2017 IEEE 3rd International Conference on Big Data Security on Cloud

Collusion Resistant Multi-Matrix Masking for Privacy-Preserving Data Collection

Samuel S. Wu
Department of Biostatistics
University of Florida
Gainesville, FL 32610, USA
Email: samwu@biostat.ufl.edu

Shigang Chen
Department of Computer &
Information Science & Engineering
University of Florida
Gainesville, FL 32610, USA

Abhishek Bhattacharjee
Department of Statistics
University of Florida
Gainesville, FL 32610, USA

Ying He
Department of Mathematics
Clarkson University
Potsdam, NY 13699, USA

***Abstract*—An integral part of any social or medical research is the availability of reliable data. For the integrity of participants' responses, a secure environment for collecting sensitive data is required. This paper introduces a novel privacy-preserving data collection method: *collusion resistant multi-matrix masking* (CRM^3). The CRM^3 method requires multiple masking service providers (MSP), each generating its own random masking matrices. The key step is that each participant's data is randomly decomposed into the sum of component vectors, and each component vector is sent to the MSPs for masking in a different order. The CRM^3 method publicly releases two sets of masked data: one being right multiplied by random invertible matrices and the other being left multiplied by random orthogonal matrices. Both MSPs and the released data may be hosted on cloud platforms. Our data collection and release procedure is designed so that MSPs and the data collector are not able to derive the original participants' data hence providing strong privacy protection. However, statistical inference on parameters of interest will produce exactly the same results from the masked data as from the original data, under commonly used statistical methods such as general linear model, contingency table analysis, logistic regression, and Cox proportional hazard regression.**

***Index Terms*—Privacy-preserving data collection, orthogonal transformation, randomized response technique, item count technique, item sum technique, matrix-masking method.**

I. Introduction

The proliferation of mobile computing devices and the ubiquity of internet access have been feeding an information explosion, leading to the big-data era. Data collection for scientific, commercial or social research is increasingly performed online. However, privacy concern presents a major obstacle to data availability when people's confidential information is involved. This problem is particularly evident in medical research [1]. As an example, modern technologies provide novel information such as genetic signatures related to potential risk of genetic related disease onset, e.g., type I diabetes, system lupus erythematosus, etc. This information is critical to research for studying and curing such diseases. But leak of sensitive data, either intentional or unintentional, may result in potential problems to people who provide the data with good intention to facilitate medical advance. For one, insurance companies will be interested in such data to guard against future risks even though these people are healthy now and the disease of concern is only a possibility in the future. Besides the medical domain, privacy concern also arises in social studies as well as surveys and voting. For instance, employees may be reluctant to voice negativity on record against their managers due to fear of reprisal if confidentiality is not guaranteed. With such concern in mind, people may simply choose not to provide their data or if they do, provide wrong information, which adversely affects the availability or quality of data.

Recognizing the importance of data confidentiality, major efforts have been made through legislation, restrictions on data access, and de-identification technologies, with limited success. Again using the medical domain as an example, since the advent of the Health Insurance Portability and Accountability Act of 1996 and subsequent rulings, privacy protection has become an important legal issue. It often takes months to get approval from Institutional Review Board (IRB) before a medical study can be launched, and even then the use of data is subject to stringent restrictions. Yet, the lengthy approval and training process, coupled with de-identification and encryption before data release, can only provide *partial privacy protection* because as long as raw data is collected, patient information leak is always a possibility due to unintentional mishandling or intentional transfer of data by those who have gained access.

Research has strived to provide solutions that alleviate the adverse impact of privacy concern. Most existing methods are designed to share obfuscated data by entities who have already collected raw data. These methods are well suited for today's practice that collect raw data to data management centers and then obfuscate the data before releasing. However, they are ineffective against the security breach of the data management centers themselves, which face real threats from the cyberspace, as is evident from the recent well-publicized online stealing of credit card information from major retailers and hacks against bank servers [2], [3]. More generally speaking, once people give out their sensitive information to data centers, they lose control of the information.

This paper complements the current practice by proposing a new option of data collection for *strong privacy protection, ensuring that raw data stay with participants and only masked data are collected, which can be distributed and shared freely*. Not only will de-identification be performed directly by the devices of data providers right after data are produced, but also the data themselves are completely masked right away,

978-1-5090-6296-6/17 $31.00 © 2017 IEEE
DOI 10.1109/BigDataSecurity.2017.10

such that sensitive information can be transferred without fear of identification or data leak. Such technologies hold the promise of removing the trust obstacle, promoting objective data collection, and helping unrestricted sharing of big data.

More specifically, we propose a new procedure called *collusion-resistent multi-matrix masking (CRM^3)* that is performed at the time of data collection, starting from the moment before information leaves the data providers' devices. The data vector at each provider is first split randomly into a number of component vectors (whose sum equals the original vector). Each component vector is masked at the provider's device and then goes through a set of independent masking servers, which may be hosted in cloud. Different component vectors will go through the servers in different orders. The masked component vectors will be combined at a data collection center, which may also be hosted in cloud. After being masked again by the collection center, the combined vector will go through the masking servers to be partially de-masked. This process, which is completely transparent to the data providers once the data leaves their devices, achieves two important properties: (1) The final masked data can be freely shared without leaking any original data of the providers as long as the data collection center and all masking servers do not collude altogether, and (2) even though the masked data do not give out individual information, they can be made in such a way that statistical inference on parameters of interest can be conducted with the exactly same results on masked data as on the original data, under general linear model, chi-squared test, logistic regression, and other statistical methods. The second property distinguishes our work from Warner's *randomized response* technique [4], [5], which only gives approximate statistical results for binary input from data providers.

The rest of the paper is organized as follows. In Section 2, we summarize the related existing methods for privacy protection, including Warner's randomized response technique [4], the item count technique [6], the item sum technique [7] and the triple matrix-masking (TM^2) method introduced by Wu et al. [8]. In Section 3, we propose a collusion resistant data collection method, which overcomes the limitations of the existing methods and provides a more secure and trustworthy data collection environment. In Section 4, we extend the proposed method to handle missing data. Section 5 draws the conclusion.

II. Related Work

In most practice, the data collector use some methods to mask the sensitive raw data before sharing. The list of commonly used methods include addition of noise [9], [10], [11], multiple imputation [12], information preserving statistical obfuscation [13], post-randomization method [14], controlled tabular adjustment [15], [16], [17], data shuffling [18], random projection based perturbation [19], and random orthogonal matrix masking [20], [21], [22]. In each of these practiced method, the participants may be reluctant to reveal sensitive information to the data collectors or may question the security of the raw data before it is obscured. The main purpose of this research is to mask the sensitive participants data before it leaves the participants' data collection devices or reaches the data collector. Distributed data masking at patients presents new technical challenges that require novel solutions beyond the existing methods. A good example is secure multi-party computation (SMC) [23], [24], [25], which allows multiple data centers to jointly perform 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 our goal of strong privacy protection. One may argue that SMC can be directly performed amongst patent devices. But that will place significant computation overhead on patient devices. More importantly, all patients have to stand by ready for any statistical analysis that may happen years into the future; if patients ever leave a study, they will take their data away if we require that private data can never leave patient devices. This makes the SMC approach practically infeasible for medical studies that collect patient data over a long time.

There are several other methods to collect data anonymously without revealing the providers' identity, including various cryptographic solutions [26], [27], [28] and anonymous communications [29], [30], [31]. These methods achieve *unlinkability*, that is, they prevent data collector and data users from learning which input came from which provider. But they do not hide the data inputs. Our research also differs from traditional approaches of removing data provider identity. We observe that even after standard patient identifiers are removed, it is still sometimes possible to deduce the patient identities from the remaining medical data due to small count. In other words, as long as the raw sensitive data are collected and some people have access to them, leak of private information is always a possibility due to unintentional mishandling or intentional transfer of data by those who have gained access, even after current standard of de-identification and encryption before data release. And oftentimes patients are not willing to participate in research if their sensitive information may be exposed, even to the investigators. In contrast, our approach requires not only the removal of identifiers, but also the masking of all other data fields, making original data completely hidden.

In addition, with growing demand of cyber physical systems and social computing, a variety of approaches have been proposed in privacy protection. For example, "security-aware efficient data transmission" was designed for cloud-based Intelligent Transportation Systems with secure real-time multimedia data sharing and transferring [32]; "intercrossed access controls" was proposed to secure accesses between various media through the multiple cloud platforms [33]; and "spoofing-jamming attack strategy" was designed to maximize the adversarial effects using an optimal power distribution [34].

The only methods in the literature that have strong privacy protection are the *randomized response* technique proposed by

Warner [4], [5], the item count technique [6] and the item sum technique [7], and the triple matrix-masking (TM^2) method [8], which are summarized in the next three sub-sections.

A. Randomized Response Technique (RRT)

Warner's method requests an interviewee to report a binary answer, truthfully or untruthfully based on a preset probability distribution. Specifically, it 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 p is the chance of "yes" answer from the random device, then the probability of getting a "yes" response is $\lambda = \pi p + (1-\pi)(1-p)$. With n randomized responses, an unbiased estimator of λ is the sample proportion $\hat{\lambda}$, hence unbiased estimator of π is $\hat{\pi} = (p-1)/(2p-1) + \hat{\lambda}/(2p-1)$, with a variance $\{\pi(1-\pi) + 1/[16(p-0.5)^2] - 1/4\}/n$. 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.

The method is well summarized in a monograph by Chaudhuri and Mukerjee [35] and has been used in many applications [5], [36], [37]. 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 is inefficient, meaning that more samples will be needed to produce the same estimation accuracy when comparing with the raw data. For example, when $\pi = 0.5$ and $p = 0.75$, 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. In addition, it is only applicable to binary data.

B. Item Count Technique (ICT) and Item Sum Technique (IST)

In the item count technique [6], interviewees are asked to provide the number of items that apply to them, without answering the questions individually, where each question is about the applicability of an item. Respondents are randomly divided into two subsamples: one subsample only responds to a short list of nonsensitive questions; the other subsample answers a long list containing both nonsensitive questions and a sensitive question of interest. The population rate of the sensitive behavior can be estimated by the mean difference of answers between the two subsamples.

Trappmann et al. [7] extended the method to the measurement of quantitative sensitive variable. The item sum technique also generates two random subsamples; and they either answer a short list (SL) of nonsensitive questions or a long list (LL)containing nonsensitive questions and a sensitive question. However, IST requests interviewees to report the *sum* of the answers to the questions in theirs list, instead of the number of items that apply to them. For example, Trappmann et al. [7] conducted an experiment in which the following questions were used: 1) How many hours did you watch TV last week? (LL and SL); and 2) On average, how many hours per week do you engage in undeclared work? (LL only). Note that the sensitive information on undeclared work remains unknown at the individual level, but the mean difference of answers between the two subsamples provides an unbiased estimate of the amount of undeclared work.

The ICT and the IST have several advantages over the RRT: i) no randomizing device is required and hence they are easier to implement; ii) for many respondents, the techniques are a lot easier to understand than the complex probabilistic concept with the RRT. The later may lead a substantial proportion of respondents to provide a self-protective answer irrespective of the outcome of the randomizing device.

C. The TM^2 Method

The TM^2 method [8] is an improvement over RRT, ICT and IST in two aspects: (i) the method is applicable to survey of multiple sensitive questions and their relationships; and (ii) it losses no efficiency for statistical inference of binary and normal data because sufficient statistics are preserved.

The method relies on the well-known facts that: (i) orthogonally record-transformed data (AX) preserve sufficient statistics for parameters of interest with the use of general linear model and contingency table analysis; and (ii) Logistic regression can be conducted on the attribute-transformed data XB with the same results as the original data if the column operator B, which transforms data variables in X, keeps the response and treatment group variables invariant. More specifically, the least-squares estimates from the original and transformed data are the same when left-multiplying the data by an orthogonal matrix. Also, the estimate of the covariance matrix as well as all inference procedures will be identical. In addition, we can hide values of discrete variables while keep the contingency table counts by orthogonal transformation. Furthermore, we can obtain the exactly same maximum likelihood estimate of the treatment effects, their estimated standard errors, and most goodness of fit statistics in logistic regression based on the attribute-transformed data. Please refer to Ting et al. [22] and Wu et al. [8] for more details.

The basic idea behind the TM^2 approach is that a masking service provider only receives masked data from data providers and then applies another mask. The data collector who holds the key to the first mask partially decrypt the doubly masked data and apply a third mask before releasing 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 collector. 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.

The method involves four parties: data providers (participants), data collector (investigators), data users (data analysts and public), and a masking service provider, which is a private business or a government entity established to promote

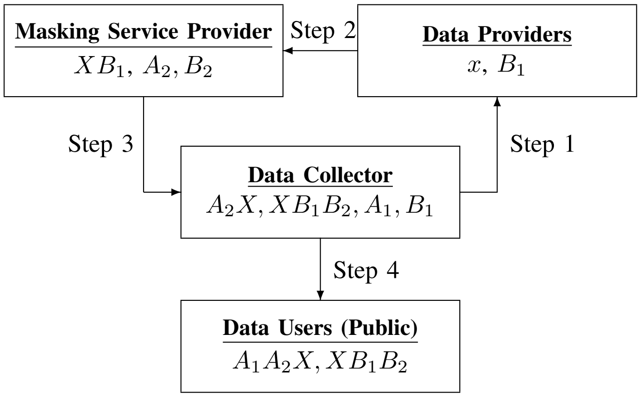

Fig. 1. The diagram above illustrates each entity's knowledge about the data and the masking matrices in the first TM^2 method. The masking service provider knows XB_1, the data collector knows A_2X, and two masked matrices (A_1A_2X & XB_1B_2) are available to everybody including the public. Nobody other than data providers (participants) knows the original data X.

security in data sharing (resembling those who provide public-key certificates for internet e-commerce). Specifically, let x be a $1 \times p$ vector containing a single participant's sensitive information and X be an $n \times p$ data matrix from a cohort of participants. The TM^2 method consists of the following steps:

Step 1. The data collector plans the data collection, creates the database structure, programs the data collection system; and chooses a key to generate a $p \times p$ 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 different keys to generate an $n \times n$ random orthogonal matrix A_2 and a random invertible matrix B_2. After receiving data from all participants, it aggregates the individual data into XB_1, and sends the doubly masked data A_2XB_1 and XB_1B_2 to the data collector after applying record and attribute transformations.

Step 4. The data collector multiplies A_2XB_1 by B_1^{-1} to get back A_2X, chooses another key to produce an $n \times n$ random orthogonal matrix A_1 and publishes A_1A_2X and XB_1B_2 , which are accessible by data users.

TM^2 however has a serious security problem: It is subject to simple collusion attacks. For instance, if the data collector colludes with the masking service provider, they together have all the keys for generating A_1, A_2, B_1 and B_2. Therefore, they can easily recover X from the masked data. For another example, if the masking service provider colludes with one of the data providers (or it registers as a data provider), it will know B_1 and therefore can recover X from XB_1 that it receives from the data providers.

III. Collusion Resistant Multi-Matrix Masking

In this section, we propose a new data collection procedure with strong privacy protection. In order to address the security problem of TM^2, the new procedure applies several technical innovations, involving multi-matrix masking for collusion resistant.

First, it is obviously insecure when the privacy of all participants' data depend on one secret matrix B_1 shared by all participants — defection of any participant breaks the security of the whole system. The question is, without using B_1 to mask, how will they send their data while preserving data privacy? Our idea is to split each participant's data x into a number k of randomly chosen component vectors, v_1 through v_k, such that $x = v_1 + ... + v_k$. Each component vector is sent to a different masking service provider, with the communication channel between the participant and each masking service provider being encrypted, through the popular TSL for example. There are k masking service providers, each of which knows only a randomized component vector. The whole system achieves k-privacy, meaning that the data privacy is compromised only when all k masking service providers collude.

Second, we want to make sure that, with multiple masking service providers, each holding a randomized piece of data from every participant, we are still able to integrate all such pieces into a masked form similar to the end results in Figure 1, AX and XB, where A $(= A_1A_2$ in the figure) is an orthogonal matrix co-generated from keys held separately by the data collector and all k masking service providers, and B $(= B_1B_2$ in the figure) is an invertible matrix co-generated from another set of keys held separately by the data collector and the k masking service providers. With AX and XB, we will be able to perform the general linear model, contingency table analysis, logistic regression and Cox regression, with the same results as if the original data were used, according to [22], [8].

It is non-trivial to design a scheme for multiple service providers to collaboratively work on random component vectors from many participants and integrate their respectively results into two masked matrices, AX and XB, without explicitly exchanging their masking matrices. This is a technical challenge that does not exist in TM^2.

The design of the new data collection system is described as follows: each masking service provider generates an $n \times n$ random orthogonal matrix for left multiplying masking and a $p \times p$ random invertible matrix for right multiplying masking. The right multiplying matrices from all service providers are commuting in product order. Let x be a $1 \times p$ vector containing a single participant's sensitive information and X be an $n \times p$ data matrix from a cohort of participants. The new masking method, illustrated in Figure 2, consists of the following steps:

Step 1. At the time of data collection, a participant's data vector x is randomly decomposed into a sum of k vectors: $x = v_1 + v_2 + \cdots + v_k$. More specifically, the jth element of v_i equals $w_{ij}x_j$, where $w = (w_{ij}, 1 \leq i \leq k, 1 \leq j \leq p)$ is a random weight matrix.

Step 2. The ith component of decomposed data (v_i) is first sent to ith masking service provider, who right multiplies it by B_i. Next, the masked data is sent to all other masking service providers for matrix masking. Then, the data masked by all masking service providers v_iB, where $B = \Pi_{i=1}^k B_i$ is the product of all right multiplying matrices (note that order of multiplication does not matter), is sent to data collector.

Step 3. Adding up all masked components $v_iB, 1 \leq i \leq k$, data collector gets xB. After receiving data from all participants, they aggregate the individual data into XB, which is sent back to masking service providers to remove right multiplying masking and to add left multiplying masking (details in the next step).

Step 4. The masked data XB is right multiplied by B_i^{-1} and left multiplied by A_i in sequence. The resulted data $AX = \Pi_{i=1}^k A_iX$ is sent to data collectors.

Step 5. The data collectors release AX and XB to data users.

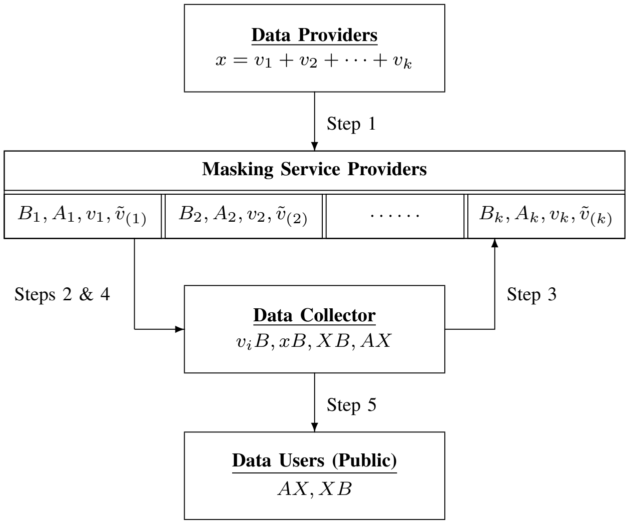
Data Providers
$x = v_1 + v_2 + \cdots + v_k$
Step 1
Masking Service Providers
$B_1, A_1, v_1, \tilde{v}_{(1)}$
$B_2, A_2, v_2, \tilde{v}_{(2)}$
$\cdots\cdots$
$B_k, A_k, v_k, \tilde{v}_{(k)}$
Steps 2 & 4
Step 3
Data Collector
v_iB, xB, XB, AX
Step 5
Data Users (Public)
AX, XB

Fig. 2. The diagram above illustrates each entity's knowledge about the data. The ith masking service provider generates a random invertible column operator B_i and a random orthogonal row operator A_i, knows a component of decomposed data v_i and data masked by other masking service providers $\tilde{v}_{(i)}$. The data collector knows masked data v_iB in addition to AX and XB, which are available to everybody including the public. Nobody other than data providers (participating patients) knows the original data X.

The orthogonal operators $A_i, i = 1, 2, \ldots, k$ can be obtained by the Gram-Schmidt orthonormalization of a random normal matrix, controlled by some random number generator seeds (i.e., keys). The resulting matrix is a draw from the uniform distribution on orthogonal matrices under the Haar measure (see p. 234 in [38]). Let M_1 and M_2 be Gram-Schmidt orthonormalization of $[1_n, Z_1]$ and $[1_n, Z_2]$, respectively. Note that both M_1 and M_2 have the first column vector parallel to 1_n, and $A = M_1M_2'$ transforms column vectors in M_2 to those in M_1. Therefore A is an orthogonal matrix that keeps 1_n invariant. More information about random orthogonal matrix can be found in [39], [40], [41].

There are two ways to obtain column operators $B_i, i = 1, 2, \ldots, k$ that are commuting in product. In the first approach, the masking service providers share a random invertible matrix T. And the ith masking service provider randomly generates a coefficient vector $(b_{i0}, b_{i1}, \ldots, b_{is})$ and let $B_i = \sum_{j=0}^s b_{ij}T^j$. In the second approach, the masking service providers share a random orthogonal matrix U. And the ith masking service provider randomly generates an invertible diagonal matrix $D_i = diag(d_{i1}, d_{i2}, \ldots, d_{ip})$ and let $B_i = UD_iU'$.

Moreover, the method can be improved to achieve $(k+1)$-privacy so that X can be recovered only when all $(k+1)$ parties (the data collector and the masking service providers) collude together. Specifically, the data collector generates an $n \times n$ random orthogonal matrix A_{k+1} for left multiplying masking and a $p \times p$ random invertible matrix B_{k+1} for right multiplying masking. And in Step 3, $A_{k+1}X\Pi_{i=1}^k B_i$ is sent back to masking service providers to remove right multiplying masking and to add left multiplying masking; while in Step 5, masked data $\Pi_{i=1}^{k+1} A_iX$ and $X\Pi_{i=1}^{k+1} B_i$ are released.

IV. Missing Data

We now extend the proposed multi-matrix masking method to account for nonresponse, where one or more data values are missing from a patient's vector x. One solution is to impute any missing value with the mean of that variable among other patients, which has the benefit of not changing the sample mean for that variable. [42] Recall that X is the matrix of the original data from the patients (participants). Our multi-matrix masking method will produce the masked data AX, while keeping the secrecy of X. In the following, we modify the method to produce $A\tilde{X}$ in the presence of missing data, where $\tilde{X}$ is the same as X except that the missing values are imputed by the means of the corresponding variables.

In Step 1, the missing values in a participant's data vector x are imputed by a predetermined constant c (say, 99999). In addition, the system collects an indicator vector y from each participant about the missing values, where an element equals 1 if the corresponding value is missing from the participant and 0 otherwise. The y vector is masked in the same way as the x vector, and hence the data collector receives yB after Step 3 and obtains AY after Step 4, where Y is an indicator matrix aggregating the row vectors (y) from all participants.

We choose A to be an orthogonal matrix that keeps 1_n invariant; see the previous section on how to produce such a matrix. We can mathematically generate $A\tilde{X}$ from AX and AY, without having to know X or Y, as follows: Let m be the row vector of sample means based on non-missing data. It is easy to check that, because A is an orthogonal matrix that keeps 1_n invariant, we have

$$m = 1_n'(AX - cAY)./(n - 1_n'AY),$$

where "./" is element-wise division. Next, we multiply each column of AY by the corresponding mean from vector m, with the resulting matrix denoted as V. In other words,

$$V = kron(m, 1_n). * AY,$$

where kron is the Kronecker tensor product and ".*" is element-wise multiplication. Finally, we are able to obtain the imputed (masked) matrix,

$$A\tilde{X} = AX - cAY + V.$$

V. Practical Discussions

Trappmann et al. [7] has shown that privacy-preserving data collection does improve participants' willingness to reveal sensitive information. It may be important to be able to convince the patients about a method's effectiveness. This is an educational issue, not a technical one, thus beyond the scope of this paper. We believe that it is not a fundamental problem in practice as long as the research community verifies a method and experts or authorities accept it; as an example, billions of people place their trust on experts and authorities when they participate in online commerce without knowing how the public key infrastructure works.

Another issue is the cost associated with multiple masking service providers. When multiple hospitals/institutes/PIs collaboratively collect data together, it will be convenient for each one of them sets up a server that acts as a masking service provider. To guard against their collusion, it will be necessary for an external trusted third party to act as an additional masking service provider. When a single PI collects data, the above setup still works, with the PI setting up a server for masking and a trusted third party acting as another masking provider. It will certainly strengthen security when more trusted third parties are used. But minimally each hospital/institute/PI only needs to set up one server, and a trusted third party is needed just as what the e-commence on the Internet needs for its security. A second approach to reduce cost is that many research projects jointly develop and share the privacy-preserving data collection system.

VI. Conclusions

A collusion resistant and privacy-preserving data collection method is proposed in this paper. Through the decomposition of participants' data into multiple components and masking them with multiple random matrices, the sensitive data are protected from the moment of data collection. Different masking service providers and the data collector separately hold keys for the generation of random matrices, hence ensuring collusion resistant privacy protection. It is imperative to note that only the data providers knows the original data, but standard statistical analysis can still be performed with the same results from the masked data as from the original data. In other words, the matrix-masking methods improve over RRT, ICT and IST in preserving statistical information on multiple sensitive questions as well as their relationships, while strongly protect privacy. 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.

Also, additional research is needed for developing methods to perform model-checking, and data exploration under more complex models while maintaining limited data disclosure. We believe 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 fitted model and to some extent perform model diagnostics.

References

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

[2] Huffington Post, *Citigroup: $2.7 Million Stolen From Customers As Result Of Hacking. http://www.huffingtonpost.com/2011/06/27/citigroup-hack_n_885045.html.* 2011.

[3] Reuters., *Target To Pay $10 Million To Settle Lawsuit From Massive Data Breach. http://www.huffingtonpost.com/2015/03/18/target-hack-settlement_n_6899290.html.* 2015.

[4] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Amer. Statist. Assoc.*, vol. 60, pp. 63–69, 1965.

[5] B. S. Everitt, "Randomized response technique," *Encyclopedia of Statistics in Behavioral Science*, 2005.

[6] J. Droitcour, R. A. Caspar, M. L. Hubbard, T. L. Parsely, and T. M. Visscher, W.and Ezzati, "The item count technique as a method of indirect questioning: A review of its development and a case study application," *In Measurement Errors in Surveys*, 1991.

[7] M. Trappmann, I. Krumpal, A. Kirchner, and B. Jann, "Item sum: A new technique for asking quantitative sensitive questions," *Journal of Survey Statistics and Methodology*, vol. 2, no. 1, pp. 58–77., 2014.

[8] S. S. Wu, S. Chen, D. Burr, and L. Zhang, "A new data collection technique for preserving privacy," *Journal of Privacy and Confidentiality*, p. ., (In press).

[9] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, "Towards privacy in public databases," *Theory of Cryptography Conference (TCC)*, pp. 556–577, Feburary 2005.

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

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

[12] D. B. Rubin, "Satisfying confidentiality constraints through the use of synthetic multiply imputed microdata," *Journal of Official Statistics*, vol. 9, pp. 461–468, 1993.

[13] J. Burridge, "Information preserving statistical obfuscation," *Statistics and Computing*, vol. 13, pp. 321–327, 2003.

[14] J. M. Gouweleeuw, P. Kooiman, L. C. R. J. Willenborg, and P. P. de Wolf, "Post randomization for statistical disclosure control: theory and implementation," *Journal of Official Statistics*, vol. 14, pp. 463–478, 1998.

[15] L. H. Cox, J. Kelly, and R. Patil, "Balancing quality and confidentiality for multivariate tabular data. in: J. domingo-ferrer and v. torra (eds.)," *Privacy in Statistical Databases 2004*, pp. 87–98, 2004.

[16] T. D. Duncan, S. E. Fienberg, R. Krishnan, R. Padman, and S. F. Roehrig, "Disclosure limitation methods and information loss for tabular data, in: P. doyle, j. lane, j. theeuwes, l. zayatz (eds.)," *Confidentiality, Disclosure and Data Access*, pp. 135–166, 2001.

[17] A. Oganian and J. Domingo-Ferrer, "A posteriori disclosure risk measure for tabular data based on conditional entropy," *SORT - Statistics and Operations Research Transactions*, vol. 27, no. 2, pp. 175–190, 2003.

[18] K. Muralidhar and R. Sarathy, "Data shuffling: a new masking approach for numerical data," *Management Science*, vol. 52, pp. 658–670, 2006.

[19] K. Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 92–106, 2006.

[20] G. T. Duncan and R. W. Pearson, "Enhancing access to microdata while protecting confidentiality: prospects for the future," *Statistical Science*, vol. 6, pp. 219–232, 1991.

[21] S. Keller-McNulty, "Comment on duncan and pearson, enhancing access to microdata while protecting confidentiality: Prospects for the future," *Statistical Science*, vol. 6, pp. 234–235, 1991.

[22] D. Ting, S. E. Fienberg, and M. Trottini, "Random orthogonal matrix masking methodology for microdata release," *International Journal of Information and Computer Securityroke*, vol. 2, no. 1, pp. 86–105, 2008.

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

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

[25] S. E. Fienberg, Y. Nardi, and A. B. Slavkovic, "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*, vol. 5661, pp. 82–94, 2009.

[26] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, 2010. Article 14, 53 pages.

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

[28] Z. Yang, S. Zhong, and R. N. Wright, "Proceedings of the 11th acm sigkdd conference," in *Anonymity-preserving data collection*, (New York), pp. 334–343, ACM, 2010.

[29] J. Brickell and V. Shmatikov, "Efficient anonymity-preserving data collection," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 76–85, 2006.

[30] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM 24*, vol. 2, pp. 84–88, 1981.

[31] M. Jakobsson, A. Juels, and R. L. Rivest, "Making mix nets robust for electronic voting by randomized partial checking," in *In Proceedings of the 11th USENIX Security Symposium*, pp. 339–353, 2002.

[32] K. Kai, L. Qiu, M. Chen, H. Zhao, and M. Qiu, "SA-EAST: Security-aware efficient data transmission for its in mobile heterogeneous cloud computing," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 16, no. 2, p. Article No. 60, 2017.

[33] Y. Li, K. Kai, Z. Ming, H. Zhao, and M. Qiu, "Intercrossed access controls for secure financial services on multimedia big data in cloud systems," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) - Special Section on Trust Management for Multimedia Big Data and Special Section on Best Papers of ACM Multimedia 2015*, vol. 12, no. 4s, p. Article No. 67, 2015.

[34] K. Kai, M. Qiu, Z. Ming, H. Zhao, and L. Qiu, "Spoofing-jamming attack strategy using optimal power distributions in wireless smart grid networks," *IEEE Transactions on Smart Grid*, vol. 99, 2017.

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

[36] M. Ostapczuk, J. Musch, and M. Moshagen, "A randomized-response investigation of the education effect in attitudes towards foreigners," *European Journal of Social Psychology*, vol. 39, pp. 920–931, 2009.

[37] D. Quercia, I. Leontiadis, L. McNamara, C. Mascolo, and J. Crowcroft, "Spotme if you can: randomized responses for location obfuscation on mobile phones," in *Proceedings of the 31st International Conference on Distributed Computing Systems (ICDCS)*, pp. 363–372, 2011.

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

[39] T. W. Anderson, I. Olkin, and L. G. Underhill, "Generation of random orthogonal matrices," *SIAM Journal of Scientific and Statistical Computing*, vol. 8, no. 4, pp. 625–629, 1987.

[40] P. Diaconis, "What is a random matrix?," *Notices of the AMS*, vol. 52, no. 11, pp. 1348–1349, 2005.

[41] G. Stewart, "The efficient generation of random orthogonal matrices with an application to condition estimators," *SIAM Journal of Numerical Analysis*, vol. 17, no. 3, pp. 403–409, 1980.

[42] R. Little and D. Rubin, *Statistical analysis with missing data, 2nd edition*. New York: Wiley & Sons, 2002.