2017 IEEE 37th International Conference on Distributed Computing Systems

Achieving Strong Privacy in Online Survey

You Zhou* Yian Zhou*‡ Shigang Chen* Samuel S. Wu†
*Department of Computer & Information Science & Engineering, University of Florida, Gainesville, FL, USA
†Department of Biostatistics, University of Florida, Gainesville, FL, USA
‡Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA, USA
Email: youzhou@cise.ufl.edu yianzhou@google.com sgchen@cise.ufl.edu samwu@biostat.ufl.edu

***Abstract*—Thanks to the proliferation of Internet access and modern digital and mobile devices, online survey has been flourishing into data collection of marketing, social, financial and medical studies. However, traditional data collection methods in online survey suffer from serious privacy issues. Existing privacy protection techniques are not adequate for online survey for lack of strong privacy. In this paper, we propose a practical strong privacy online survey scheme SPS based on a novel data collection technique called *dual matrix masking* (DM^2), which guarantees the correctness of the tallying results with low computation overhead, and achieves universal verifiability, robustness and strong privacy. We also propose a more robust scheme RSPS, which incorporates multiple distributed survey managers. The RSPS scheme preserves the nice properties of SPS, and further achieves robust strong privacy against joint collusion attack. Through extensive analyses, we demonstrate our proposed schemes can be efficiently applied to online survey with accuracy and strong privacy.**

I. Introduction

The proliferation of Internet access and modern digital and mobile devices has sparked interest in online data collection, which has resulted in a vast amount of online surveys conducted among different individuals, organizations and institutions across the world. According to the ESOMAR report [1], global marketing research spent over 40 billion dollars in 2013 to collect people's marketing strategies. Creating new and expanded meanings to surveys, online survey [2]–[5] has many advantages over the conventional data collection methods such as face-to-face, mail and telephone surveys: It can establish asynchronous contacts with respondents on the move, achieve faster, simpler and cheaper surveys, improve the quality of survey responses, extend contacts across national boundaries, and adjust through different situations.

However, great benefit comes with great risk. With its tremendous advantages, online survey also face major challenges. One serious concern is privacy [5]–[7]. As the collected data become more vulnerable and can be abused easier than before, many online survey participants worry about unauthorized disclosure of their submitted responses. Due to the lack of trust in confidentiality protection [8], they may refuse to participate in online survey or consent to research but purposely provide wrong information. Therefore, protecting the privacy of survey data is crucial to avoid reluctance in online survey participation. Today's common practice of online survey data management is to collect data by trustworthy parties such as authorized institutions and organizations, and store the data at their servers, which will be protected through means of access control, personnel training, encryption and de-identification. However, it has been demonstrated that these traditional methods can hardly provide the high level of confidence [9]. They are ineffective against internal attacks by system administrators, principal investigators or data analysts who have access to the raw data and intend to steal them for personal benefits.

We define two privacy models for online survey: ***weak privacy*** and ***strong privacy***. With weak privacy, although the raw survey data from each participant are collected, anonymous data submission is ensured by hiding the identity of each participant. However, linkage attacks [10] [11] can occur in small-count cases where the adversary is able to guess out the identities of some participants based on the raw survey data. With strong privacy, which is the subject studied in this paper, anonymous data submission is not required, but the raw data must be randomized before leaving the participant, and no one in the system can recover the raw data from the randomized data. In other words, *you know someone provides data, but you do not know what the data are*. We do not find any prior work that achieves strong privacy for online survey. There only exists work on weak privacy [12]–[15]: *you know what the data are, but you do not know which person provides which data.* This paper attempts to achieve strong privacy.

In this paper, we propose two practical strong privacy preserving schemes, ***SPS*** and ***RSPS***, to remove the major obstacles in online survey. Our SPS scheme is based on a novel and efficient data collection technique called *dual matrix masking* (DM^2). Ensuring that the raw response data stay with their original sources and the data collector (survey manager) only collects masked data, SPS retains the utility of the survey data from the tallying point of view, and guarantees strong privacy for individual participants because the raw survey response will not be available to any adversary. SPS also achieves the universal verifiability such that any participant can independently check whether the survey outcome corresponds to the published result. We then propose a more robust RSPS scheme, which incorporates multiple distributed survey managers. RSPS preserves the nice properties of SPS, and further achieves robust strong privacy against joint collusion attack. Through extensive analyses on correctness, efficiency, universal verifiability and robustness, we demonstrate our proposed schemes can be efficiently applied to online survey with accuracy and strong privacy.

1063-6927/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDCS.2017.247

The rest of the paper is organized as follows. Section II formalizes the problem. Section III gives some preliminary information. Section IV and Section V present our novel SPS and RSPS schemes, respectively. Section VI summarizes the related work. Section VII draws the conclusion.

II. Problem Formalization

A. System model

We consider an online survey system model as illustrated in Fig. 1, which consists of three software components: client, participant, and survey manager.

- **Client:** Representing a government, an organization, a company or an individual who wants to perform an online survey, the client-side software initiates the survey through a survey manager, receives the masked survey data via the manager, and produces the final tally from those data. The client-side software may be run by a third-party company specialized for online survey design and execution, or by a server set up by the party that conducts the survey. In either case, we conveniently use the term "client" for the entity carrying out these operations. We use V to denote a client.
- **Participant:** Representing people who are invited to participate in an online survey, the participant software (executed as a Javascript through browsers or as an app on mobile devices) masks the user's response data and submits the masked data through a survey manager. Depending on the context, we also use the term "participant" to refer to a person that submits survey data. Let U denote an arbitrary participant.
- **Survey Manager:** A server (set up by a commercial company or non-profit organization) for carrying out online surveys. It takes survey requests from the client, configures the surveys based on client requirements, invites and authenticates participants, collects valid response data from authorized participants, aggregates the received data, and forwards the aggregated data to the client. We will first consider the case of one survey manager and later expand it to multiple managers for better security. In the latter case, let s be the number of available managers, which are denoted as $M_1, M_2, \cdots, M_s$.

Each survey involves a client, a survey manager, and a certain number of participants. The client is responsible for the initialization of the online survey system. The participants can communicate with the client through a web interface or a mobile app. They need to first register themselves with the client to obtain masking keys. After that, the participants can use the masking keys to generate masked responses and submit them to the survey manager. The survey manager aggregates masked data from the participants into *group survey data*, which are then sent to the client. Finally, based on the received group data, the client will generate the final survey result.

B. Problem Statement

Under the above online survey system, we formally define our problem. The goal is to allow the client to efficiently obtain the survey result from the participants while protecting the privacy of each individual participant throughout the survey process. More specifically, the client first sets up an

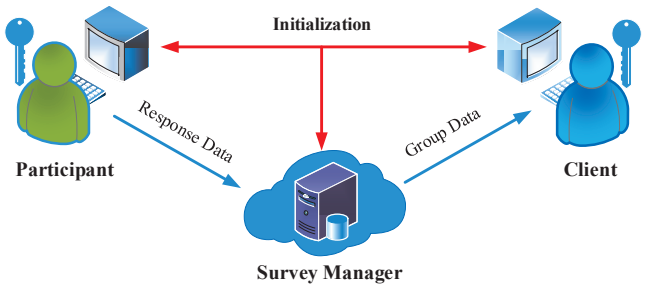

Fig. 1. Online survey system model.

appropriate web-based questionnaire, which contains a list of multiple-choice questions, each allowing the participants to select one or more response choices. During the data collecting period, the participants submit their responses to the survey manager, which will aggregate the responses into group data and transfer to the client. We want to achieve the following objectives:

1) *Accuracy*: The client should get the exact survey result of the questionnaire with no bias.
2) *Efficiency*: The computation overhead for the three parties in the online survey system should be as small as possible.
3) *Privacy*: The original survey response should never leave the participant's device. The adversary should never know about the original response of any individual participant. As discussed in the introduction, we define two privacy models for online survey, weak privacy and strong privacy. This paper attempts to achieve strong privacy.
4) *Fairness*: Each authenticated participant can only submit one valid response. If a participant submits multiple times, only one response should be counted into the final tally.
5) *Universal Verifiability*: Participants can check if the survey outcome corresponds to published result independently.

C. Threat Model

We assume a semi-honest model for the survey manager and the client, which follow the operations of the proposed scheme but are curious about the data reported from the participants. We also assume that the key sharing process between the client and the participants is secure, which can be achieved through secure authentication protocols and encrypted communications. Therefore, any external eavesdropper will not be able to learn the participants' masking keys. If an adversary compromises the survey manager, it does not have more capability than the survey manager in learning the participants' data. If an adversary compromises the client, it does not have more capability than the client in learning the participants' data. If an adversary compromises both the survey manager and the client, it is equivalent to a joint collusion attack, which we will introduce multiple survey managers to mitigate.

Our schemes are not designed to guard against participants from reporting wrong data to the survey manager. However,

if a compromised survey manager attempts to report wrong data to its client, or a compromised client attempts to report wrong survey result, our schemes can easily detect those misbehaviors. And even when these happen, our schemes still make sure that their design goal holds — no raw survey data is leaked.

Note that there are other active attacks that will affect the normal usage of the online survey system, such as denial-of-service (DoS) attack. Those attacks are beyond the scope of this paper. We focus on preventing privacy disclosure caused by the online survey scheme itself.

III. Preliminaries

Before presenting our novel schemes for strong privacy online survey, we first give some preliminary information, including the matrix masking technique, response formatting and a simple solution as well as its limitation.

A. Matrix Masking

Matrix masking, which refers to a class of statistical disclosure limitation (SDL) methods, is one of the most popular techniques used for data collecting and publishing with disclosure limitations [16]–[18]. It uses some specific matrices transforming a data matrix to a masked matrix via pre- and post- multiplication and a possible addition of noise or perturbations to protect the confidentiality of statistical data. For example, Duncan [16] proposes to transform an $n \times p$ data matrix X to the masked data of the form:

$$X \rightarrow AXB + C,$$

where matrix A is a row operator, matrix B is a column operator, and matrix C is the noise or perturbations added to the data. There are a wide variety of variations of the standard matrix masking approach. In our scheme, we adopt matrix masking as a building block to propose a novel *dual matrix masking* (DM^2) technique. In particular, two matrices A and B are used to mask the original data X:

$$X \rightarrow AXB, \tag{1}$$

where A is a random invertible matrix as a row operator, and B is a random invertible matrix as a column operator. Both masking matrices are only known to the participants and the client. We will prove that our online survey scheme based on DM^2 achieves strong privacy and guarantees the same final tallying result as the original data X.

Lemma III.1 (DM^2 - masked data disclosure limitation)**.** *The adversary cannot recover the original data X from the* DM^2 *masked data AXB if it has no knowledge about the masking matrices A and B.*

Proof: Since A and B are invertible, there exits a sequence of row operations and column operations such that

$$P_{s_A}^A \cdots P_2^A P_1^A A Q_1^A Q_2^A \cdots Q_{t_A}^A = E,$$
$$P_{s_B}^B \cdots P_2^B P_1^B B Q_1^B Q_2^B \cdots Q_{t_B}^B = E,$$

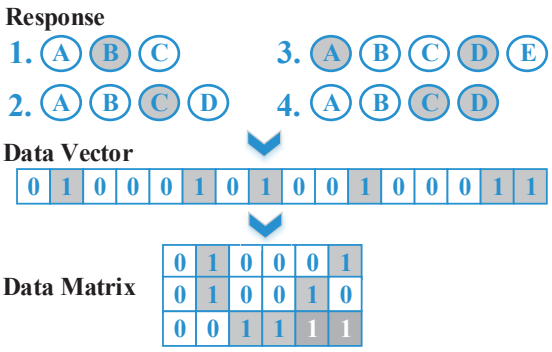

Fig. 2. An example of the response formatting. In this example, a response to $N = 16$ choices is first formatted to a $1 \times N$ data vector, which is then formatted to an $n \times p$ data matrix ($p = 6, n = 3$). The last 2 empty cells are both set to be 1 (flag parameters).

and then

$$\begin{aligned} A &= P_1^{A^{-1}} P_2^{A^{-1}} \cdots P_{s_A}^{A^{-1}} E Q_{t_A}^{A^{-1}} \cdots Q_2^{A^{-1}} Q_1^{A^{-1}}, \\ B &= P_1^{B^{-1}} P_2^{B^{-1}} \cdots P_{s_B}^{B^{-1}} E Q_{t_B}^{B^{-1}} \cdots Q_2^{B^{-1}} Q_1^{B^{-1}}, \end{aligned} \tag{2}$$

where $P_1^A, P_2^A, \cdots, P_{s_A}^A, P_1^B, P_2^B, \cdots, P_{s_B}^B$ and $Q_1^A, Q_2^A, \cdots, Q_{t_A}^A, Q_1^B, Q_2^B, \cdots, Q_{t_B}^B$ are elementary invertible matrices, and E is the unit matrix. Therefore, from (1) and (2), the masked data can be presented as follows:

$$\begin{aligned} AXB = P_1^{A^{-1}} \ldots P_{s_A}^{A^{-1}} Q_{t_A}^{A^{-1}} \ldots Q_1^{A^{-1}} X \\ P_1^{B^{-1}} \ldots P_{s_B}^{B^{-1}} Q_{t_B}^{B^{-1}} \ldots Q_1^{B^{-1}}. \end{aligned} \tag{3}$$

Since $P_1^{A^{-1}}, \cdots, P_{s_A}^{A^{-1}}$ and $Q_1^{A^{-1}}, \cdots, Q_{t_A}^{A^{-1}}$ are all random row operation elementary matrices used for random row switching, multiplication and adding, without knowledge of all these random matrices in advance, the adversary cannot get any information related to the rows of the original data X. Similarly, since $P_1^{B^{-1}}, \cdots, P_{s_B}^{B^{-1}}$ and $Q_1^{B^{-1}}, \cdots, Q_{t_B}^{B^{-1}}$ are all random column operation elementary matrices unknown to the adversary, the information related to the columns of the original data X is also protected. More specifically, if the original data X has a uniform distribution over $\mathbb{R}^{\mathfrak{D}}$, where $\mathfrak{D}$ is the dimensions of X, the masked data AXB also has a uniform distribution over $\mathbb{R}^{\mathfrak{D}}$. Since the masking matrices A and B are randomly selected from the set of all invertible matrices, from the adversary's point of view, it can only guess the original data X from $\mathbb{R}^{\mathfrak{D}}$ with equal probability. Therefore, the DM^2 technique protects the original information of X as long as the masking matrices A and B are unknown to the adversary. Note that the data which are masked in online survey will be made pseudo-random in our schemes. [17] provides more details, which can be used to prove that DM^2 is complete statistically defensible method of disclosure limitation. This completes the proof. ■

B. Response Formatting

As we described in Section II-B, the survey questionnaire contains several multiple-choice questions that may or may not specify the number of choices for the participants to select. For

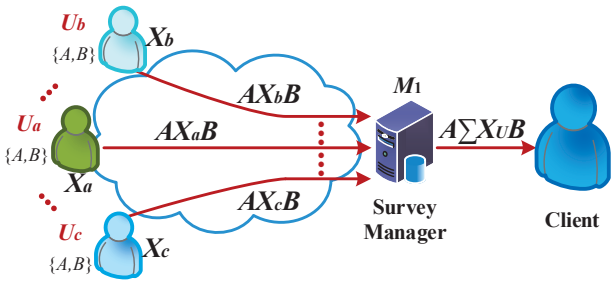

Fig. 3. An example of our simple solution.

the purpose of facilitating data transmission and processing, the response of a participant will first be formatted to an N-dimensional binary vector x, where each 0/1 value in x denotes the participant unselects/selects the corresponding choice and N is the total number of all choices in the questionnaire.

Sometimes, the value of N can be very large. Directly using the data vector x as input data X will cause a scalability problem since the DM^2 masking matrix B will be high-dimensional, i.e., $N \times N$. To improve the computation and communication efficiency, the N-dimensional binary vector x will be further formatted to an $n \times p$ binary matrix X by positioning all binary values from x to X in sequence and appending flag parameters to the last few empty slots (if any) in X. Here p is a parameter determined by the client and $n = \lceil \frac{N}{p} \rceil$. The flag appending will be discussed later. For now, we just fill the empty slots with 1's. An example of response formatting is shown in Fig. 2.

C. A Simple Solution and Its Limitation

With the response of each participant formatted into a data matrix X, we now propose a simple solution based on the DM^2 technique (i.e., $X \rightarrow AXB$). The solution contains the following three steps, and an example is illustrated in Fig. 3.

- **Step 1:** Client chooses keys to generate an $n \times n$ random invertible matrix A and a $p \times p$ random invertible matrix B, and distributes A and B to registered participants. For example, in Fig. 3, participants U_a, U_b and U_c all receive the masking matrices $\{A, B\}$ from the client.
- **Step 2:** Each participant takes the survey, and its response is formatted to an $n \times p$ data matrix X as discussed in Section III-B. Later X will be masked by the masking matrices A and B before leaving the participant's devices; only the masked data AXB are sent to the survey manager. As shown in Fig. 3, participants U_a, U_b and U_c send their masked data AX_aB, AX_bB and AX_cB to the survey manager M_1, respectively.
- **Step 3:** After receiving data from all participants, the survey manager aggregates these data into masked group data $AX_gB = \sum AX_UB = A\sum X_UB$, and sends to the client. Finally, the client employs de-masking matrices A^{-1} and B^{-1} to recover X_g (i.e., $\sum X_U$), which is the final result for this survey.

Through the DM^2 technique, the simple solution masks the original data X from the time of data generation at participant's devices. The original responses never leave the participants, and the tallying can be performed directly on the masked data. However, it may still leak some important information since all participants use the same masking matrices and the data matrix X comes from a small binary space. For instance, some participants may submit a same response. If one response is revealed, the adversary can obtain the responses of all people who submit the same response, which is not acceptable. To address this problem, our online survey schemes should take into consideration the following statements: (i) the participants with the same response should submit different masked data; (ii) the data space of X should be extended to $\mathbb{R}^{n \times p}$.

IV. Strong Privacy Online Survey Scheme With Single Survey Manager

In this section, we propose a ***S**trong **P**rivacy online **S**urvey scheme with single survey manager* (***SPS***) based on DM^2. We first describe SPS in detail, then analyze its correctness, efficiency, universal verifiability, robustness and strong privacy.

A. SPS: Strong Privacy Online Survey

Here we present our SPS scheme, which includes three phases: initialization, data collecting, and data tallying.

1) Initialization Phase: The initialization phase occurs at the beginning of the online survey. The client V first sets up the questionnaire, initializes the system parameters, and defines the data formats. The survey manager M_1 then configures a website for this questionnaire, and sends invitations to the participants.

Having accepted the invitation, a participant U will register with the client V using its identification information ID_U (e.g., email address). Upon receipt of U's registration request, the client generates a piece of registration information Reg_U, and sends U an encrypted certification $E(\kappa, Cert_U)$, a survey identifier SID_U, and some system parameters $Params = \{Seed, k, n, p\}$, where κ is a shared symmetric key between the client V and the survey manager M_1, $Seed$ is a universal random seed to generate DM^2 masking keys, k is a decomposition factor, n and p are the data matrix dimensions for response formatting as discussed in Section III-B. The SID_U serves an important purpose of assuring double-submission detection, which we will explain more later. Finally, the client stores the information $\{Reg_U, SID_U, Cert_U\}$, and shares the certification information $(SID_U \,|\, Cert_U)$ with the survey manager M_1 for it to later verify participants. The initialization phase is summarized below:

$$\boldsymbol{U} \xrightarrow{ID_U} \boldsymbol{V} \xrightarrow{(Params \,|\, SID_U \,|\, E(\kappa, Cert_U))} \boldsymbol{U}$$
$$\downarrow (SID_U \,|\, Cert_U)$$
$$\boldsymbol{M_1}$$

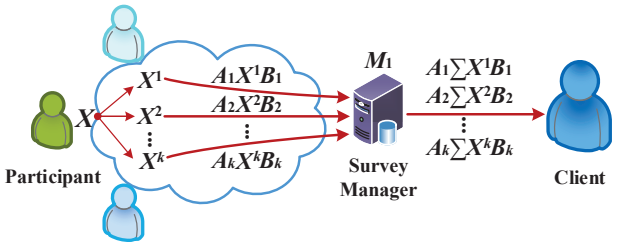

Fig. 4. An example of our SPS scheme.

2) Data Collecting Phase: In this phase, participants submit masked survey data to the survey manager M_1. The survey manager gathers the masked data, further generates masked group survey data, and sends to the client V. There are four steps as shown in the following. An example for the data collecting phase with a particular participant is given in Fig. 4.

a) Step 1: Participants submit masked survey data. From the previous initialization phase, each participant U receives $Params$, SID_U, and $E(\kappa, Cert_U)$ from the client. It then uses the random seed $Seed$ to generate a set of $n \times n$ random invertible matrices $\{A_i\}_{i\in[1,k]}$ and a set of $p \times p$ random invertible matrices $\{B_i\}_{i\in[1,k]}$ as its DM^2 masking keys. Note that the $Seed$ for each participant U is the same, so the DM^2 masking keys are also the same for all participants.

During the survey process, the participant U takes the survey and the response is formatted into an $n \times p$ data matrix X_U as described in Section III-B. Then, the data matrix X_U is randomly decomposed into k $n \times p$ matrices $\{X_U^1, X_U^2, ..., X_U^k\}$ such that $X_U = X_U^1 + X_U^2 + \cdots + X_U^k$. More specifically, $\forall 1 \le i \le k$, $X_U^i[u][v] = w_{uv}^i X_U[u][v]$, where $w = \{w^i | 1 \le i \le k\}$ with each w^i denoting an $n \times p$ matrix of random weight parameters such that $\sum_{i=1}^{k} w_{uv}^i = 1$, $\forall 1 \le u \le n, 1 \le v \le p$. Through this, the original binary data matrix X_U is converted to k data matrices in the $\mathbb{R}^{n\times p}$ space.

After that, each decomposition data matrix X_U^i $(1 \le i \le k)$ is left multiplied by A_i and right multiplied by B_i, which generates the masked survey data MD_U of U:

$$MD_U = \{MD_U^i \mid MD_U^i = A_i X_U^i B_i,\ i \in [1,k]\}. \quad (4)$$

Finally, U submits its result by sending MD_U, SID_U and $E(\kappa, Cert_U)$ in a message Msg_1 to the survey manager M_1:

$$\boldsymbol{U} \xrightarrow{Msg_1 = (MD_U \,|\, SID_U \,|\, E(\kappa, Cert_U))} \boldsymbol{M_1}$$

b) Step 2: Survey manager verifies each submission. Upon receiving a submission message Msg_1 from a participant, the survey manager M_1 first decrypt the certification information with the shared key κ:

$$D(\kappa, E(\kappa, Cert_U)) = (Cert_U), \quad (5)$$

and obtain the information $(MD_U|SID_U|Cert_U)$. After that, the survey manager looks up the $(SID_U|Cert_U)$ in its database, and checks whether the survey identifier SID_U is in the set S, which contains the SID of all already-responded participants to detect potential double submission. If the $(SID_U|Cert_U)$ exists in the survey manager's database, and SID_U does not exist in S, it means this submission is the first response of a registered participant. So the survey manager will accept the response, store MD_U, and insert SID_U into S. Otherwise, the survey manager will reject the submission and do nothing. Note that it can be easily adjusted to support response update: the survey manager will only keep the latest response of each participant. We omit this for space limitation.

c) Step 3: Survey manager generates masked group data. When the survey period ends, the survey manager M_1 finishes collecting all participants' masked survey data MD_U. Next, it will aggregate these data to generate the masked group data GD. Suppose in the online survey system, there are m participants $\{U_1, U_2, \cdots, U_m\}$ who have submitted valid survey responses (i.e., their survey identifiers are in the set S), and their masked survey data are $\{MD_{U_1}, MD_{U_2}, \cdots, MD_{U_m}\}$. The masked group survey data GD contains k $n \times p$ matrices, which is the sum of the masked survey data:

$$GD = \left\{ GD^i \,\middle|\, GD^i = \sum_{j=1}^{m} MD_{U_j}^i = \sum_{j=1}^{m} A_i X_{U_j}^i B_i,\ i \in [1,k] \right\} \quad (6)$$

d) Step 4: Survey manager uploads masked group survey data. Now the survey manager M_1 has generated the masked group survey data GD. Next, it will send GD with the set S in a message Msg_2 to the client V:

$$\boldsymbol{M_1} \xrightarrow{Msg_2 = (GD \,|\, S)} \boldsymbol{V}$$

3) Data Tallying Phase: Upon receiving the message Msg_2, the client fetches the $Seed$ to obtain the DM^2 masking keys used by the participants, $\{A_i\}_{i\in[1,k]}$ and $\{B_i\}_{i\in[1,k]}$, and then computes their inverse matrices, $\{A_i^{-1}\}_{i\in[1,k]}$ and $\{B_i^{-1}\}_{i\in[1,k]}$. To obtain the aggregate tally of the responses from all participants, the client simply left multiplies the ith element of GD (i.e., GD^i) by A_i^{-1}, and then right multiplies it by B_i^{-1} to recover the ith de-masked group survey data G^i:

$$\begin{aligned} G^i &= A_i^{-1} \times GD^i \times B_i^{-1} \\ &= A_i^{-1} \times \sum_{j=1}^{m} A_i X_{U_j}^i B_i \times B_i^{-1} = \sum_{j=1}^{m} X_{U_j}^i. \end{aligned} \quad (7)$$

The final tally result is simply

$$G = \sum_{i=1}^{k} G^i. \quad (8)$$

In the next subsection, we will prove $G = R$, where $R = \sum_{j=1}^{m} X_{U_j}$ is the sum of the original data matrices of all participants who have submitted valid survey data. Finally, the client verifies if the number of participants matches the number of elements in the set S. If they match, the client obtains the final result on how many participants have selected each individual question choice. It will also publish the final tally and the set S if necessary.

B. Correctness

We prove that the final tally G of the client equals the actual survey result $R = \sum_{j=1}^{m} X_{U_j}$.

Proof: From (8) and (7), we have

$$\begin{aligned} G &= \sum_{i=1}^{k} G^i = \sum_{i=1}^{k} \sum_{j=1}^{m} X_{U_j}^i \\ &= \sum_{j=1}^{m} \sum_{i=1}^{k} X_{U_j}^i = \sum_{j=1}^{m} X_{U_j} = R. \end{aligned} \tag{9}$$

This completes the proof. ∎

C. Efficiency

Suppose there are a total of m participants. The original response x of each participant is a $1 \times N$ vector, and the formatted data matrix X is $n \times p$. The computation overhead of each entity in SPS is given as follows.

1) Participant: Each participant U first needs to perform response formatting (from x to X), whose cost is $O(n \times p)$. Then U splits the response X to k $n \times p$ data matrices, whose cost is $O(k \times n \times p)$. Finally, U generates k masked data $\{AX^iB\}_{1 \le i \le k}$ by left- and right- matrix multiplication, whose computation cost is $O(k \times (n^2 \times p + n \times p^2))$. So the total computation overhead for each participant U is $O(kN(n+p))$.

2) Survey manager: The survey manager M_1 only needs to aggregate m participants' masked data (k $n \times p$ matrices). Clearly, its total computation overhead is $O(k \times m \times N)$.

3) Client: For each masked group data GD^i, the client V needs to left multiply it by an $n \times n$ matrix A_i^{-1} and right multiply it by a $p \times p$ matrix B_i^{-1} to recover the de-masked $n \times p$ group data G^i, whose cost is $O(k \times (n^2 \times p + n \times p^2))$. Then it adds up all k de-masked group data, whose cost is $O(k \times n \times p)$. So its total computation overhead is $O(kN(n+p))$.

Since k is a constant number far smaller than m and N, the computation complexity for each participant, the survey manager, and the client are actually $O((n+p) \times N)$, $O(m \times N)$, and $O((n+p) \times N)$, respectively. One can see that our SPS scheme is indeed very efficient.

D. Universal Verifiability

When the client V publishes the final result, the participants should be able to verify it. Our SPS scheme achieves the universal verifiability property: any participant U can independently check if the survey outcome corresponds to the result published by the client. This optional verification occurs after V publishes the result, and it includes two steps as follows.

Step 1 - Publishing Information: After accomplishing the data tallying task, the client V publishes the tally result R_V of the survey. Also, the survey manager M_1 will publish the masked group survey data $GD_M = \{GD_M^i\}_{1 \le i \le k}$.

Step 2 - Verifying Result: Each participant U can independently perform the same procedure to verify the correctness of the published result R_V. More specifically, U first computes the inverse of the DM^2 masking keys, $\{A_i^{-1}\}_{i \in [1,k]}$ and $\{B_i^{-1}\}_{i \in [1,k]}$, and then calculates the de-masked group survey data G_U similar to (7),

$$G_U = \{G_U^i \mid G_U^i = A_i^{-1} GD_M^i B_i^{-1},\ i \in [1,k]\}, \tag{10}$$

and the survey result $R_U = \sum_{i=1}^{k} G_U^i$. Finally, the participant U can verify the final result by comparing R_U with R_V. If $R_U = R_V$, U verifies that the final result is correct.

E. Robustness

1) Double submission detection: In SPS, only registered participants are allowed to submit responses, and each participant can only submit one valid response. In particular, each participant obtains a unique survey identifier SID from the client, which must accompany its response for submission. The survey manager keeps track of a set S of the survey identifiers of already-responded participants. If a participant U submits multiple times, the survey manager will only accept one response and ignore others. Even if the survey manager is compromised to cooperate with the participant to cast multiple submissions, the client can easily detect this cheating behavior through mismatched number of participants and the final tally. Therefore, SPS achieves double submission detection.

2) Submission tamper detection: If a compromised survey manager attempts to duplicate, modify or replace any individual response, the number of participants and the final tally will be mismatched during the data tallying phase, which can be easily detected by the client. The survey manager cannot remove any individual response either. When the client publishes the set S, any participant U who has submitted its response can check if its survey identifer SID_U is in the set S. If $SID_U \notin S$, then U detects its response has been removed and report to the client V. In summary, no entity can modify, duplicate, or remove any individual submission without being detected.

3) Error detection: In the response formatting, the original response x, which is a $1 \times N$ vector, is converted to an $n \times p$ data matrix X. The remaining $r = n \times p - N$ empty slots in X can be filled by some flag parameters for error detection. For example, each empty slot can be set as a constant c_i for decomposition X^i. In the data tallying phase, the client can check whether the value of this slot in G^i equals $m \times c_i$, where m is the total number of participants. If it does not match, the client will detect the error in decomposition X^i.

F. Privacy Analysis

We now demonstrate that SPS achieves strong privacy under the pre-defined threat model in Section II-C: the adversary compromises either the survey manager or the client, but not both. When the adversary compromises both the survey manager and the client, it is equivalent to a joint collusion attack, which we will introduce a more robust scheme RSPS with multiple survey managers to mitigate in the next section.

1) Strong privacy: As described in Section II-C, any external eavesdropper will not be able to learn the participants' masking keys. If an adversary compromises the client, it does not have more capability than the client in learning the

participants' data. If an adversary compromises the survey manager, it does not have more capability than the survey manager in learning the participants' data.

In the former case where the adversary compromises the client, since the client V only gets the masked group response of all participants from the survey manager M_1, the only information that the adversary can obtain is the masked group response and final results, from which it cannot derive any original response of any individual participant.

Consider the latter case where the adversary compromises the survey manager. Recall that in SPS, the response data matrix X of each participant is split into k decomposition matrices $\{X^i\}_{i\in[1,k]}$, with each X^i being masked by a different pair of DM^2 masking keys $\{A_i, B_i\}$. Each participant U only submits its masked survey data $MD_U = \{A_i X_U^i B_i\}_{i\in[1,k]}$ to the survey manager M_1. Therefore, through compromising the survey manager, the adversary can only get MD_U, from which the adversary still cannot extract U's original response. We will demonstrate this using a game between a challenger and an adversary $\mathcal{A}$. The game proceeds in the following:

- The survey response space $\mathcal{D} = \{SR_1, SR_2, \cdots, SR_\alpha\}$ is available to both the adversary $\mathcal{A}$ and the challenger. The adversary knows that there is only one participant U in the game, and the challenger knows U's DM^2 masking matrices, $\{A_i\}_{i\in[1,k]}$ and $\{B_i\}_{i\in[1,k]}$.
- At some time, the challenger picks a random index $j \in \{1, 2, \cdots, \alpha\}$ to obtain the survey data $X_U = SR_j$, and randomly split X_U into k matrices $\{X_U^i\}_{i\in[1,k]}$ such that $X_U = \sum_{i=1}^{k} X_U^i$. Then the challenger generates a message $\Theta = \{\Theta^i \mid \Theta^i = A_i X_U^i B_i,\ i \in [1,k]\}$, and sends Θ to the adversary $\mathcal{A}$.
- After receiving the message Θ, the adversary $\mathcal{A}$ returns a guess $j^* \in \{1, \cdots, \alpha\}$ on j and wins the game if $j^* = j$.

Definition IV.1. *We define that the advantage of $\mathcal{A}$ breaking the strong privacy property of Θ is*

$$\mathbf{Adv}_{\mathcal{A}} = \alpha \cdot \left(\Pr[j = j^*] - \frac{1}{\alpha}\right) = \alpha \cdot \Pr[j = j^*] - 1.$$

The strong privacy of Θ is achieved if the advantage $\mathbf{Adv}_{\mathcal{A}}$ is negligible for an arbitrary adversary $\mathcal{A}$. Furthermore, if $\mathbf{Adv}_{\mathcal{A}}$ is exactly 0, then the strong privacy is unconditional. We now prove that the information Θ achieves the unconditional strong privacy in the following theorem.

Theorem IV.1. *The information Θ achieves unconditional strong privacy through the decomposition DM^2 masking.*

Proof: From Lemma III.1, with the DM^2 technique, the adversary cannot recover the original data X from the masked data AXB if it has no knowledge about the masking matrices A and B. Since the adversary has no knowledge about U's DM^2 masking matrices, $\{A_i\}_{i\in[1,k]}$ and $\{B_i\}_{i\in[1,k]}$, it cannot extract any X_U^i from $\Theta^i = A_i X_U^i B_i$, and thereby cannot obtain the survey response X_U picked by the challenger. In the eye of the adversary, each survey response in $\mathcal{D}$ is equally suspicious by the decomposition DM^2 masking. Therefore, $\Pr[j = j^*] = \frac{1}{\alpha}$. Then, by Definition IV.1, we have

$$\mathbf{Adv}_{\mathcal{A}} = \alpha \cdot \Pr[j = j^*] - 1 = \alpha \cdot \frac{1}{\alpha} - 1 = 0. \tag{11}$$

This completes the proof. ■

From Theorem IV.1, the information Θ achieves the unconditional strong privacy, which means the information MD_U submitted by a participant U to the survey manager M_1 in SPS also achieves the unconditional strong privacy. Therefore, even if the adversary compromises the survey manager, the adversary still cannot derive any original response of any participant.

2) Joint collusion attack: We have demonstrated that SPS preserves strong privacy when the adversary compromises either the client or the survey manager. When the adversary compromises both the survey manager and the client, it is equivalent to a joint collusion attack, where the compromised survey manager M_1 cooperates with the compromised client V to gain the confidential survey response of a targeted group of one or more participants. More specifically, the survey manager M_1 acquires the DM^2 masking keys $\{A_i\}_{i\in[1,k]}$ and $\{B_i\}_{i\in[1,k]}$ from the client V, thereby it can get their inverse matrices and further calculate $X^1, X^2, \cdots, X^k$ from the masked data MD of any participant to recover its original survey data X, which is equal to $X^1 + X^2 + \cdots + X^k$. In other words, the original response of every participant will be revealed to the joint adversaries. To guard against the joint collusion attack, we propose a more robust strong privacy scheme RSPS with multiple survey managers, which will be discussed in the next section.

V. Robust Strong Privacy Online Survey Scheme With Distributed Survey Managers

In this section, we propose a ***R**obust **S**trong **P**rivacy online **S**urvey scheme with distributed survey managers* (**RSPS**). We first introduce the key idea of additive secret sharing, and then describe RSPS in detail and analyze its properties.

A. Additive Secret Sharing

Secret sharing refers to securely distributing a secret among a group of entities, each of whom is allocated a share of the secret. More specifically, there is one dealer and some players. The dealer gives a share of the secret to each player by a sharing algorithm $Share$. The secret can be reconstructed from all the shares by the algorithm Rec only when specific conditions are fulfilled. Different implementations of the $Share$ algorithm will lead to different secret sharing schemes with different security properties. In our RSPS scheme, we design a specific additive secret sharing among s distributed survey managers. A secret data matrix X is split to s shares $X^1, X^2, \cdots, X^s$ such that

$$X^1 + X^2 + \cdots + X^s = X, \tag{12}$$

where $X^1, X^2, \cdots, X^s$ are determined by Algorithm 1.

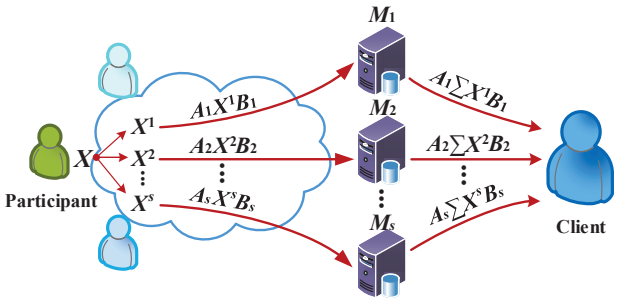

Fig. 5. An example of our RSPS scheme.

Algorithm 1 Share algorithm of RSPS additive secret sharing

1: **Inputs:** X, s, n, p
2: **for** $i \leftarrow 1$ to $s-1$ **do**
3: $\quad X^i \leftarrow$ uniformly chosen from $\mathbb{R}^{n\times p}$
4: **end for**
5: $X^s = X - \sum_{i=1}^{s-1} X^i$

B. RSPS: Robust Strong Privacy Online Survey

In order to further improve the confidentiality protection of the participants, we utilize s distributed survey managers to propose a more robust RSPS scheme. Below we will explain the three phases in RSPS: initialization, distributed data collecting, and data tallying. An example of RSPS is illustrated in Fig. 5.

1) Initialization Phase: Similar to SPS, in the initialization phase, the client initializes the online survey system, and segregates the data collection task to s distributed survey managers. Specifically, each survey manager is responsible for collecting some part of the survey data of all participants. Each participant U does almost the same work as in the SPS: It obtains the system parameters $Params = \{Seed, s, n, p\}$, a survey identifier SID_U, and a series of encrypted certifications $(E(\kappa^1, Cert_U^1)|\cdots|E(\kappa^s, Cert_U^s))$ from the client V, where κ^i is a shared symmetric key between V and the survey manager M_i, $i \in [1, s]$. Finally, the client V also shares the certification information $(SID_U|Cert_U^i)$ with the survey manager M_i for it to later verify participants. The initialization phase of RSPS is summarized in the following:

$$\boldsymbol{U} \xrightarrow{ID_U} \boldsymbol{V} \xrightarrow{(Params|SID_U|E(\kappa^1,Cert_U^1)|\cdots|E(\kappa^s,Cert_U^s))} \boldsymbol{U}$$
$$\downarrow (SID_U \,|\, Cert_U^i)$$
$$\boldsymbol{M_i}$$

2) Distributed Data Collecting Phase: In this phase, each participant U takes the survey and formats its response to a data matrix X_U similar to SPS, and then applies Algorithm 1 to obtain s shares $X_U^1, \cdots, X_U^s$. Next, U uses the random seed $Seed$ to generate a set of random invertible matrices $\{A_i\}_{i\in[1,s]}$ and $\{B_i\}_{i\in[1,s]}$ as its DM^2 masking keys to produce its masked survey data $MD_U = \{A_i X_U^i B_i\}_{i\in[1,s]}$. Finally, for each survey manager M_i, U distributes the ith masked share of MD_U, $MD_U^i = A_i X_U^i B_i$, with SID_U and $E(\kappa^i, Cert_U^i)$ in a message Msg_1^i to the survey manager M_i:

$$\boldsymbol{U} \xrightarrow{Msg_1^i = (MD_U^i \,|\, SID_U \,|\, E(\kappa^i, Cert_U^i))} \boldsymbol{M_i}$$

Upon receiving a message Msg_1^i from a participant, the survey manager M_i decodes the message Msg_1^i with κ^i, and gets $(MD_U^i|SID_U|Cert_U^i)$. Then it takes the same process as in SPS to verify each survey response and maintain a set S^i, which contains the SID of all already-responded participants to detect potential double submission.

When the survey period ends, each survey manager M_i starts to aggregate received responses from all m participants to generate the ith masked group survey data GD^i,

$$GD^i = \sum_{j=1}^{m} MD_{U_j}^i = \sum_{j=1}^{m} A_i X_{U_j}^i B_i, \tag{13}$$

and then sends GD^i with the set S^i in a message Msg_2^i to the client V:

$$\boldsymbol{M_i} \xrightarrow{Msg_2^i = (GD^i \,|\, S^i)} \boldsymbol{V}$$

3) Data Tallying Phase: After receiving the message Msg_2^i from every distributed survey managers M_i, the client uses $\{A_i^{-1}\}_{i\in[1,s]}$ and $\{B_i^{-1}\}_{i\in[1,s]}$ to obtain the de-masked group survey data G^i for each survey manager M_i:

$$G^i = A_i^{-1} \times \sum_{j=1}^{m} A_i X_{U_j}^i B_i \times B_i^{-1} = \sum_{j=1}^{m} X_{U_j}^i. \tag{14}$$

The final tally is simply

$$G = \sum_{i=1}^{s} G^i = \sum_{j=1}^{m} \sum_{i=1}^{s} X_{U_j}^i = \sum_{j=1}^{m} X_{U_j} = R, \tag{15}$$

where $R = \sum_{j=1}^{m} X_{U_j}$ is the sum of the original data matrices of all valid participants. Finally, the client verifies if the number of participants matches the number of elements in each set S^i. If they all match, the client gets the final tally for every question choice.

C. Property Analysis

RSPS preserves the nice properties of SPS. First, the correctness of RSPS is obvious from (15), similar to the analysis of SPS in Section IV-B. In addition, the participant and the client work almost the same as in SPS. Our two schemes diverge from the data collecting phase, where RSPS utilizes multiple survey managers to each aggregate a share of masked data, but the total work for all survey managers does not change. Similar to the analysis of SPS in Section IV-C, IV-D and IV-E, our RSPS scheme also achieves the universal verifiability and robustness with comparable efficiency as SPS. We will omit the duplicate discussion.

D. Privacy Analysis

In this subsection, we show that RSPS achieves strong privacy as SPS, and demonstrate that it is also robust against the joint collusion attack.

1) Strong privacy: In SPS, all masked data can be monitored if the adversary compromises the survey manager. However, in RSPS, the adversary cannot obtain such information until it compromises all s survey managers. Clearly, RSPS also achieves comparable strong privacy as SPS, if not stronger.

2) Robustness against joint collusion attack: We now demonstrate that RSPS is robust against the joint collusion attack: The adversary cannot derive any original response of any individual participant unless it compromises the client and all s survey managers.

Lemma V.1. *The additive secret sharing scheme by Algorithm 1 splits the secret X to s shares $X^1, \cdots, X^s$. Any subset of $s-1$ shares of X is uniformly distributed.*

Proof: According to the secret sharing in Algorithm 1, the shares $X^1, \cdots, X^{s-1}$ are uniformly distributed and independent in $\mathbb{R}^{n\times p}$. Let $X^{'} = X^2 + \cdots + X^{s-1}$ for a fixed $X^2, \cdots, X^{s-1}$. From Algorithm 1, we have

$$\begin{aligned} X^s &= X - X^1 - (X^2 + \cdots + X^{s-1}) \\ &= (X - X^{'}) - X^1. \end{aligned} \tag{16}$$

As X^1 is still uniformly distributed when $X^2, \cdots, X^{s-1}$ are fixed, we get that X^s is uniformly distributed and independent from $X^2, \cdots, X^{s-1}$. Therefore, $X^2, \cdots, X^s$ are also uniformly distributed and independent in $\mathbb{R}^{n\times p}$. The same argument can be extended for any other $s-1$ different shares and the proof is completed. ■

Theorem V.2. *In the joint collusion attack, the adversary cannot learn anything about the original survey data X for any coalition of up to $s-1$ survey managers.*

Proof: If $s-1$ survey managers become compromised and cooperate with the compromised client to get the DM^2 masking keys, the adversary can reveal at most $s-1$ shares $\{X^{i_1}, \cdots, X^{i_{s-1}}\}$ of the participants' response. According to Lemma V.1, any $s-1$ element subset $\{X^{i_1}, \cdots, X^{i_{s-1}}\}$ is uniformly distributed. Therefore, for any two secret values in $\mathbb{R}^{n\times p}$, their secret shared forms are indistinguishable for any coalition of up to $s-1$ survey managers. The coalition cannot get the remaining share to recover the original survey data X. This completes the proof. ■

According to Theorem V.2, the adversary cannot learn anything about the original survey data X unless all s survey managers are compromised to launch the joint collusion attack together with a compromised client, which demonstrate RSPS is more robust than SPS.

VI. Related Work

In this section, we review some existing privacy protection techniques, which can be adopted in online survey to address the privacy issues to some degree. The related research efforts can be briefly categorized into the following groups.

Secure Multi-party Computation (SMC): SMC [19] [20] allows multiple data sources to jointly compute a function over their input without revealing their original data to each other. The computation and communication overhead of SMC is prohibitively high when the number of participants is large. Moreover, the data sources have to directly involve in any joint computation, and stand by ready for any data analysis that may happen for a long time, which is not applicable to online survey due to lack of submit-and-go property.

Privacy-preserving Data Mining (PPDM): PPDM techniques [21]–[23] target at extracting statistical results from perturbed data without compromising the privacy of the data sources. They can be applied to online survey, but with great limitations. For example, the "random response" technique proposed by Warner [21] has not been widely used in practice because it is only applicable to binary data. Other perturbation-based approaches [22] [23] add noise directly to the raw data before collecting. But these perturbation techniques reduce the precision of the final result, and the data collector can use privacy intrusion techniques [24] [25] to filter noise from the perturbed data, thereby rediscovering part of the original private data. Therefore, they are not adequate in providing strong privacy or exact final result for online survey.

De-identification and Anonymous Data Collection: Traditional approaches of identity removal cannot achieve strong privacy because even after standard participant identifiers are removed, it is still sometimes possible to deduce the participant identities from the remaining data. Anonymous data collection methods are designed with the intention to collect data anonymously without revealing the participants' identities, including cryptographic solutions [12] [13] and anonymous communications [14] [15]. With a goal of unlinkability, these methods try to prevent data collector from learning which input came from which participant. But they do not hide the data values, and linkage attack [10] [11] can still occur in many situations. So these methods still cannot achieve strong privacy.

In summary, the past research focuses on secure multi-party computation, privacy-preserving data mining, de-identification and anonymous data collection, with success in various degrees. However, they are inadequate in providing the strong privacy or exact survey results, which are essential to online survey. We point out that as long as the raw survey data are collected, leaking of private information is always a possibility. Our paper takes a bold step to avoid this problem by developing new solutions based on data masking. Our solutions ensure that the original survey responses never leave the participants and the survey manager only collects masked responses, which retain the tallying utility of the original survey data and also achieve strong privacy for participants.

VII. Conclusion

In this paper, we propose two novel efficient strong privacy online survey schemes SPS and RSPS. Unlike many existing privacy protection techniques in other scenarios, our schemes apply a novel efficient dual matrix masking (DM^2)

technique, and achieve accurate outcomes and strong privacy. In particular, RSPS is more robust against the joint collusion attack. Through extensive analysis on correctness, efficiency, universal verifiability and robustness, we demonstrate that our schemes can be efficiently applied to online survey in a variety of situations with accuracy and strong privacy.

VIII. Acknowledgment

This work is supported in part by the National Science Foundation under grant STC-1562485, and a grant from Florida Center for Cybersecurity.

References

[1] ESOMAR, "GLOBAL MARKET RESEARCH 2014," 2014. [Online]. Available: http://www.mrsnz.org.nz

[2] J. Bethlehem and S. Biffignandi, "Web Surveys and Other Modes of Data Collection," *Handbook of Web Surveys*, pp. 147–188, 2011.

[3] V. M. Sue and L. A. Ritter, *Conducting Online Surveys*. Sage, 2012.

[4] "Top 5 Benefits of Online Surveys," 2014. [Online]. Available: http://obsurvey.com/blog/top-5-benefits-of-online-surveys-pt1

[5] V. Vehovar and K. L. Manfreda, "Overview: Online Surveys," *The SAGE handbook of online research methods*, pp. 177–194, 2008.

[6] T. D. Baker, "Confidentiality and Electronic Surveys: How IRBs Address Ethical and Technical Issues," *IRB*, vol. 34, no. 5, p. 8, 2012.

[7] F. John, "Privacy and Market Research, At A Glance," 2012. [Online]. Available: https://www.surveymonkey.com/blog/2012/04/19/privacy-and-market-research

[8] D. Schultz and S. Schultz, *Theories Of Personality*. Cengage Learning, 2012.

[9] M. Orcutt, "Hackers Are Homing In on Hospitals," 2014. [Online]. Available: http://www.technologyreview.com/news/530411/hackers-are-homing-in-on-hospitals

[10] I. Dinur and K. Nissim, "Revealing Information while Preserving Privacy," *Proc. of ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 202–210, 2003.

[11] C. Dwork and M. Naor, "On the Difficulties of Disclosure Prevention in Statistical Databases or The Case for Differential Privacy," *Journal of Privacy and Confidentiality*, vol. 2, no. 1, p. 8, 2008.

[12] Z. Yang, S. Zhong, and R. N. Wright, "Anonymity-Preserving Data Collection," *Proc. of ACM SIGKDD*, pp. 334–343, 2005.

[13] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," *ACM Computing Surveys (CSUR)*, vol. 42, no. 4, p. 14, 2010.

[14] B. Adida, "Helios: Web-based Open-Audit Voting," *Proc. of USENIX Security Symposium*, vol. 17, pp. 335–348, 2008.

[15] S. Chow, J. Liu, and D. Wong, "Robust Receipt-Free Election System with Ballot Secrecy and Verifiability," *Proc. of NDSS*, vol. 8, pp. 81–94, 2008.

[16] G. T. Duncan and R. W. Pearson, "Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future," *Statistical Science*, vol. 6, no. 3, pp. 219–232, 1991.

[17] D. Ting, S. E. Fienberg, and M. Trottini, "Random Orthogonal Matrix Masking Methodology for Microdata Release," *International Journal of Information and Computer Security*, vol. 2, no. 1, pp. 86–105, 2008.

[18] Y. Zhou, Y. Zhou, S. Chen, and S. S. Wu, "MVP: An Efficient Anonymous E-voting Protocol," *Proc. of IEEE GLOBECOM*, 2016.

[19] S. E. Fienberg, Y. Nardi, and A. B. Slavković, "Valid Statistical Analysis for Logistic Regression with Multiple Sources," *Protecting Persons While Protecting the People*, pp. 82–94, 2009.

[20] I. Damgård, V. Pastro, N. Smart, and S. Zakarias, "Multiparty Computation from Somewhat Homomorphic Encryption," *Advances in Cryptology–CRYPTO 2012*, pp. 643–662, 2012.

[21] S. L. Warner, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[22] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in Privacy Preserving Data Mining," *Proc. of ACM Sigmod Record*, vol. 33, no. 1, pp. 50–57, 2004.

[23] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, "Toward Privacy in Public Databases," *Theory of Cryptography*, pp. 363–385, 2005.

[24] H. Kargupta, S. Datta, W. Qi, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," *Proc. of ICDM*, pp. 99–106, 2003.

[25] Z. Huang, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," *Proc. of ACM SIGMOD Int'l Conf. on Management of Data*, pp. 37–48, 2005.