and 1 in Z_2, c with 1 in Z_1 and 0 in Z_2, and d with 1's in both Z_1 and Z_2. The contingency table can also be computed as follows: $Z_1'Z_1 = c + d$ is the number of 1's in vector Z_1, $Z_2'Z_2 = b + d$ is the number of 1's in vector Z_2, and $Z_1'Z_2 = d$ is the number of 1's that Z_1 and Z_2 have in common. From these three values and the sample size n, we can easily compute a, b, c and d.

Table 1. Correspondence between two forms of counts in 2×2 table

		Usual			Vector		
		Values of Z_2		Totals	Values of Z_2		Totals
		0	1		0	1	
Values	0	a	b	$a+b$	–	–	–
of Z_1	1	c	d	$c+d$	–	$Z_1'Z_2$	$Z_1'Z_1$
	Totals	$a+c$	$b+d$	n	–	$Z_2'Z_2$	n

If we want to hide values of Z_1 and Z_2, we can transform the data by multiplying them with an orthogonal matrix A before release. Note that even though the transformed data take *real* values, we can obtain the same contingency table from AZ_1 and AZ_2 as we would have gotten from the original data Z_1 and Z_2. Specifically, because $(AZ_1)'(AZ_1) = Z_1'Z_1$, $(AZ_2)'(AZ_2) = Z_2'Z_2$, and $(AZ_1)'(AZ_2) = Z_1'Z_2$, we have the same counts for the three quantities considered previously. However, with the transformed data, nobody knows the original value in Z_1 and Z_2 for any of the participants. Moreover, the usual analysis, including the chi-squared test and estimation of relative risk and odds ratio, will yield identical results for the transformed data as for the original data.

Remark 1. (Categorical variables with multiple levels and high-dimensional contingency tables) *Contingency tables, whose cells contain frequency counts from cross-classifying a sample or a population according to a collection of categorical variables (attributes), are among the most prevalent forms of statistical data. It is easy*

to check that, for variables with multiple levels and for high-dimensional contingency tables, the cell counts remain invariant if we include multiple dummy binary indicator variables. For an extensive literature on the contingency table analysis such as logit and log-linear models, see Bishop et al. (1975), Fienberg (1980) and Agresti (1990).

In certain applications, it is not enough to hide the values of the variables. For example, a particular contingency table cell may be too sensitive to be released if the number of respondents is smaller than a threshold. In such a case, we should protect privacy by combined use of the TM^2 method and other disclosure limitation techniques, including cell suppression, rounding, sampling, data swapping, and other sampling and simulation techniques (for more details see Duncan et al., 2001; Oganian and Domingo-Ferrer, 2003; Domingo-Ferrer and Saygin, 2008; Fienberg and Slavkovic, 2008; and Slavkovic, 2010). The TM^2 method makes sure that the data collectors do not see the raw patient data (Z_1 and Z_2) but they can still derive the correct contingency table (a, b, c and d). If the data collectors find that some cells in the contingency table are sensitive according to a threshold rule, they can use the disclosure limitation techniques to protect these cells from being disclosed to others.

2.2 Attribute-Transformed Data Enable Logistic Regression

In many applications, we study the association between a binary outcome and a continuous variable, or it is necessary to adjust for some covariates in the investigation of relationship between a binary outcome and a categorical variable. In such cases, we employ a logistic regression model, in which $\text{logit}[\pi(Z)] = Z\beta$, where $\pi(Z) = Pr(Y = 1|Z)$ for binary response Y. One usually estimates the parameter β by the method of maximum likelihood and estimates the covariance matrix by $\widehat{\text{Cov}}(\hat{\beta}) = (Z'\hat{D}Z)^{-1}$, where $\hat{D}$ is a diagonal matrix with $\hat{\pi}_i(1 - \hat{\pi}_i)$ on the main diagonal and $\hat{\pi}_i$ is the maximum likelihood estimate of the response probability for the ith subject (Agresti, 1990; p. 114).

We consider a data transformation XB where B is a $(p + 1) \times (p + 1)$ matrix constructed so that some of the analyses for logistic regression can be carried out on the transformed data with the same results as for the original data. Specifically, we choose the column operator B to be a block diagonal invertible matrix that keeps the response variable invariant, i.e., $B = diag(I_1, C)$. Now we fit the logistic regression model based on $W = ZC$ rather than the original model based on Z for the same response, i.e., $\text{logit}[\pi(W)] = W\beta_{\text{new}} = ZC\beta_{\text{new}}$. It is easy to see that: (i) the maximum likelihood estimates satisfy $\hat{\beta}_{\text{new}} = C^{-1}\hat{\beta}$; (ii) $\hat{D}$ is the same under two models; and (iii) $\widehat{\text{Cov}}(\hat{\beta}_{\text{new}}) = (W'\hat{D}W)^{-1} = C^{-1}(Z'\hat{D}Z)^{-1}C'^{-1} = C^{-1}\widehat{\text{Cov}}(\hat{\beta})C'^{-1}$. Therefore, the maximum likelihood estimate of the treatment effects and their estimated standard errors are the same for the original data and the matrix-masked data if we choose C from block diagonal matrices with an identity matrix on the top left corresponding to variables of treatment effects. That is, the column operator B keeps the response and treatment group variables invariant and applies the column transformation only to other

covariates. However, it should be acknowledged that the results may be different for other estimators of variance in the logistic regression and the effects of other covariates cannot be estimated based on the above masking procedure.

Because the binary response and treatment group variables are kept invariant, we can calculate the exact residuals and log likelihood for the fitted and null models. Consequently, we can perform most goodness of fit assessments, including the Pearson or likelihood-ratio chi-squared statistics (Agresti, 1990; p. 107 – 112). For example, for the fitted model the maximized log likelihood is $\sum_{i=1}^{n}[Y_i \log(\hat{\pi}_i) + (1 - Y_i) \log(1 - \hat{\pi}_i)]$; and for the null model it is $n[\bar{Y} \log(\bar{Y}) + (1 - \bar{Y}) \log(1 - \bar{Y})]$, where $\bar{Y} = \sum Y_i/n$. In addition, we can evaluate the association between the observed binary responses $\{Y_i\}$ and their fitted values $\{\hat{\pi}_i\}$, as well as the proportional reduction in error obtained by using $\hat{\pi}_i$ instead of $\bar{Y}$ as a predictor of Y_i. However, much work remains to be done in this area, including diagnostic analysis on the relationship between the response and the covariate variables and the appropriate choice of link function.

3 TM^2 Hides Original Data from Everyone

As Duncan & Pearson (1991) and Du et al. (2004) pointed out, matrix masks are powerful and they encompass many commonly proposed disclosure-limitation methods. In this section, we propose two implementations of the TM^2 method, which perform data masking at the time of data collection so that the original data are hidden from everyone, while statistical analysis can still be performed with the same results from the masked data as if they were from the original data. These new methods will be attractive to both investigators and participants in studies that involve sensitive personal information.

3.1 The First TM^2 Method

Consider stroke rehabilitation research as an application example. Dobkin & Dorsch (2011) describe technology for continuously monitoring patient mobility and community activity, which are essential to optimization of therapies and development of new treatments for patients with neurological problems. These data can be used to construct an accurate measure of daily living, an objective version of the usual "Activities of Daily Living" variable, described in Duncan et al. (1999) and elsewhere. One such system consists of an ankle accelerometer and smartphone, with the smartphone programmed to continuously compute and transmit positions and activity variables to a clinic, using a geographical positioning system (GPS). The collected data give detailed information about time and type of places the patient visits (e.g., shopping, active recreation such as sports and travel, spiritual or religious activities, and hospital visit), total distance and geographic area traveled, movement patterns, etc. Such information can be sensitive to some patients. In order to include privacy-sensitive patients, it is worthwhile to develop a smartphone program that directly converts GPS coordinates to activity variables and then masks the resulting mobility and activity data before sending them out.

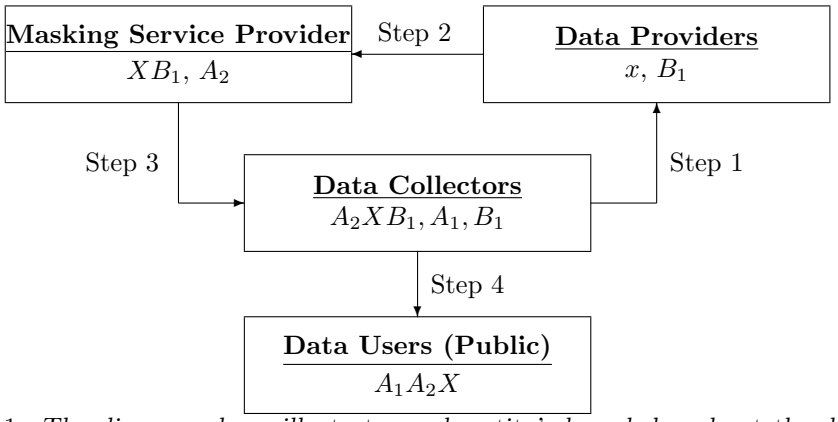

Figure 1: *The diagram above illustrates each entity's knowledge about the data and the masking matrices in the first TM*^{2} *method. The masking service provider knows* XB_1*, the data collectors know* A_2X*, and* A_1A_2X *is available to everybody including the public. Nobody other than data providers (participants) knows the original data* X*.*

We propose a triple matrix-masking method to address the above requirement. In addition to data providers, data collectors and data users, the method requires a masking service provider (see Figure 1). In the previous example, data providers are patient participants, and data users are study investigators as well as other researchers who can access the information. Typically, the data managers and statistical analysts in the study investigative team are in charge of data collection. Also, they release transformed data to the data users once the data have been collected. The masking service provider may be a private business or a government entity established to promote data sharing. It is the first entity that receives the data in a masked form; and it applies another mask before sending the doubly masked data to the data collectors. Because the data collectors hold the key to the first mask, they can partially decrypt the doubly masked data and apply a third mask before releasing them to the public.

Specifically, let x be a $1 \times (p+1)$ vector containing a single participant's sensitive information and X be an $n \times (p+1)$ data matrix from a cohort of participants. The TM^2 method consists of the following steps:

Step 1. The data collectors plan the data collection, create the database structure, program the data collection system. They choose a key to generate a $(p+1)\times(p+1)$ random invertible matrix B_1, which is distributed to the participants' data collection devices.

Step 2. At the time of data collection, a participant's data x are immediately transformed by B_1 before leaving the participant's device; only masked data xB_1 are sent to the masking service provider.

Step 3. The masking service provider chooses a different key to generate an $n \times n$

random orthogonal matrix A_2, using the algorithm given in Appendix 1. After receiving data from all participants, it aggregates the individual data into XB_1, applies record transformation and sends the doubly masked data A_2XB_1 to the data collectors.

Step 4. The data collectors multiply A_2XB_1 by B_1^{-1} to get back A_2X, choose another key to produce an $n \times n$ random orthogonal matrix A_1 and publish A_1A_2X, which is accessible by data users.

Remark 2. (Choice of Orthogonal Operator) *Both orthogonal operators A_1 and A_2 can be obtained by the Gram-Schmidt orthonormalization of a random normal matrix, which is controlled by some random number generator seed (i.e., key). The resulting matrix is a draw from the uniform distribution on orthogonal matrices under the Haar measure (see Eaton, 1983; p. 234). Let Z_1 and Z_2 be two $n \times (n-1)$ random normal matrices, and M_1 and M_2 be Gram-Schmidt orthonormalization of $[1_n, Z_1]$ and $[1_n, Z_2]$, which have the first column vector parallel to 1_n. Note that orthogonal matrix $A = M_1M_2'$ transforms column vectors in M_2 to those in M_1, hence A keeps 1_n invariant. Appendix 1 presents a Matlab program for generating such an orthogonal operator. More information about random orthogonal matrices can be found in Steward (1980), Anderson et al. (1987), and Diaconis (2005).*

Remark 3. (Improvement of Initial Masking at Step 2) *When the data matrix X has few columns, the masking service provider (or any data intruder who has access to XB_1) may be able to recover B_1 and hence the full data if he or she knows a sufficient number of original records. To improve the level of privacy protection offered by the column operator B_1, a participant's data x can be augmented with extra columns of random noise. These additional columns will not affect the statistical analysis of A_1A_2X.*

The above method protects the privacy of individual participants because nobody other than data providers knows the original data X. As illustrated in Figure 1, the masking service provider only knows XB_1 and A_1A_2X, but has no access to B_1 and A_1; the data collectors only know A_2X and A_1A_2X, but have no access to A_2; while the public knows A_1A_2X but does not know A_1 and A_2. The privacy protection depends on the distribution of keys: the data collectors have keys to generate matrices A_1 and B_1, while the masking service provider holds the key to generate matrix A_2.

The security of the TM^2 method is briefly given as follows. Let S be the set consisting of all data matrices that are orthogonal transformations of X, which are equivalent to orthogonal transformations of A_1A_2X. Because any member in S may result in the masked data (namely, A_1A_2X), for data users who have access to A_1A_2X and only know that A_1 and A_2 are random orthogonal matrices, they only know that X belongs to the set S. That is, for any $W = \Gamma X$ from S where Γ is an orthogonal matrix, there exist two orthogonal matrices $\tilde{A_1}$ and $\tilde{A_2}$ (for example, $\tilde{A_1} = A_1$ and $\tilde{A_2} = A_2\Gamma'$) such that data users receive $\tilde{A_1}\tilde{A_2}W = A_1A_2X$. Similarly, the data collectors who have access to A_2X and A_1A_2X only know that the original data matrix is an element in S. Lastly, the masking service provider has access to XB_1 in addition to A_1A_2X, thus it knows that each column vector of X belongs to the subspace spanned by the column

vectors of XB_1 and that X is an element in S. Therefore it does not have enough information to disclose values of data in X because B_1 is a general invertible matrix.

On the other hand, because row operators A_1 and A_2 are orthogonal matrices, A_1A_2X preserves sufficient statistics for the general linear model and for contingency table analysis. In other words, A_1A_2X can be analyzed to obtain the same results as if X was used under either the general linear model or contingency table analysis. The main reason for right-multiplying the column operator B_1 in the first step is that this operation can be done one row of X at a time. That is, the masking operation can be done independently at each participant's device, allowing the collection of masked data one record at a time.

Furthermore, the TM^2 method can be designed to enable partial masking, allowing data users to access part of the data (such as treatment group), while keeping other sensitive information hidden. Specifically, let X_1 be an $n \times p_1$ matrix for insensitive data, and X_2 be an $n \times p_2$ matrix for sensitive information. The data collectors are required to choose B_1 from the set of block diagonal matrices with a $p_1 \times p_1$ identity matrix at the top left corner and a $p_2 \times p_2$ invertible matrix B_1^* at the bottom right corner, i.e., $B_1 = diag(I_{p_1}, B_1^*)$. Hence the masking service provider will receive $XB_1 = [X_1, X_2B_1^*]$, where the sensitive information is masked through attribute-transformation with B_1^*. In addition, the masking service provider and the data collectors are required to generate orthogonal matrices A_1 and A_2 that keep X_1 invariant, which guarantees that data users have access to X_1 because $A_1A_2X = [X_1, A_1A_2X_2]$. Here, it is important to choose A_1 and A_2 that keep X_1 invariant, which guarantees that statistical associations between variables in X_1 and X_2 are the same as those between X_1 and $A_1A_2X_2$. Also, in this case, the data users gain more information than $X'X$ because of their access to X_1.

In addition, a quality assurance technique can be easily implemented in the proposed privacy-preserving data collection method to aid the data collectors in checking whether appropriate transformations were applied to the original data X in Steps 2 and 3. To do so, we require the matrix X to add a column of 1s (i.e., 1_n) as the first column, as well as a column of constants (say, c) as the last column. Then after the data collectors reverse the B_1 transformation to get A_2X, the last column of A_2X should be c times the first column of A_2X. Also, in the case that A_2 is an orthogonal matrix that keeps 1_n invariant, the last column of A_2X should equal to $c\,1_n$.

3.2 An Illustrative Example of the 1st TM^2 Method

In a medical or social study, individuals are often unwilling to share sensitive information such as illegal activities, medical conditions or personal finance. If the investigators can convince the individuals that their data will be used only in an aggregate study and cannot be linked back to them, it could increase their willingness to participate. In this subsection, we demonstrate the first TM^2 method using a random subset of 20 observations from the LEAPS study described in Duncan et. al (2011). Table 2 presents the original data of eight variables as explained below.

Table 2: Random subset of 20 observations from LEAPS, X

Obs No	Response	Group	Δ	Age	BBS	IH	MIF	ADL/iADL	QA
1	0	1	0.08	63	30	1	1	50	888
2	1	0	0.67	57	40	0	1	62.5	888
3	1	0	0.20	47	43	0	1	87.5	888
4	1	1	0.52	38	39	1	1	80	888
5	1	1	0.47	83	36	0	0	60	888
6	1	0	0.34	54	29	0	0	80	888
7	0	1	-0.07	50	13	0	1	47.5	888
8	1	1	0.34	68	48	0	0	72.5	888
9	1	0	0.25	57	47	0	0	72.5	888
10	1	0	0.48	65	39	0	1	47.5	888
11	1	1	0.15	43	9	1	0	50	888
12	1	0	0.12	81	40	0	0	67.5	888
13	0	1	-0.13	76	48	1	1	32.5	888
14	1	1	0.15	84	29	0	0	42.5	888
15	1	0	0.29	75	39	0	0	85	888
16	1	1	0.20	65	33	0	0	42.5	888
17	1	1	0.67	65	45	0	1	75	888
18	1	1	0.15	66	24	0	1	55	888
19	1	0	0.33	51	40	0	0	50	888
20	1	1	0.22	90	44	0	0	100	888

Variable	Description
Response	Improved functional level of walking 1 year after the stroke (Yes=1/No=0)
Δ	Change in walking speed from 2-month to 12-month post-stroke (m/s)
Group	Treatment group, 1 = Locomotor Training Program; 0 = Home Exercise Program
Age	Age at stroke onset (years)
BBS	Berg Balance Scale in sitting, standing, reaching, shifting weight, and turning
IH	Inpatient Hospitalization post randomization (Yes=1/No=0)
MIF	Multiple or Injurious Falls post randomization (Yes=1/No=0)
ADL/iADL	Activities of daily living (ADL's) and instrumental activities of daily life (iADL's)

The data include two sensitive medical conditions: inpatient hospitalization (IH) and multiple or injurious falls (MIF). Recall that our goal is to enable the secure release of data to anyone (see Figure 1), so that not only the data collectors but also other researchers can use the data. However, some patients may not want information of their hospitalization or injuries to be made public, which could adversely affect their opportunities of employment or insurance policies. The proposed TM^2 methods address this problem by collecting and publishing only the masked data through the following four steps:

Step 1. The data collectors plan the data collection and create a database consisting of the eight variables listed above and a variable for quality assurance. Also, a web-based data entry system is developed for each participant to enter the data. In addition, the data collectors choose a key of 535 as the random seed to generate a 9×9 random invertible matrix

$$B_1 = \begin{pmatrix} 0.3622 & 0.5330 & 0.5465 & 0.6382 & 0.5198 & 0.1257 & 0.9477 & 0.9711 & 0.0889 \\ 0.7470 & 0.5532 & 0.1052 & 0.5047 & 0.7759 & 0.6993 & 0.4742 & 0.8163 & 0.3183 \\ 0.1635 & 0.1752 & 0.9745 & 0.7202 & 0.6283 & 0.8917 & 0.3486 & 0.8989 & 0.8635 \\ 0.6691 & 0.1261 & 0.6600 & 0.5385 & 0.1014 & 0.6139 & 0.8303 & 0.6335 & 0.9892 \\ 0.6674 & 0.1946 & 0.7629 & 0.4894 & 0.1891 & 0.0904 & 0.0578 & 0.8739 & 0.6303 \\ 0.4392 & 0.4399 & 0.6468 & 0.6252 & 0.3250 & 0.1620 & 0.6275 & 0.3957 & 0.3935 \\ 0.3429 & 0.4247 & 0.5300 & 0.2512 & 0.0221 & 0.1629 & 0.8318 & 0.0557 & 0.1729 \\ 0.4811 & 0.6877 & 0.6486 & 0.1597 & 0.6365 & 0.3162 & 0.8877 & 0.3551 & 0.0631 \\ 0.8146 & 0.9458 & 0.9722 & 0.4226 & 0.9869 & 0.6940 & 0.8043 & 0.5670 & 0.8160 \end{pmatrix}, \tag{1}$$

which is incorporated to the data entry system.

Step 2. At the time of data collection, the first participant enters its data which are shown in the first row of Table 2. The record is immediately transformed by B_1 and only masked data, which are shown in the first row of Table 3, are sent to the masking service provider. This is repeated for all 20 subjects.

Step 3. The masking service provider chooses a different key 536, and uses the Matlab program described in the Appendix 1 to generate a 20×20 random orthogonal

Table 3: Attribute-transformed data, XB_1

Obs No	Response	Group	Δ	Age	BBS	IH	MIF	ADL/iADL	QA
1	811.12	889.45	961.55	433.27	921.40	674.52	814.59	588.70	809.91
2	819.06	898.88	973.69	436.86	930.43	675.49	821.34	598.36	810.96
3	826.32	915.31	985.14	436.60	945.60	677.11	835.24	603.11	804.13
4	815.26	909.29	972.35	429.96	940.46	670.00	822.09	592.74	793.22
5	832.96	899.75	985.56	448.62	931.35	690.65	840.05	611.03	833.97
6	817.73	907.91	973.82	432.17	938.96	677.73	832.80	592.71	801.70
7	789.40	882.32	937.58	416.82	914.86	663.92	799.92	564.19	785.66
8	836.92	908.77	992.80	448.32	939.97	686.37	839.34	616.34	827.37
9	828.13	906.62	984.58	441.33	937.83	678.75	829.64	607.60	815.46
10	816.50	889.34	968.30	438.15	921.39	675.40	814.54	597.06	817.13
11	783.75	882.96	932.41	412.66	915.95	660.37	796.91	558.65	776.87
12	837.09	904.82	991.71	449.94	935.68	691.15	844.68	614.79	834.36
13	823.37	882.52	972.31	446.14	914.86	678.41	810.82	606.26	832.83
14	820.48	886.42	969.22	442.71	918.79	684.82	824.83	599.05	829.17
15	840.86	915.93	998.50	449.14	946.13	693.06	855.23	616.48	829.05
16	810.45	884.82	959.78	434.47	917.65	673.56	809.30	590.55	812.94
17	834.51	910.01	991.00	446.12	940.92	685.50	839.84	613.06	823.13
18	811.46	892.20	962.16	432.81	923.99	677.43	821.52	587.77	809.17
19	808.63	889.04	960.77	431.14	921.63	667.39	804.31	589.76	803.77
20	862.18	929.65	1021.98	462.51	958.87	708.10	881.74	636.44	848.24

matrix A_2 = GenerateROM(536, 20). Due to space limit, we omit the A_2 matrix here but readers can easily get the matrix by ranning the Matlab program. After receiving attribute-transformed data from all participants (XB_1 shown in Table 3), the masking service provider applies record transformation and sends the doubly masked data (A_2XB_1 shown in Table 4) to the data collectors.

Step 4. The data collectors choose another key 537 to produce a 20×20 random orthogonal matrix A_1= GenerateROM(537, 20), which is once again omitted. They multiply A_2XB_1 by B_1^{-1} to get back A_2X, left-multiply A_2X by A_1, and then publish masked data A_1A_2X (see Table 5) so that data users have access to orthogonally-transformed data.

Table 4: Doubly masked data transmitted to the data collectors, A_2XB_1

Obs No	Response	Group	Δ	Age	BBS	IH	MIF	ADL/iADL	QA
1	795.60	887.49	945.96	420.82	919.73	664.97	804.02	571.99	788.72
2	840.43	919.19	1000.48	447.63	949.45	687.48	849.44	618.21	822.65
3	791.58	881.74	939.84	419.67	914.62	664.87	800.04	567.99	789.47
4	796.13	879.39	943.55	424.02	912.19	667.12	800.10	573.20	798.15
5	841.80	923.57	1001.96	447.49	953.84	686.33	849.39	619.67	818.79
6	805.99	893.06	958.68	427.91	925.37	669.24	811.34	584.31	797.94
7	824.78	904.14	979.76	439.20	935.27	680.03	830.65	602.15	814.55
8	832.45	906.93	989.09	444.66	937.86	680.75	831.87	612.44	821.24
9	817.76	901.68	972.02	434.31	932.73	676.24	825.40	594.34	806.79
10	853.09	917.85	1009.35	459.15	947.97	705.96	871.74	627.80	847.91
11	812.33	883.40	962.23	436.56	916.05	673.45	807.41	593.59	816.73
12	834.91	896.09	986.64	451.41	927.64	691.55	838.55	613.97	840.33
13	834.47	900.47	988.38	449.43	932.12	685.13	831.75	615.64	832.04
14	832.90	908.16	987.47	445.17	938.90	690.72	847.27	608.08	825.70
15	806.27	903.66	960.74	423.93	935.32	672.25	824.64	580.30	789.34
16	821.46	888.51	970.92	442.41	920.68	681.94	821.66	600.61	826.57
17	799.74	877.59	948.42	427.91	911.01	662.34	789.62	582.39	801.58
18	834.54	907.82	991.00	446.67	938.49	683.35	835.70	614.02	824.29
19	818.12	910.93	975.28	431.66	941.56	676.32	832.92	593.05	798.56
20	831.83	894.32	983.43	449.71	925.90	689.65	835.22	610.91	837.67

Table 5: Matrix-masked data released to data users, A_1A_2X

Obs No	Response	Group	Δ	Age	BBS	IH	MIF	ADL/iADL	QA
1	0.82	0.75	0.09	60.91	33.20	0.42	1.06	67.76	888
2	0.31	1.43	0.07	84.76	39.35	0.72	0.60	50.69	888
3	1.34	0.27	0.29	108.09	43.04	-0.76	-0.76	81.61	888
4	0.50	0.93	0.30	68.94	17.70	0.35	1.40	28.43	888
5	0.49	1.23	0.14	80.43	49.46	0.26	0.82	89.87	888
6	0.31	0.81	0.07	47.74	23.39	0.85	0.58	50.08	888
7	0.61	-0.15	0.27	68.08	52.09	-0.06	1.09	51.89	888
8	1.36	0.39	0.72	47.80	47.55	0.54	0.57	98.78	888
9	1.49	-0.06	0.46	48.19	31.22	-0.29	0.51	90.43	888
10	0.28	-0.20	0.05	52.30	33.94	0.55	0.60	64.95	888
11	0.82	0.91	-0.24	63.64	15.00	0.35	-0.31	50.64	888
12	1.14	0.03	0.41	68.22	54.36	-0.06	0.38	45.89	888
13	1.05	0.85	0.14	60.35	42.17	0.21	0.38	53.95	888
14	1.30	0.64	0.56	65.24	31.98	0.49	0.41	70.15	888
15	0.73	0.99	0.45	58.41	36.90	-0.20	0.83	75.78	888
16	1.12	0.74	0.47	68.55	36.60	0.19	-0.32	57.37	888
17	0.94	-0.22	0.29	59.64	35.54	-0.15	0.15	76.90	888
18	0.80	0.54	0.25	54.06	35.94	0.16	0.10	59.08	888
19	0.78	1.13	0.28	61.54	34.54	0.72	0.29	46.80	888
20	0.81	0.97	0.35	51.12	21.03	-0.28	0.64	48.95	888

Table 6. Correspondence between two forms of counts in 2×2 table

		Vector			Usual		
		Multiple or Injurious Falls			Multiple or Injurious Falls		
		Yes	No	Totals	Yes	No	Totals
Group	LTP	$V_1'V_2 = 6$	–	$V_1'V_1 = 12$	6	6	12
	HEP	–	–	–	3	5	8
	Totals	$V_2'V_2 = 9$	–	$n = 20$	9	11	20

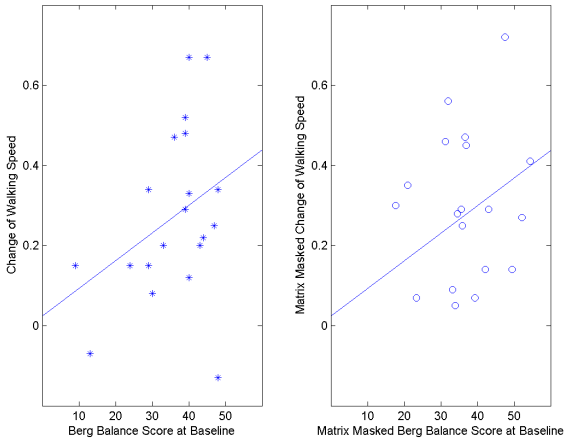

Figure 2: Scatter plots and fitted least-squares lines for the original and matrix masked data. The left panel is the actual data and its model fit; the right panel is the masked data and its model fit. The points in the matrix masked data have been completely scrambled and bear no relationship with the original data points; yet the regression line is exactly the same.

Figure 2 shows that regression lines is exactly the same for the actual data X and masked data A_1A_2X. Also, the residuals from both regressions would have the same distribution if they are normally distributed.

Table 5 shows that the transformed data for the binary variables (Response, Group, IH and MIF) take real values. From these masked data, users can only guess each participant's sensitive medical conditions - whether she or he had IH and MIF post randomization. However, for statistical inference, users have access to the exact counts for contingency tables. For example, the frequency counts can be obtained from masked data of Group (V_1) and MIF (V_2) as shown in Table 6.

3.3 The 2nd TM^2 Method

In many applications, we would like to conduct logistic regression. As stated in Section 2, it is sufficient to have access to data XB, where B is a block diagonal invertible matrix that keeps the response and treatment variables invariant. The first TM^2 procedure can be modified so that the data users know XB but nobody except for participants knows the original data X. In this case, we reverse the usage of the two random matrices, i.e., the data collectors generate the row operator A_0 and the masking service provider applies the column operator B_1. Both operators are invertible matrices, but not required to be orthogonal. The new procedure is as follows:

Step 1. The data collectors plan the data collection, create the database structure, program the data collection system. They choose a key to generate an $r \times r$ random invertible matrix A_0, which is distributed to the participants' data collection devices.

Step 2. At the time of data collection, a participant's data x are independently augmented to x^* with $(r-1)$ extra rows of random noise (which the data collectors do not know), and only the transformed data A_0x^* is sent by the participant to the masking service provider. The extra rows are necessary so that the left-multiplication of A_0 can be performed.

Step 3. The masking service provider chooses a different key to generate a $(p+1) \times (p+1)$ random invertible matrix B_1 that is block diagonal and keeps invariant the variables representing the response and treatment groups, applies attribute-transformation and sends the doubly masked data $A_0x^*B_1$ to the data collectors.

Step 4. The data collectors left-multiply $A_0x^*B_1$ by A_0^{-1} to get back x^*B_1, extract the first row of x^*B_1 to get xB_1, and aggregate data xB_1 from all participants to get XB_1. Then, they choose another key to produce a $(p+1) \times (p+1)$ block diagonal random invertible matrix B_2 that has the same invariant property as B_1, right-multiply XB_1 by B_2, and publish XB_1B_2, which is made publicly accessible to data users.

Remark 4. (Quality Assurance of the 2nd TM^2 method) *Similar to the first TM^2 method, we can add a device for the data collectors to check whether appropriate transformations were applied to the augmented data x^*. The trick is to add a row of constants (say, c) as the last row among the extra rows of noise appended to the original data x and use column operator B_1 that satisfies $1_n'B_1 = 1_n'$. After the data collectors remove the A_0 transformation to obtain x^*B_1, the last row of x^*B_1 should equal to $c\,1_n'$.*

Because logistic regression is a widely used method in biomedical and social research, many people have investigated approaches to conduct privacy preserved logistic regression with multiple data sources. For example, Fienberg et al. (2006) described "secure" logistic regression when all variables are categorical. And Fienberg et al. (2009) proposed an approach to carry out "valid" logistic regression with quantitative covariates using secure multi-party computation (SMC). Their approach proceeds in two steps:

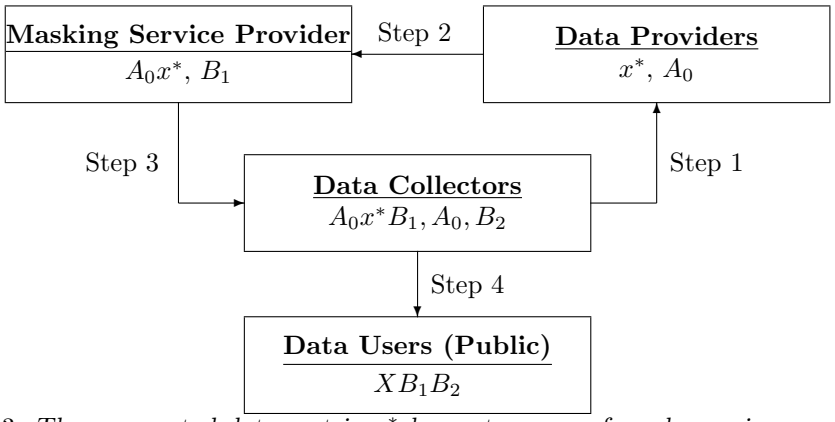

Figure 3: *The augmented data matrix x^* has extra rows of random noise appended to record x. The masking service provider knows A_0x^*, the data collectors know x^*B_1, and XB_1B_2 is available to everybody including the public.*

1) An initial estimate of regression coefficients is chosen; 2) for every iteration of the Newton-Raphson algorithm, a new estimate of regression coefficients is found using the following secure summation process: the first party shares its intermediate statistics with the addition of a random matrix; each remaining parties add its intermediate statistics to the updated sum; and at the last step the first party removes random noise and shares the global sum as well as the updated estimate.

TM^2 and SMC are designed for different purposes. The former ensures that certain statistical investigations can be carried out without requiring data providers to reveal their private data to data collectors. The latter ensures that multiple data collectors can perform joint statistical investigations without revealing their data to each other. For example, three hospitals collect private data from their patients respectively and then perform joint data mining without exchanging their raw data. In this example, each hospital still holds its patients' private data, which is against the design goal of TM^2.

If we perform SMC directly among the patients' devices, the two methods would remain different. The TM^2 method is distributive in data collection but centralized in data storage and data analysis. By contrast, the SMC approach requires distributed storage of data as well as distributed computation, which is practically infeasible when data storage and computation are performed directly by patient devices. Specifically, if we require that the private data of patients never leave their devices, the SMC method will place significant computation overhead on patient devices, particularly when a study involves thousands or more patients. More importantly, all patients have to stand by ready for any statistical analysis that may happen years into the future, which makes the SMC approach not feasible for medical studies that collect patient data over a long time - when patients leave a study they take their data away if we require that private data can never leave patient devices. There is no such issue with the TM^2 method since

it keeps the patients' data in a masked form, and the data is available for analysis at any time into the future after the patients have left the study.

TM^2 and SMC methods may appear to be complementary to each other. With multiple data collectors, TM^2 can be used to collect data from patients in a masked form to their respective data collectors, which may then use SMC to perform joint mining. However, we point out that since the masked data collected by TM^2 can be made publicly available, it becomes unnecessary to use SMC for joint mining over already masked data.

Finally, we can modify the second TM^2 method to allow data users to perform different types of statistical analysis. Suppose the masking service provider chooses an $n \times n$ random *orthogonal* matrix A_1 in addition to the block diagonal random invertible matrix B_1, while the data collectors hold keys to generate an $n \times n$ random *orthogonal* matrix A_2 in addition to the random invertible matrix A_0 and the block diagonal random invertible matrix B_2. Once the data collectors recover XB_1, they left-multiply A_2 and send A_2XB_1 back to the masking service provider, who removes B_1 and returns A_1A_2X. Then, the data collectors release A_1A_2X and XB_1B_2 to data users, who can conduct general linear regression, contingency table analysis or logistic regression. The first TM^2 method can be modified similarly to let the data users access both attribute-transformed data and orthogonally record-transformed data. Specifically, the masking service provider generates a block diagonal random invertible matrix B_2 in addition to the $n \times n$ random orthogonal matrix A_2 and sends A_2XB_1 and XB_1B_2 to the data collectors, who then publish A_1A_2X and XB_1B_2. It should be pointed out that, while release of two data products enables different types of statistical analysis, it could increase the disclosure risk since the data intruders may combine the different products to disclose confidential information. Further research is needed to assess disclosure risk in such scenarios.

3.4 An Illustrative Example of the 2nd TM^2 Method

Next, we illustrate the second TM^2 Method using the 1st and 11th observations of the LEAPS data. The procedure consists of the following four steps:

Step 1. The data collectors plan data collection similar to the first step of the first TM^2 Method, except that there is no variable for quality assurance. The data collectors choose key 535 as a random seed to generate an 8×8 random invertible

matrix

$$A_0 = \begin{pmatrix} 0.3622 & 0.8146 & 0.6877 & 0.5300 & 0.6252 & 0.1891 & 0.6139 & 0.3486 \\ 0.7470 & 0.5330 & 0.9458 & 0.6486 & 0.2512 & 0.3250 & 0.0904 & 0.8303 \\ 0.1635 & 0.5532 & 0.5465 & 0.9722 & 0.1597 & 0.0221 & 0.1620 & 0.0578 \\ 0.6691 & 0.1752 & 0.1052 & 0.6382 & 0.4226 & 0.6365 & 0.1629 & 0.6275 \\ 0.6674 & 0.1261 & 0.9745 & 0.5047 & 0.5198 & 0.9869 & 0.3162 & 0.8318 \\ 0.4392 & 0.1946 & 0.6600 & 0.7202 & 0.7759 & 0.1257 & 0.6940 & 0.8877 \\ 0.3429 & 0.4399 & 0.7629 & 0.5385 & 0.6283 & 0.6993 & 0.9477 & 0.8043 \\ 0.4811 & 0.4247 & 0.6468 & 0.4894 & 0.1014 & 0.8917 & 0.4742 & 0.9711 \end{pmatrix}. \quad (2)$$

Step 2. At the time of data collection, the first participant's data are independently augmented to x^* with six extra rows of normal random noise and a row of quality assurance data (see Table 7). The record is immediately masked and only the record-transformed data (A_0x^* shown in Table 8) are sent to the masking service provider. This is repeated for subject 11.

Step 3. The masking service provider chooses the column operator B_1, which is constructed to be block diagonal so that it keeps the first two columns invariant with the lower 6×6 block being transpose of the matrix generated by GenerateROM(536, 6):

$$B_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.6297 & 0.1342 & 0.3644 & 0.5092 & 0.3058 & 0.3160 \\ 0 & 0 & 0 & 0.2396 & 0.0874 & 0.7071 & -0.5469 & 0.2091 & 0.3038 \\ 0 & 0 & 0 & 0.5462 & -0.5825 & 0.1089 & 0.4796 & 0.3246 & 0.1233 \\ 0 & 0 & 0 & 0.1989 & 0.4472 & -0.5018 & -0.0815 & 0.6460 & 0.2912 \\ 0 & 0 & 0 & 0.3124 & 0.5508 & 0.3219 & 0.3360 & 0.0907 & -0.6118 \\ 0 & 0 & 0 & 0.3326 & 0.3629 & -0.0006 & 0.3036 & -0.5761 & 0.5775 \end{pmatrix}. \quad (3)$$

It applies attribute-transformation B_1, and sends the doubly masked data $A_0x^*B_1$ (see Table 9) to the data collectors.

Step 4. The data collectors left-multiply $A_0x^*B_1$ by A_0^{-1} to get back x^*B_1, extract the first row of x^*B_1 to get xB_1, aggregate data xB_1 from both participants to get XB_1. Then, the data collectors choose another key 537 to produce B_2, which has the same diagonal structure as B_1 but the lower 6×6 block is the transpose of the

Table 7: Two selected records of augmented data, x^*

Obs No	Response	Group	Δ	Age	BBS	IH	MIF	ADL/iADL
1	0	1	0.08	63.00	30.00	1	1	50.00
1	0	1	-0.73	-0.65	-1.52	0.10	0.17	0.18
1	0	1	0.43	-0.07	-1.34	-0.97	0.18	-0.07
1	0	1	-0.65	0.42	-0.20	1.84	0.96	0.44
1	0	1	0.30	0.13	-0.41	0.52	0.33	-0.37
1	0	1	0.56	0.26	0.63	0.49	0.02	-1.15
1	0	1	-0.23	-0.59	0.94	0.11	0.33	-0.14
1	0	1	777	777	777	777	777	777
11	1	1	0.15	43.00	9.00	1	0	50.00
11	1	1	-1.25	1.23	0.67	-0.15	-0.44	-0.03
11	1	1	-1.30	0.24	-1.76	1.70	-0.96	2.31
11	1	1	-0.35	0.49	1.14	0.70	-0.01	-1.10
11	1	1	-0.60	1.11	2.32	0.59	0.48	1.49
11	1	1	0.28	0.23	-0.36	0.85	1.33	0.94
11	1	1	-0.50	-0.03	-1.16	1.55	0.79	-0.36
11	1	1	888	888	888	888	888	888

matrix generated by GenerateROM2(537, 6):

$$B_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0986 & 0.4196 & -0.0204 & 0.6730 & -0.5015 & 0.3307 \\ 0 & 0 & 0 & -0.1314 & 0.6584 & 0.5712 & 0.0176 & 0.2706 & -0.3865 \\ 0 & 0 & 0 & -0.3756 & -0.2481 & 0.1295 & 0.5055 & 0.6290 & 0.3598 \\ 0 & 0 & 0 & 0.1437 & 0.4009 & -0.0136 & -0.4787 & 0.2090 & 0.7387 \\ 0 & 0 & 0 & 0.7082 & 0.1505 & -0.3814 & 0.2465 & 0.4783 & -0.2021 \\ 0 & 0 & 0 & 0.5566 & -0.3813 & 0.7147 & 0.0361 & -0.0854 & 0.1594 \end{pmatrix}. \quad (4)$$

Finally, the data collectors right-multiply XB_1 by B_2, and publish the selected rows of XB_1B_2 that correspond to the transformed data but not transformed noise (see Table 10) so that data users have access to the transformed data.

Table 8: Initially masked data for the two selected records, A_0x^*

Obs No	Response	Group	Δ	Age	BBS	IH	MIF	ADL/iADL
1	0	1	270.41	293.10	279.91	272.11	272.41	288.78
1	0	1	645.01	692.09	665.50	646.49	646.86	682.30
1	0	1	44.16	55.16	48.16	46.51	46.31	53.48
1	0	1	487.56	529.97	507.47	489.86	489.09	520.40
1	0	1	647.02	688.57	665.46	647.78	647.97	678.50
1	0	1	689.58	717.24	701.99	691.41	691.49	711.46
1	0	1	624.98	646.12	634.50	626.35	626.54	641.17
1	0	1	754.69	784.73	768.37	755.92	755.92	777.74
11	1	1	306.89	327.30	313.44	312.82	309.57	329.55
11	1	1	735.15	770.95	743.79	740.56	736.74	776.73
11	1	1	49.45	59.84	53.51	53.39	50.79	59.90
11	1	1	556.56	587.14	564.45	559.51	558.19	591.36
11	1	1	736.96	768.79	744.09	742.95	739.47	775.31
11	1	1	786.19	808.77	792.95	791.93	788.61	811.97
11	1	1	711.87	730.78	716.97	718.59	715.25	733.77
11	1	1	860.85	884.28	865.77	865.78	863.16	888.19

Table 9: Doubly masked data transmitted to the data collectors, $A_0x^*B_1$

Obs No	Response	Group	Δ	Age	BBS	IH	MIF	ADL/iADL
1	0	1	288.11	275.39	287.27	268.67	268.94	288.35
1	0	1	680.76	652.40	680.36	640.91	641.15	682.68
1	0	1	53.22	48.41	51.88	43.43	44.10	52.75
1	0	1	520.46	493.46	519.04	484.21	485.60	521.60
1	0	1	677.97	652.20	678.31	642.95	644.13	679.74
1	0	1	711.23	694.58	710.17	687.54	688.14	711.51
1	0	1	641.39	628.63	640.77	623.31	624.19	641.37
1	0	1	777.65	758.94	777.18	752.13	753.03	778.42
11	1	1	324.91	317.21	319.91	306.17	304.30	327.07
11	1	1	763.85	751.63	759.18	732.43	725.12	771.71
11	1	1	58.84	54.29	55.69	49.01	49.59	59.47
11	1	1	580.87	569.49	578.07	554.50	547.52	586.78
11	1	1	763.21	753.57	758.02	734.97	727.93	769.87
11	1	1	805.77	797.48	800.78	785.26	782.18	808.96
11	1	1	728.86	723.37	723.50	711.21	709.51	730.78
11	1	1	879.95	873.43	876.21	859.07	854.99	884.38

Table 10: Matrix-masked data released to data users, XB_1B_2

Obs No	Response	Group	Δ	Age	BBS	IH	MIF	ADL/iADL
1	0	1	10.27	-8.93	48.05	59.68	-0.26	36.26
11	1	1	1.38	-3.58	50.38	36.93	-4.24	22.28

4 Differences between TM^2 Method and Related Work

The TM^2 method is different from the standard frameworks in the literature on statistical confidentiality. Most disclosure limitation methods in previous research assume trustworthy data collectors who have full access to original data, and the goal of data masking is to prevent data users from obtaining confidential information. In this *trusted* model, data providers are willing to provide their sensitive information to data collectors. In our case, we assume an *untrusted model* treating everyone (including the data collectors) as potential intruders, and data providers are reluctant to share their sensitive information unless their answers will be used only in aggregate and cannot be linked back to them. The system is designed so that nobody other than data providers knows the original data.

Our method is an improvement of Warner's *randomized response* technique, which requests an interviewee to report whether or not his true binary answer to a sensitive question is the same as a randomly generated response that only the interviewee sees. Let π be the true proportion of interest (probability of "yes" answer to the sensitive question if truthfully disclosed) and c is the chance of "yes" answer from the random device. Then the probability of getting a "yes" response is $\lambda = \pi c + (1-\pi)(1-c)$. With n randomized responses, an unbiased estimator of λ is the sample proportion $\hat{\lambda}$, and hence the unbiased estimator of π is $\hat{\pi} = (c-1)/(2c-1) + \hat{\lambda}/(2c-1)$, with a variance $\{\pi(1-\pi) + 1/[16(c-0.5)^2 - 1/4]\}/n$. The data collectors may guess but cannot determine absolutely the interviewee's response.

Both Warner's technique and our TM^2 method meet the dual objectives of generating enough reliable data to yield fruitful inference and protecting respondents' privacy despite their truthful replies. However, Warner's randomized response technique is inefficient if there are ways to obtain truthful answers from all interviewees. Note that, when $\pi = 0.5$ and $c = 0.75$, the variance of $\hat{\pi}$ based on a randomized response survey is $1/n$, which is 4 times of the variance from a direct response survey, provided that all interviewees told the truth. The TM^2 method provides nearly the same privacy protection for interviewees as the Warner's technique, but it loses no efficiency for statistical inference of binary and normal data because sufficient statistics are preserved.

There are several other methods that are designed with the intention to collect data anonymously without revealing the providers' identities, including various cryptographic solutions (Yang et al., 2005; Gehrke, 2006; Fung et al., 2010) and anonymous communications (Chaum, 1981; Jakobsson et al., 2002; Brickell and Shmatikov, 2006). These methods try to achieve *unlinkability*, that is, they try to prevent data collectors and data users from learning which input came from which provider. But they do not hide the data values – they merely make it impossible (or very difficult) to link data

values to the providers. However, linkage attack can still occur in many situations. Dinur and Nissim (2003) showed that an attacker can reproduce the original database almost exactly based on queries answered with bounded noise. Dwork and Naor (2010) have several results stating that it is not possible to provide privacy and utility without making assumptions about how the data are generated. For example, they proved that it is not possible to publish anonymized data that prevents an attacker from learning information about people who are not even part of the data unless the anonymized data has very little utility or some assumptions are made about the attacker's background knowledge. For more information, see Kifer and Lin (2012) and Lin and Kifer (2014), which proposed a framework for extracting semantic guarantees from privacy definitions (or sets of data sanitizing algorithms). Also, as long as the raw sensitive data are collected and some people have access to them, leaking of private information is always a possibility due to unintentional mishandling or intentional transfer of data by those who have gained access; these mishaps occur even when de-identification and sanitizing before data release is done according to the current standard.

5 Conclusions

In this article, we propose the use of triple matrix-masking to protect participant privacy from the moment of data collection. The method lets the masking service provider and the data collectors separately hold keys for the generation of random matrices. It ensures that nobody other than the data providers sees the original data, but standard statistical analysis can still be performed with the same results from the masked data as from the original data. Therefore, confidentiality of the data and privacy of participants are well protected. In addition, an error checking mechanism is built in the data collection method to make sure that the data used for analysis are an appropriate transformation of the original data and a partial masking technique is introduced to grant data users access to non-sensitive personal information. The new technique holds the promise of removing the lack of trust obstacle and promoting privacy-preserving data collection. With the ever growing amount of data generated by electronic devices and the increasing demand for privacy protection, the method can be a great tool for survey research or clinical studies.

There are several relevant research questions not fully addressed in this article. First, further research is needed to evaluate the effectiveness of obtaining truthful answers using the new approach. Intuitively, people should be more willing to reveal truthful data if they know that nobody has access to their sensitive information. However, one drawback of the TM^2 method is that the masking service provider and the data collectors jointly can reconstruct exactly the individual records by sharing their keys, which is different from the randomized response technique of Warner (1965). Second, additional research is needed for developing methods to perform model-checking, missing data imputation, and data exploration under more complex models while maintaining limited data disclosure. We believe that the partial masking technique may offer help here. In many applications, it is enough for privacy protection to release the original main outcome while masking all other sensitive information. This will allow statistical

analysts to access residuals of the fitted model and to some extent perform model diagnostics.

Acknowledgement

The authors would like to thank the editor and two anonymous referees for their insightful comments and valuable suggestions, which led to development of improved security of the methods and much clearer explanations.

References

[1] Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley.

[2] American Association of Medical Colleges (2010). Report of the working group on information technology security and privacy in VA and NIH-sponsored research.

[3] Anderson, T. W., Olkin, I. & Underhill, L. G.(1987). Generation of random orthogonal matrices. *SIAM Journal of Scientific and Statistical Computing* **8**(4), 625-629.

[4] Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., & Talwar, K. (2007). Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems.*

[5] Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press. Reprinted (2007), New York: Springer.

[6] Blum, A., Dwork, C., McSherry, F., & Nissim, K. (2005). Practical privacy: The SuLQ framework. In *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS).* ACM Press. 128-138.

[7] Brickell, J. & Shmatikov, V. (2006). Efficient anonymity-preserving data collection. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* 76-85.

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

[9] Chaudhuri, A. & Mukerjee, R. (1987). *Randomized response: theory and techniques.* CRC Press, Marcel Dekker, Inc., New York.

[10] Chaum, D. (1981). Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM 24* **2**, 84-88.

[11] Chawla, S., Dwork, C.,McSherry, F., Smith, A. & Wee, H. (2005). *Towards Privacy in Public Databases, Theory of Cryptography Conference (TCC) 2005.* Cambridge, MA: Springer-Verlag, February, pp.556-577.

[12] Cox, L. H., Kelly, J. & Patil, R. (2004). Balancing quality and confidentiality for multivariate tabular data. In: J. Domingo-Ferrer and V. Torra (Eds.) *Privacy in Statistical Databases 2004*, Springer-Verlag, pp.87-98.

[13] Diaconis, P. (2005). What is a random matrix? *Notices of the AMS* **52**(11), 1348-1349.

[14] Dinur, I. & Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'03).* ACM Press. 202-210.

[15] Dobkin, B. H. & Dorsch, A. (2011). The promise of mHealth: daily activity monitoring and outcome assessments by wearable sensors. *Neurorehabilitation and Neural Repair* **25**(9), 788-798.

[16] Dobra, A. & Fienberg, S. E. (2009). The generalized shuttle algorithm. In P. Gibilisco, E. Riccomagno, M. Piera Rogantin, and H. P. Wynn, eds., *Algebraic and Geometric Methods in Statistics,* 135-156. Cambridge University Press.

[17] Dobra, A., Fienberg, S. E., Rinaldo, A., Slavkovic, A. B., & Zhou, Y. (2008). Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation, and disclosure limitation. In M. Putinar and S. Sullivant, eds., *Emerging Applications of Algebraic Geometry,* IMA Series in Applied Mathematics. New York: Springer. 63-88.

[18] Domingo-Ferrer, J., & Saygin, Y. eds. (2008). *Privacy in Statistical Databases, UNESCO Chair in Data Privacy International Conference, PSD 2008, Istanbul, Turkey, September 24-26, 2008. Proceedings, volume 5262 of Lecture Notes in Computer Science.* Springer, Berlin / Heidelberg.

[19] Du, W. S., Han, Y. S. & Chen, S. (2004). Privacy-preserving multivariate statistical analysis:linear regression and classification. *Proceedings 2004 SIAM International Conference on Data Mining (SDM04).*

[20] Duncan, T. D., Fienberg, S. E., Krishnan, R., Padman, R.,& Roehrig, S. F. (2001). Disclosure limitation methods and information loss for tabular data, in: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), *Confidentiality, Disclosure and Data Access,* North-Holland, Amsterdam, 135-166.

[21] Duncan, G. T. & Pearson, R. W. (1991). Enhancing access to microdata while protecting confidentiality: prospects for the future. *Statistical Science* **6**, 219-232.

[22] Duncan, P. W., Sullivan, K. J., Behrman, A. L., Azen, S. P., Wu, S. S., Nadeau, S. E., Dobkin, B. H., Rose, D. K., Tilson, J. K., Cen, S., Hayden, S. K., for The LEAPS Investigative Team. (2011). Body-weight-supported

treadmill rehabilitation after stroke. *New England Journal of Medicine* **364**(21), 2026-2036.

[23] Duncan, P. W., Wallace, D., Lai, S. M., Johnson, D., Embretson, S., & Laster, L. J. (1999). The stroke impact scale version 2.0. Evaluation of reliability, validity, and sensitivity to change. *Stroke* **30**(10), 2131-40.

[24] Dwork, C. (2006). Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, eds., *ICALP (2),* vol. 4052 of Lecture Notes in Computer Science. Berlin: Springer. 1-12.

[25] Dwork, C.& Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality* **2**(2), 93-107.

[26] Eaton, M. (1983). *Multivariate Statistics: A Vector Space Approach.* New York: Wiley.

[27] Fienberg, S. E. (1980). *The Analysis of Cross-classified Categorical Data.* Cambridge, MA: MIT Press. Reprinted (2007), New York: Springer.

[28] Fienberg, S. E., Fulp, W. , Slavkovic, A. & Wrobel, T. (2006). Secure log-linear and logistic regression analysis of distributed databases. In: Domingo-Ferrer, J., Franconi, L. (eds.) *Privacy in Statistical Databases PSD 2006.*, LNCS, vol. 4302, 277-290. Springer, Heidelberg.

[29] Fienberg, S. E., Nardi, Y. , & Slavkovic, A. B. (2009). Valid statistical analysis for logistic regression with multiple sources. In Gal, C.S., Kantor, P.B., Lesk, M.E., eds.,*Protecting Persons While Protecting the People*, Lecture Notes in Computer Science No. 5661, 82-94. Springer, Heidelberg.

[30] Fienberg, S. E., Rinaldo, A., & Yang, X. (2010). Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In J. Domingo-Ferrer and E. Magkos, eds., *Privacy in Statistical Databases 2010 (PSD 2010),* vol. 6344 of Lecture Notes in Computer Science. Berlin: Springer. 187-199.

[31] Fienberg, S. E. & Slavkovic, A. B. (2008). A survey of statistical approaches to preserving confidentiality of contingency table entries. In C. Aggarwal and P. S. Yu, eds., *Privacy Preserving Data Mining: Models and Algorithms.* New York: Springer. 289-310.

[32] Fung, B. C. M, Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-Preserving data publishing: A survey of recent developments. *ACM Comput. Surv.* **42**, 4, Article 14, 53 pages.

[33] Gehrke, J. (2006). Models and methods for privacy-preserving data publishing and analysis. *Tutorial at the 12th ACM SIGKDD.*

[34] Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J., & de Wolf, P. P. (1998). Post randomization for statistical disclosure control: theory and implementation. *Journal of Official Statistics* **14**, 463-478.

[35] Jakobsson, M., Juels, A., & Rivest, R. L. (2002). Making mix nets robust for electronic voting by randomized partial checking. *In Proceedings of the 11th USENIX Security Symposium* 339-353.

[36] Keller-McNulty, S. (1991). Comment on Duncan and Pearson, Enhancing Access to Microdata while Protecting Confidentiality: Prospects for the Future. *Statistical Science* **6**, 234-235.

[37] Kifer, D. & Lin, B.-R. (2012). An axiomatic view of statistical privacy and utility. *Journal of Privacy and Confidentiality* **4**(1), 5-49.

[38] Kim, J. (1986). A method for limiting disclosure in microdata based on random noise and transformation. *American Statistical Association Proceedings of the Section on Survey Research Methods* 370-374.

[39] Kim, J. & Winkler, W. (1995). Masking IRS income data on a merged file between 1990 CPS file and IRS income tax return file. *American Statistical Association Proceedings of the Section of Survey Research Methods* 114-119.

[40] Lin, B.-R.& Kifer, D. (2014). Towards a Systematic Analysis of Privacy Definitions. *Journal of Privacy and Confidentiality* **5**(2), 57-109.

[41] Liu, K., Kargupta, H., & Ryan, J. (2006). Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering* **18**(1), 92-106.

[42] Muralidhar, K. & Sarathy, R. (2006). Data shuffling: a new masking approach for numerical data. *Management Science* **52**, 658-670.

[43] Oganian, A., & Domingo-Ferrer, J. (2003). A posteriori disclosure risk measure for tabular data based on conditional entropy. *SORT - Statistics and Operations Research Transactions* **27**(2), 175-190.

[44] Ostapczuk, M., Musch, J., & Moshagen, M. (2009). A randomized-response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology* **39**, 920-931.

[45] Quercia, D., Leontiadis, I., McNamara, L., Mascolo, C., & Crowcroft, J. (2011). SpotME if you can: randomized responses for location obfuscation on mobile phones. In *Proceedings of the 31st International Conference on Distributed Computing Systems (ICDCS)*, 363-372.

[46] Rubin, D. B. (1993). Satisfying confidentiality constraints through the use of synthetic multiply imputed microdata. *Journal of Official Statistics* **9**, 461-468.

[47] Slavkovic, A. B. (2010). Partial information releases for confidential contingency table entries: present and future research efforts. *Journal of Privacy and Confidentiality* **1**(2), 253-264.

[48] Slavkovic, A. B. & Fienberg, S.E. (2009). Algebraic geometry of 2×2 contingency tables. In P. Gibilisco, E. Riccomagno, M.P. Rogantin, H.P. Wynn, eds., *Algebraic and Geometric Methods in Statistics,* pages 63-81. Cambridge University Press, UK.

[49] Slavkovic, A. B. & Lee, J. (2010). Synthetic two-way contingency tables that preserve conditional frequencies. *Statistical Methodology. Special Issue on Statistical Methods for Social Sciences* **7**(3), 225-239.

[50] Stewart, G. (1980). The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal of Numerical Analysis* **17**(3), 403-409.

[51] Ting, D., Fienberg, S. E., & Trottini, M. (2008). Random orthogonal matrix masking methodology for microdata release. *International Journal of Information and Computer Security* **2**(1), 86-105.

[52] Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* **60**, 63-69.

[53] Yang, X., Fienberg, S. E., & Rinaldo, A. (2012). Differential privacy for protecting multi-dimensional contingency table data: extensions and applications. *Journal of Privacy and Confidentiality* **4**(1), 101-125.

[54] Yang, Z., Zhong, S., & Wright, R. N. (2005). Anonymity-preserving data collection. In *Proceedings of the 11th ACM SIGKDD Conference.* ACM, New York, 334-343.

[55] Winkler, W. (2008). General discrete-data modeling methods for producing synthetic data with reduced re-identification risk that preserve analytic properties. *Statistics Research Report Series, 2010-02,* U.S. Bureau of the Census, Washington, DC.

Appendix 1. A Matlab Program for Generating Random Orthogonal Matrix

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%   The following function generates a p by p orthogonal operator,  %%
%%   which keeps the column vector of ones invariant, by the         %%
%%   Gram-Schmidt orthonormalization of a random normal matrix.      %%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function M = GenerateROM(SeedValue, p)
```

```
rng(SeedValue);
Z1 = [ones(p, 1) randn(p, p-1)]; M1 = GramSchmidtOrthonorm(Z1);
rng(SeedValue+2);
Z2 = [ones(p, 1) randn(p, p-1)]; M2 = GramSchmidtOrthonorm(Z2);
M = M1 * M2';

function M = GramSchmidtOrthonorm(Z)
[p, col] = size(Z); Y = []; M = [];
for i = 1:p
    v = Z(:, i); u = v;
    for j = 1:(i-1)
        y = Y(:, j);    u = u - (y' * v) / (y' *y) *y;
    end
    Y = [Y u];    M = [M u/sqrt(u'*u)];
end;
```