MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

A Framework for Privacy Preserving Statistical Analysis on Distributed Databases

Lin, B-R; Wang, Y.; Rane, S.

TR2012-064 December 2012

Abstract

Alice and Bob are mutually untrusting curators who possess separate databases containing information about a set of respondents. This data is to be sanitized and published to enable accurate statistical analysis, while retaining the privacy of the individual respondents in the databases. Further, an adversary who looks at the published data must not even be able to compute statistical measures on it. Only an authorized researcher should be able to compute marginal and joint statistics. This work is an attempt toward providing a theoretical formulation of privacy and utility for problems of this type. Privacy of the individual respondents is formulated using differential privacy. Privacy of the marginal and joint statistics on the distributed databases is formulated using a new model called distributional differential privacy. Finally, a constructive scheme based on randomized response is presented as an example mechanism that satisfies the formulated privacy requirements.

IEEE International Workshop on Information Forensics and Security (WIFS)

A Framework for Privacy Preserving Statistical Analysis on Distributed Databases

Bing-Rong Lin
Pennsylvania State University
blin@cse.psu.edu

Ye Wang and Shantanu Rane
Mitsubishi Electric Research Laboratories
{yewang,rane}@merl.com

***Abstract*—Alice and Bob are mutually untrusting curators who possess separate databases containing information about a set of respondents. This data is to be sanitized and published to enable accurate statistical analysis, while retaining the privacy of the individual respondents in the databases. Further, an adversary who looks at the published data must not even be able to compute statistical measures on it. Only an authorized researcher should be able to compute marginal and joint statistics. This work is an attempt toward providing a theoretical formulation of privacy and utility for problems of this type. Privacy of the individual respondents is formulated using $\epsilon-$differential privacy. Privacy of the marginal and joint statistics on the distributed databases is formulated using a new model called $\delta-$distributional $\epsilon-$differential privacy. Finally, a constructive scheme based on randomized response is presented as an example mechanism that satisfies the formulated privacy requirements.**

I. Introduction

With the rapid emergence and penetration of "Big Data" into everyday human lives, it has become extremely important for public and private enterprises to perform statistical analysis on large databases. These enterprises include governments, medical universities, hospitals, financial institutions, and private companies. For example, a medical researcher may be able to determine the efficacy of a drug in the treatment of a disease or find correlations between the occurrence of disease and the food habits, ages, or geographical locations of the patients. This would clearly be beneficial in improving the efficacy of healthcare provided to the patients. In the corporate world, companies may be able to leverage statistical studies on customer data for targeted advertising and product placement.

In all such applications, it is imperative that the privacy of individuals be maintained. Indeed, unless the public is satisfied that their privacy is being preserved, they would not allow their data to be collected or used. The mechanism of randomized response [1], [2] was developed to address this problem, and was originally used for collecting data from surveys involving uncomfortable questions. In this mechanism, the individual respondents are allowed to change their responses with a certain probability. The result is that the data that gets recorded or published does not unambiguously reveal the response of a particular respondent, but aggregate statistical measures, such as the mean or variance, can still be computed from the "perturbed" data. Thus, randomized response provides privacy to the respondents but does not hide the probability distribution or "type" of the data from an adversary.

In recent years, the notion of differential privacy has received much attention [3], [4]. Differential privacy (DP) is a strict privacy formulation applied to functions computed from the respondent data, such as classifiers. Informally, differential privacy means that the result of a function computed on a database of respondents is almost insensitive to the presence or absence of a particular respondent. A more formal way of stating this is that when the function is evaluated on adjacent databases (differing in only one respondent), the probability of outputting the same result is almost unchanged. It has been shown that adding Laplacian noise to the result of a function computed on a database provides differential privacy to the individual respondents in the database [5] [6]. The variance of the added noise trades off the privacy of the respondents with the utility of the function computed on the database.

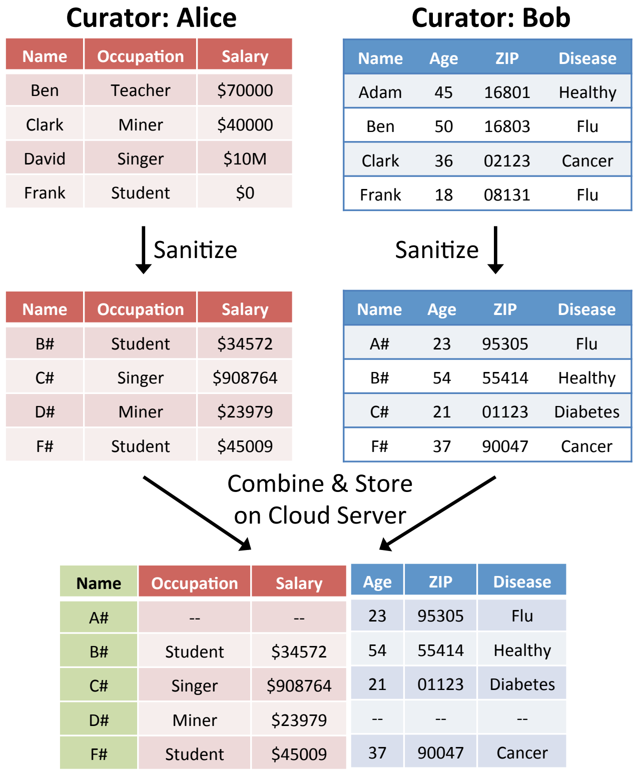
Conventional mechanisms for privacy, such as k-anonymization [7] [8] are not differentially private, because an adversary can link an arbitrary amount of side information to the anonymized database, and defeat the anonymization mechanism [9]. Mechanisms used to provide differential privacy typically involve *output* perturbation, e.g., noise is added to a function of the data. Nevertheless, it can be shown that the randomized response mechanism – where noise is added to the data itself – provides differential privacy to the respondents. Unfortunately, while differential privacy provides a rigorous and worst-case characterization for the privacy of the respondents, it is not enough to formulate privacy of the empirical probability distribution or "type" of the data. In particular, if an adversary has accessed anonymized adjacent databases, a differentially private mechanism ensures that he cannot de-anonymize any respondent, but by construction, possessing an anonymized database reveals the distribution of the data.

In this work, we are interested in a privacy formulation that protects the privacy of the database respondents while also protecting the empirical probability distribution from unauthorized parties. To do this, we have to first answer the question: "What does it mean for a probability distribution or type to be private?" To this end, first we describe the underlying multiparty privacy framework consisting of several database curators and fix our notation in Section II. Notions of respondent privacy and distribution privacy are made explicit in Section III along with a discussion on the utility of privacy mechanisms for this problem. Section IV is devoted to deriving the conditions that privacy mechanisms must satisfy in order to simultaneously achieve the twin goals of respondent privacy and distributional privacy. In Section V, we analyze a mechanism in which each curator independently sanitizes its database to protect the privacy of the database respondents, while allowing marginal and joint statistics to be computed by an authorized party who is given certain randomization parameters. Section VI concludes the paper with a discussion of the results and contributions.

II. Multiparty Problem Setting and Notation

For ease of exposition, we present our problem formulation and results with two data curators, Alice and Bob, however our methods can easily be generalized to more than two curators.

WIFS'2012, December, 2-5, 2012, Tenerife, Spain. 978-1-4244-9080-6/10/\$26.00 ©2012 IEEE.



Name	Occupation	Salary
Ben	Teacher	$70000
Clark	Miner	$40000
David	Singer	$10M
Frank	Student	$0

Name	Age	ZIP	Disease
Adam	45	16801	Healthy
Ben	50	16803	Flu
Clark	36	02123	Cancer
Frank	18	08131	Flu

Name	Occupation	Salary
B#	Student	$34572
C#	Singer	$908764
D#	Miner	$23979
F#	Student	$45009

Name	Age	ZIP	Disease
A#	23	95305	Flu
B#	54	55414	Healthy
C#	21	01123	Diabetes
F#	37	90047	Cancer

Name	Occupation	Salary	Age	ZIP	Disease
A#	--	--	23	95305	Flu
B#	Student	$34572	54	55414	Healthy
C#	Singer	$908764	21	01123	Diabetes
D#	Miner	$23979	--	--	--
F#	Student	$45009	37	90047	Cancer

Fig. 1. An example in which two curators Alice and Bob independently sanitize their databases to protect the privacy of their respondents, and make the combined data available on a cloud server for statistical research. The key of join operation is the hashed name generated from a cryptographic hash function.

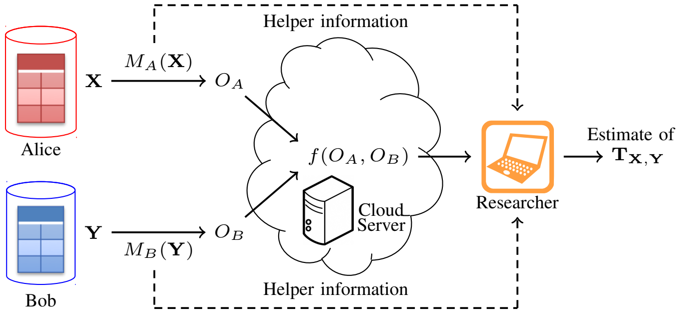


Fig. 2. Curators Alice and Bob independently sanitize their databases and provide it to a cloud server. A researcher can derive joint statistics or joint type based on the sanitized data, without compromising the privacy of individual database respondents. Neither the statistics nor the individual data entries are revealed to the cloud server.

Consider a medical data mining application in which Alice and Bob are mutually untrusting curators who have collected data from a large number of respondents. For example, as shown in Fig. 1, Alice's database contains the occupation and salary of the respondents, and Bob's database contains the age and location of the respondents along with the disease that they are suffering from, if any. If this data is combined and analyzed, medical researchers may be able to determine for example, the correlations among age, occupation and cancer. However, before the data is combined and made available to a medical researcher, it must be properly sanitized to preserve the privacy of the respondents; it is well-known that simply removing the respondent names to anonymize the data may not do enough to preserve the privacy of the respondents [10].

To facilitate the storage, transmission and computation required on these potentially large databases, the curators submit their sanitized data to an untrusted cloud server, as shown in Fig. 2. The server thus holds a large vertically partitioned database, in which some columns are contributed by Alice and other columns are contributed by Bob. The objective of the authorized researcher is to the extract useful statistics about the underlying database from computation provided on the sanitized data by the curator and low-rate helper information from the curators. Altogether, we have the following privacy and utility requirements:

1) **Individual Privacy**: The individual data of the respondents should not be revealed to the cloud server or the researchers.
2) **Distributional Privacy**: The statistics of the data provided by Alice and Bob should not be revealed to the cloud server.
3) **Statistical Utility**: The *authorized* researcher should be able to compute the joint and marginal distributions of the data provided by Alice and Bob.

In the following sections, we formalize these notions in our problem framework.

A. Problem Framework and Notation

The data table held by Alice is modeled as a sequence of random variables $\mathbf{X} := (X_1, X_2, \ldots, X_n)$, with each X_i taking values in the finite-alphabet $\mathcal{X}$. Likewise, Bob's data table is modeled as a sequence of random variables $\mathbf{Y} := (Y_1, Y_2, \ldots, Y_n)$, with each Y_i taking values in the finite-alphabet $\mathcal{Y}$. The length of the sequences, n, represents the total number of respondents in the database, and each (X_i, Y_i) pair represents the data of the respondent i collectively held by Alice and Bob, with the alphabet $\mathcal{X} \times \mathcal{Y}$ representing the domain of each respondents's data.

We assume that the sequence of data pairs (X_i, Y_i) are independently and identically distributed (i.i.d.) according to some joint distribution $P_{X,Y}$ over $\mathcal{X} \times \mathcal{Y}$, that is, for $\mathbf{x} := (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $\mathbf{y} := (y_1, \ldots, y_n) \in \mathcal{Y}^n$, we have that

$$P_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} P_{X,Y}(x_i, y_i).$$

A privacy mechanism is a randomized mapping $M : \mathcal{I} \to \mathcal{O}$ from some finite input space $\mathcal{I}$ to a finite output space $\mathcal{O}$, and governed by a conditional distribution $P_{O|I}$. *Post-randomization* (PRAM) is a specific class of privacy mechanisms where the input and output are both sequences, i.e., $\mathcal{I} = \mathcal{O} = \mathcal{D}^n$ for some finite alphabet $\mathcal{D}$, and each element of the input sequence is randomized independently and identically according to an element-wise conditional distribution.

In this paper, we consider that Alice and Bob each independently apply a PRAM mechanism to their data tables. Denoting these mechanisms as $R_A : \mathcal{X}^n \to \mathcal{X}^n$ and $R_B : \mathcal{Y}^n \to \mathcal{Y}^n$, their respective outputs as $\widetilde{\mathbf{X}} := (\widetilde{X}_1, \ldots, \widetilde{X}_n) := R_A(\mathbf{X})$ and $\widetilde{\mathbf{Y}} := (\widetilde{Y}_1, \ldots, \widetilde{Y}_n) := R_B(\mathbf{Y})$, and their governing distributions as $P_{\widetilde{X}|X}$ and $P_{\widetilde{Y}|Y}$, we have that

$$\begin{aligned} P_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}|\mathbf{X},\mathbf{Y}}(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}|\mathbf{x},\mathbf{y}) &= P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\mathbf{x})P_{\widetilde{\mathbf{Y}}|\mathbf{Y}}(\widetilde{\mathbf{y}}|\mathbf{y}) \\ &= \prod_{i=1}^{n} P_{\widetilde{X}|X}(\widetilde{x}_i|x_i)P_{\widetilde{Y}|Y}(\widetilde{y}_i|y_i). \end{aligned}$$

We also use $R_{AB} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{X}^n \times \mathcal{Y}^n$, defined by $R_{AB}(\mathbf{X},\mathbf{Y}) := (\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}) := (R_A(\mathbf{X}), R_B(\mathbf{Y}))$ to denote the mechanism that arises from the concatenation of each individual mechanism. R_{AB} is also a PRAM mechanism and is governed by the conditional distribution $P_{\widetilde{X}|X}P_{\widetilde{Y}|Y}$.

B. Type Notation

The *type* (or empirical distribution) of a sequence of random variables $\mathbf{X} = (X_1, \ldots, X_n)$ is the mapping $T_{\mathbf{X}} : \mathcal{X} \to [0,1]$ defined by

$$T_{\mathbf{X}}(x) := \frac{|\{i : X_i = x\}|}{n}, \quad \forall x \in \mathcal{X}.$$

The *joint type* of two sequences $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is the mapping $T_{\mathbf{X},\mathbf{Y}} : \mathcal{X} \times \mathcal{Y} \to [0,1]$ defined by

$$T_{\mathbf{X},\mathbf{Y}}(x,y) := \frac{|\{i : (X_i, Y_i) = (x,y)\}|}{n}, \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y}.$$

The *conditional type* of a sequence $\mathbf{Y} = (Y_1, \ldots, Y_n)$ given another $\mathbf{X} = (X_1, \ldots, X_n)$ is the mapping $T_{\mathbf{Y}|\mathbf{X}} : \mathcal{Y} \times \mathcal{X} \to [0,1]$ defined by

$$T_{\mathbf{Y}|\mathbf{X}}(y|x) := \frac{T_{\mathbf{Y},\mathbf{X}}(y,x)}{T_{\mathbf{X}}(x)} = \frac{|\{i : (Y_i, X_i) = (y,x)\}|}{|\{i : X_i = x\}|}$$

Note that the values of these type mappings are determined given the underlying sequences, and are potentially random if the sequences are random.

C. Matrix Notation for Distributions and Types

The various (marginal, conditional and joint) distributions and types of finite-alphabet random variables can be represented as vectors and/or matrices. By fixing a consistent ordering on their finite domains, these mappings can be written vectors/matrices indexed by their domains. The distribution $P_X : \mathcal{X} \to [0,1]$ can be written as an $|\mathcal{X}| \times 1$ column-vector $\mathbf{P}_X$, whose "x"-th element, for $x \in \mathcal{X}$, is given by $\mathbf{P}_X[x] := P_X(x)$. A conditional distribution $P_{Y|X} : \mathcal{Y} \times \mathcal{X} \to [0,1]$ can be written as a $|\mathcal{Y}| \times |\mathcal{X}|$ matrix $\mathbf{P}_{Y|X}$, defined by $\mathbf{P}_{Y|X}[y,x] := P_{Y|X}(y|x)$. A joint distribution $P_{X,Y} : \mathcal{X} \times \mathcal{Y} \to [0,1]$ can be written as a $|\mathcal{X}| \times |\mathcal{Y}|$ matrix $\mathbf{P}_{X,Y}$, defined by $\mathbf{P}_{X,Y}[x,y] := P_{X,Y}(x,y)$, or as a $|\mathcal{X}||\mathcal{Y}| \times 1$ column-vector $\overline{\mathbf{P}}_{X,Y}$, formed by stacking the columns of $\mathbf{P}_{X,Y}$.

We can similarly develop the matrix notation for types, with $\mathbf{T}_{\mathbf{X}}$, $\mathbf{T}_{\mathbf{Y}|\mathbf{X}}$, $\mathbf{T}_{\mathbf{X},\mathbf{Y}}$, and $\overline{\mathbf{T}}_{\mathbf{X},\mathbf{Y}}$ similarly defined for sequences $\mathbf{X}$ and $\mathbf{Y}$ with respect to the corresponding type mappings. Note that these type vectors/matrices are random quantities.

III. Privacy and Utility Conditions

We now formulate the privacy and utility requirements for this problem of computing joint and marginal statistics on independently sanitized data. According to the requirements described in Section II, the formulation will consider in turn, privacy of the respondents, privacy of the distribution and finally the utility for an authorized medical researcher.

A. Privacy of the Respondents

The data pertaining to a respondent should be kept private from all other parties, including any authorized researchers who aim to recover the distributions. We formalize this notion using ϵ-differential privacy for the respondents.

Definition III.1. *[11] For $\epsilon \geq 0$, a randomized mechanism $M : \mathcal{D}^n \to \mathcal{O}$ gives ϵ-differential privacy if for all data sets $\mathbf{d}, \mathbf{d}' \in \mathcal{D}^n$, within Hamming distance $d_H(\mathbf{d}, \mathbf{d}') \leq 1$, and all $\mathcal{S} \subseteq \mathcal{O}$,*

$$\Pr[M(\mathbf{d}) \in \mathcal{S}] \leq e^{\epsilon} \Pr[M(\mathbf{d}') \in \mathcal{S}]$$

Under the assumption that the respondents are sampled i.i.d., a privacy mechanism that satisfies DP results in a strong privacy guarantee: an attacker with knowledge of all respondents except one, cannot discover the data of the sole missing respondent [12]. This notion of privacy is rigorous and widely accepted and satisfies the privacy axioms of [13], [14].

B. Privacy of the Distribution

The data curators, Alice and Bob, do not want to reveal the marginal and joint statistics of the data to attackers or to the cloud server. Hence they must ensure that the marginal and joint empirical distribution, i.e., the marginal and joint types cannot be learned from $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$. As explained in the introduction, ϵ-DP cannot be used to characterize privacy in this case. To formulate a privacy notion for the empirical probability distribution, we extend ϵ-differential privacy as follows.[1]

Definition III.2. *(δ-Distributional ϵ-differential privacy) Let $d(\cdot,\cdot)$ be a distance metric on the space of distributions. For $\epsilon, \delta \geq 0$, a randomized mechanism $M : \mathcal{D}^n \to \mathcal{O}$ gives δ-distributional ϵ-differential privacy if for all data sets $\mathbf{d}, \mathbf{d}' \in \mathcal{D}^n$, with $d(\mathbf{T_d}, \mathbf{T_{d'}}) \leq \delta$, and all $\mathcal{S} \subseteq \mathcal{O}$,*

$$\Pr[M(\mathbf{d}) \in \mathcal{S}] \leq e^{\epsilon} \Pr[M(\mathbf{d}') \in \mathcal{S}]$$

Note that the larger the δ is, and the smaller the ϵ is, the better the distribution is protected. We also want to point out that Definition III.2 satisfies privacy axioms [13], [14].

C. Utility for Authorized Researchers

The objective of the authorized researcher is to the extract useful statistics about the underlying database $(\mathbf{X}, \mathbf{Y})$. We model this problem as the reconstruction of the joint and marginal type functions $T_{\mathbf{X},\mathbf{Y}}(x,y)$, $T_{\mathbf{X}}(x)$, and $T_{\mathbf{Y}}(y)$, or (equivalently) the matrices $\mathbf{T}_{\mathbf{X},\mathbf{Y}}$, $\mathbf{T}_{\mathbf{X}}$, and $\mathbf{T}_{\mathbf{Y}}$. The cloud server facilitates this reconstruction by providing computation based on the sanitized data $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$. The data curators, Alice and Bob, also assist by providing some low-rate, independently generated helper information. Given the cloud server's computation and the helper information from the curators, the researcher produces the estimates $\dot{\mathbf{T}}_{\mathbf{X},\mathbf{Y}}$, $\dot{\mathbf{T}}_{\mathbf{X}}$, and $\dot{\mathbf{T}}_{\mathbf{Y}}$.

For a given distance metric $d(\cdot,\cdot)$ over the space of distributions, we define the *expected utility* of the estimates as

$$\mu_{\mathbf{T}_{\mathbf{X},\mathbf{Y}}} := E[-d(\dot{\mathbf{T}}_{\mathbf{X},\mathbf{Y}}, \mathbf{T}_{\mathbf{X},\mathbf{Y}})]$$

$$\mu_{\mathbf{T}_{\mathbf{X}}} := E[-d(\dot{\mathbf{T}}_{\mathbf{X}}, \mathbf{T}_{\mathbf{X}})]$$

$$\mu_{\mathbf{T}_{\mathbf{Y}}} := E[-d(\dot{\mathbf{T}}_{\mathbf{Y}}, \mathbf{T}_{\mathbf{Y}})]$$

[1] A similar definition appeared in [15]. However, the conditions in our definition specializes to distributional distance metrics, and does not require additional technical conditions.

IV. Analysis of Privacy Requirements

In this section, we analyze our PRAM mechanisms with respect to the privacy requirements.

A natural question to ask is whether privacy protection of the marginal types of the database implies privacy protection for the joint type. We now show that if the distance function d satisfies a general property shared by common distribution distance measures, then this is indeed the case.

Lemma IV.1. *Let $d(\cdot,\cdot)$ be a distance function such that*

$$d(\mathbf{T}_{\mathbf{X},\mathbf{Y}},\mathbf{T}_{\mathbf{X}',\mathbf{Y}'}) \geq \max(d(\mathbf{T}_{\mathbf{X}},\mathbf{T}_{\mathbf{X}'}), d(\mathbf{T}_{\mathbf{Y}},\mathbf{T}_{\mathbf{Y}'})). \qquad (1)$$

Let M_{AB} be the privacy mechanism defined by $M_{AB}(\mathbf{X},\mathbf{Y}) := (M_A(\mathbf{X}), M_B(\mathbf{Y}))$. If M_A satisfies δ-distributional ϵ_1-differential privacy and M_B satisfies δ-distributional ϵ_2-differential privacy, then M_{AB} satisfies δ-distributional $(\epsilon_1+\epsilon_2)$-differential privacy.

Proof: For any databases $(\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}') \in \mathcal{X}^n \times \mathcal{Y}^n$ and $d(\mathbf{T}_{\mathbf{x},\mathbf{y}},\mathbf{T}_{\mathbf{x}',\mathbf{y}'}) \leq \delta$, Equation 1 implies $d(\mathbf{T}_{\mathbf{x}},\mathbf{T}_{\mathbf{x}}) \leq \delta$ and $d(\mathbf{T}_{\mathbf{y}},\mathbf{T}_{\mathbf{y}'}) \leq \delta$ Then, for any $\mathcal{S} \subseteq \mathcal{X}^n \times \mathcal{Y}^n$, the result follows from

$$\begin{aligned}\sum_{(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})\in\mathcal{S}} P_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}|\mathbf{X},\mathbf{Y}}(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}|\mathbf{x},\mathbf{y}) &= \sum_{(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})\in\mathcal{S}} P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\mathbf{x})P_{\widetilde{\mathbf{Y}}|\mathbf{Y}}(\widetilde{\mathbf{y}}|\mathbf{y})\\ &\leq \sum_{(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})\in\mathcal{S}} e^{\epsilon_1}P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\mathbf{x}')e^{\epsilon_2}P_{\widetilde{\mathbf{Y}}|\mathbf{Y}}(\widetilde{\mathbf{y}}|\mathbf{y}')\\ &= e^{\epsilon_1+\epsilon_2}\sum_{(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})\in\mathcal{S}} P_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}|\mathbf{X},\mathbf{Y}}(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}|\mathbf{x}',\mathbf{y}').\end{aligned}$$

■

The condition (1) is fairly general; it can be shown that the ℓ_1 distance and KL divergence between distributions (or types) satisfies this property.

We argue that if vertically partitioned tables are sanitized independently and we want to recover joint distribution from the sanitized table, the choice of privacy mechanisms are restricted to the class of PRAM algorithms. We now analyze the constraints that should be placed on PRAM algorithms so that they satisfy the privacy constraints of Section III. First, consider the privacy requirement of the respondents in Alice and Bob's databases.

Lemma IV.2. *Let $R : \mathcal{X}^n \to \mathcal{X}^n$ be a PRAM mechanism governed by conditional distribution $P_{\widetilde{X}|X}$. R satisfies ϵ-DP if*

$$\epsilon = \max_{x_1,x_2,\widetilde{x}\in\mathcal{X}} \ln(P_{\widetilde{X}|X}(\widetilde{x}|x_1)) - \ln(P_{\widetilde{X}|X}(\widetilde{x}|x_2)). \qquad (2)$$

Proof: Let $\mathcal{S} \subseteq \mathcal{X}^n$ and let $\mathbf{x},\mathbf{x}' \in \mathcal{X}^n$ and $d_H(\mathbf{x},\mathbf{x}') = 1$. Then, we have

$$\begin{aligned}P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}\in\mathcal{S}|\mathbf{x}) &= \sum_{\widetilde{\mathbf{x}}\in\mathcal{S}}\prod_{i=1}^{n} P_{\widetilde{X}|X}(\widetilde{x}_i|x_i)\\ \leq \quad &\sum_{\widetilde{\mathbf{x}}\in\mathcal{S}} e^{\epsilon}\prod_{i=1}^{n} P_{\widetilde{X}|X}(\widetilde{x}_i|x_i') = e^{\epsilon}P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}\in\mathcal{S}|\mathbf{x}')\end{aligned}$$

The inequality holds because $d_H(\mathbf{x},\mathbf{x}') = 1$ and Equation 2. The result follows from the definition of ϵ-DP. ■

Lemma IV.3. *Define $M_{AB}(\mathbf{x},\mathbf{y}) = (M_A(\mathbf{x}), M_B(\mathbf{y}))$. If M_A satisfies ϵ_1-DP and M_B satisfies ϵ_2-DP, then M_{AB} satisfies $(\epsilon_1+\epsilon_2)$-DP.*

Proof: The result follows from composition lemma [16]. ■

The lemma can be extended to k curators where if i^{th} curator's sanitized data satisfies ϵ_i-DP, then the joint system provides $(\sum_{i=1}^{k}\epsilon_i)$-DP. Next, we consider the privacy requirement for the joint and marginal types.

Lemma IV.4. *Let $d(\cdot,\cdot)$ be the distance metric on the space of distributions. Let $R : \mathcal{X}^n \to \mathcal{X}^n$ be a PRAM mechanism governed by conditional distribution $P_{\widetilde{X}|X}$.*

- ***Necessary Condition:*** *If R satisfies δ-distributional ϵ-DP, then R must satisfy $\frac{\epsilon}{\lfloor n/2\rfloor}$-DP for the respondents.*
- ***Sufficient Condition:*** *If R satisfies $\frac{\epsilon}{n}$-DP for the respondents, then R satisfies δ-distributional ϵ-DP.*

Proof: We first prove the necessary condition. Assume that n is even. By way of contradiction, let $\widetilde{x}_1, x_i, x_j \in \mathcal{X}$ and $\frac{P_{\widetilde{X}|X}(\widetilde{x}_1|x_i)}{P_{\widetilde{X}|X}(\widetilde{x}_1|x_j)} > e^{\frac{\epsilon}{n/2}}$. Let $\mathbf{x} \in \mathcal{X}^n$ and the first $n/2$ respondents have value x_i and the last $n/2$ respondents have value x_j. Let $\mathbf{x}' \in \mathcal{X}^n$ and the first $n/2$ respondents have value x_j and the last $n/2$ respondents have value x_i. Pick $\widetilde{x}_2$ such that $P_{\widetilde{X}|X}(\widetilde{x}_2|x_j) \geq P_{\widetilde{X}|X}(\widetilde{x}_2|x_i)$. The existence of $\widetilde{x}_2$ can be proved by plugging $\frac{P_{\widetilde{X}|X}(\widetilde{x}_1|x_i)}{P_{\widetilde{X}|X}(\widetilde{x}_1|x_j)} > e^{\frac{\epsilon}{n/2}}$ into $\sum_{\widetilde{x}\in\mathcal{X}} P_{\widetilde{X}|X}(\widetilde{x}|x_i) = \sum_{\widetilde{x}\in\mathcal{X}} P_{\widetilde{X}|X}(\widetilde{x}|x_j)$. Let $\widetilde{\mathbf{x}} \in \mathcal{X}^n$ and the first $n/2$ respondents have value $\widetilde{x}_1$ and the last $n/2$ respondents have value $\widetilde{x}_2$. Then, the result follows from contradiction since

$$\frac{P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\mathbf{x})}{P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\mathbf{x}')} > \prod_{i=1}^{n/2} e^{\frac{\epsilon}{n/2}} = e^{\epsilon}.$$

The case of odd n can be proved by carefully picking the last respondent's data as x_1, x_2 and $\widetilde{x}$ in $\mathbf{x}$, $\mathbf{x}'$ and $\widetilde{\mathbf{x}}$ respectively such that $P_{\widetilde{X}|X}(\widetilde{x}|x_1) \geq P_{\widetilde{X}|X}(\widetilde{x}|x_2)$.

We now prove the sufficient condition. Let $\mathcal{S} \subseteq \mathcal{X}^n$.

$$\frac{\sum_{\widetilde{\mathbf{x}}\in\mathcal{S}} P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\mathbf{x})}{\sum_{\widetilde{\mathbf{x}}\in\mathcal{S}} P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\mathbf{x}')} = \frac{\sum_{\widetilde{\mathbf{x}}\in\mathcal{S}}\prod_{i=1}^{n} P_{\widetilde{X}|X}(\widetilde{x}_i|x_i)}{\sum_{\widetilde{\mathbf{x}}\in\mathcal{S}}\prod_{i=1}^{n} P_{\widetilde{X}|X}(\widetilde{x}_i|x_i')} \leq e^{n\frac{\epsilon}{n}} = e^{\epsilon}$$

■

The proof can be easily extended to the case of k curators, where the necessary condition is that R satisfies $\frac{\epsilon}{k\lfloor n/2\rfloor}$-DP for the respondents and the sufficient condition is that R satisfies $\frac{\epsilon}{kn}$-DP for the respondents.

Supplementing PRAM with an additional random permutation step can enhance the level of privacy. However, this permutation must be performed by and synchronized between the curators[2], and kept private from the cloud (or any other attacker) in order to obtain the privacy enhancement. Hence, the curators must have a secure channel or other mechanism to obtain a private random permutation. The following lemma characterizes the privacy gains.

Lemma IV.5. *Let $d(\cdot,\cdot)$ be the norm one distance on the space of distributions. Let $\Pi : \mathcal{X}^n \to \mathcal{X}^n$ be a uniformly random permutation function. Let $R : \mathcal{X}^n \to \mathcal{X}^n$ be a PRAM mechanism governed by conditional distribution $P_{\widetilde{X}|X}$. Let $M = R \circ \Pi$ be a privacy mechanism that first randomly permutes the data and then applies the PRAM mechanism R.*

- ***Sufficient Condition:*** *If R satisfies $\frac{\epsilon}{\lfloor\frac{\delta n}{2}\rfloor}$-DP, then M satisfies δ-distributional ϵ-DP.*

Proof: Let $\mathbf{x},\mathbf{x}',\widetilde{\mathbf{x}} \in \mathcal{X}^n$ be such that $d(\mathbf{T}_{\mathbf{x}},\mathbf{T}_{\mathbf{x}'}) \leq \delta$, and let $\{\pi_1,...,\pi_{n!}\}$ be the set of possible permutation realizations for

[2] This synchronization is necessary to preserve joint statistics.

n respondents. The probability that mechanism M maps $\mathbf{x}$ into $\widetilde{\mathbf{x}}$ is given by

$$\Pr[M(\mathbf{x}) = \widetilde{\mathbf{x}}] = \frac{1}{n!}\sum_{i=1}^{n!} P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\pi_i(\mathbf{x})). \tag{3}$$

Since $d(\mathbf{T}_{\mathbf{x}}, \mathbf{T}_{\mathbf{x}'}) \leq \delta$, there exists a permutation π such that $\pi(\mathbf{x})$ differs from $\mathbf{x}'$ by at most $\frac{\delta n}{2}$ respondents. Thus, we have that

$$\begin{aligned}
\frac{\Pr[M(\mathbf{x}) = \widetilde{\mathbf{x}}]}{\Pr[M(\mathbf{x}') = \widetilde{\mathbf{x}}]} &= \frac{\frac{1}{n!}\sum_{i=1}^{n!} P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\pi_i(\mathbf{x}))}{\frac{1}{n!}\sum_{i=1}^{n!} P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\pi_i(\mathbf{x}'))} \\
&= \frac{\sum_{i=1}^{n!} P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\pi_i(\mathbf{x}))}{\sum_{i=1}^{n!} P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\pi_i(\pi(\mathbf{x}')))} \\
&\leq \max_i \frac{P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\pi_i(\mathbf{x}))}{P_{\widetilde{\mathbf{X}}|\mathbf{X}}(\widetilde{\mathbf{x}}|\pi_i(\pi(\mathbf{x}')))}, \\
&\leq e^{\epsilon}
\end{aligned}$$

where the last inequality holds since $\pi_i(\mathbf{x})$ and $\pi_i(\pi(\mathbf{x}'))$ differ by at most $\frac{\delta n}{2}$ respondents, and given that R satisfies $\frac{\epsilon}{\lfloor \frac{\delta n}{2} \rfloor}$-DP. ■

V. An Example Mechanism

We now present an example realization of the system framework given in Section II, where the privacy mechanisms are chosen to satisfy the privacy and utility requirements of Section III. The key requirements of this system can be summarized as follows:

(I) R_{AB} is a δ-distributional ϵ-differentially private mechanism.
(II) Helper information is generated by a ϵ-DP algorithm.
(III) R_A and R_B are PRAM mechanisms.

Since the perturbed data $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ are generated by a δ-distributional ϵ-differentially private mechanism, helper information is necessary to accurately estimate the marginal and joint type. To generate outputs that preserve different levels of privacy, the curators adopt a multilevel privacy approach [17]. As shown in Fig. 3, the databases are sanitized by a two-pass process. The first pass (i.e., $R_{AB,1}$) takes the true data (i.e., $\mathbf{X}, \mathbf{Y}$) as input and guarantees the respondent privacy while the second pass (i.e., $R_{AB,2}$) takes the sanitized output (i.e., $\widehat{\mathbf{X}}, \widehat{\mathbf{Y}}$) of the first pass as input and provides distributional privacy. The helper information is extracted during the second pass so as not to diminish respondent privacy. The mechanisms are constructed with the following constraints:

(i) $R_{A,2}$ and $R_{B,2}$ are $\frac{\epsilon}{2n}$-DP.
(ii) $R_{A,1}$ and $R_{B,1}$ are $\frac{\epsilon}{2}$-DP.
(iii) $R_{A,1}$, $R_{A,2}$, $R_{B,1}$ and $R_{B,2}$ are PRAM mechanisms.

By Lemma IV.3, the constraint (ii) implies that $R_{AB,1}$ is ϵ-DP and hence satisfies requirement (II) above[3]. Note that $R_A(\mathbf{X})$ can be viewed as $R_{A,2}(R_{A,1}(\mathbf{X}))$ and is governed by the conditional distribution (in matrix notation)

$$\mathbf{P}_{\widetilde{X}|X} = \mathbf{P}_{\widetilde{X}|\widehat{X}} \mathbf{P}_{\widehat{X}|X}.$$

Hence, constraint (iii) implies that requirement (III) is satisfied. By Lemma IV.1, IV.4, the constraint (i) implies that the requirement (I) is satisfied[4]. Now that we have shown that privacy requirements are satisfied, we proceed to show how researchers can compute the estimated types.

Recall that without presenting helper information, researchers cannot accurately estimate types due to requirement (I). In this example, the helper information consists of the conditional types $\mathbf{T}_{\widetilde{\mathbf{X}}|\widehat{\mathbf{X}}}$ and $\mathbf{T}_{\widetilde{\mathbf{Y}}|\widehat{\mathbf{Y}}}$ computed during the second pass. By [18], an unbiased estimate of $\mathbf{T}_{\mathbf{X}}$ computed from $\widetilde{\mathbf{X}}$ is given by $\mathbf{P}^{-1}_{\widetilde{X}|X}\mathbf{T}_{\widetilde{\mathbf{X}}}$ and the exact types can be recovered by $\mathbf{T}^{-1}_{\widetilde{\mathbf{X}}|\mathbf{X}}\mathbf{T}_{\widetilde{\mathbf{X}}}$. Thus, we have the following identities and estimators:

$$\mathbf{T}_{\widehat{\mathbf{X}}} = \mathbf{T}^{-1}_{\widetilde{\mathbf{X}}|\widehat{\mathbf{X}}}\mathbf{T}_{\widetilde{\mathbf{X}}}, \tag{4}$$

$$\dot{\mathbf{T}}_{\mathbf{X}} = \mathbf{P}^{-1}_{\widetilde{X}|\widehat{X}}\mathbf{T}_{\widehat{\mathbf{X}}} = \mathbf{P}^{-1}_{\widetilde{X}|\widehat{X}}\mathbf{T}^{-1}_{\widetilde{\mathbf{X}}|\widehat{\mathbf{X}}}\mathbf{T}_{\widetilde{\mathbf{X}}},$$

$$\mathbf{T}_{\widehat{\mathbf{Y}}} = \mathbf{T}^{-1}_{\widetilde{\mathbf{Y}}|\widehat{\mathbf{Y}}}\mathbf{T}_{\widetilde{\mathbf{Y}}}, \tag{5}$$

$$\dot{\mathbf{T}}_{\mathbf{Y}} = \mathbf{P}^{-1}_{\widetilde{Y}|\widehat{Y}}\mathbf{T}_{\widehat{\mathbf{Y}}} = \mathbf{P}^{-1}_{\widetilde{Y}|\widehat{Y}}\mathbf{T}^{-1}_{\widetilde{\mathbf{Y}}|\widehat{\mathbf{Y}}}\mathbf{T}_{\widetilde{\mathbf{Y}}}.$$

Note that the invertiblity of $\mathbf{P}_{\widetilde{X}|\widehat{X}}$ can be guaranteed by curators when selecting privacy mechanisms and the invertibility of $\mathbf{T}_{\widetilde{\mathbf{X}}|\widehat{\mathbf{X}}}$ can be guaranteed by sequence manipulation[5]. The estimation error in $\dot{\mathbf{T}}_{\mathbf{X}}$ and $\dot{\mathbf{T}}_{\mathbf{Y}}$ can be computed using the techniques in [19], [18].

Extending the results to compute the joint type presents some challenges. The matrix form of the conditional distribution of the collective mechanism R_{AB} is given by $\mathbf{P}_{\widetilde{X},\widetilde{Y}|X,Y} = \mathbf{P}_{\widetilde{X}|X} \otimes \mathbf{P}_{\widetilde{Y}|Y}$ where $\otimes$ is the Kronecker product [19]. An unbiased estimate of the joint type is given by

$$\begin{aligned}
\dot{\mathbf{T}}_{\mathbf{X},\mathbf{Y}} &= \mathbf{P}^{-1}_{\widetilde{X}\widetilde{Y}|X,Y}\mathbf{T}_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}} \\
&= ((\mathbf{P}_{\widetilde{X}|\widehat{X}}\mathbf{P}_{\widehat{X}|X}) \otimes (\mathbf{P}_{\widetilde{Y}|\widehat{Y}}\mathbf{P}_{\widehat{Y}|Y}))^{-1}\mathbf{T}_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}} \\
&= (\mathbf{P}_{\widetilde{X}|\widehat{X}}\mathbf{P}_{\widehat{X}|X})^{-1} \otimes (\mathbf{P}_{\widetilde{Y}|\widehat{Y}}\mathbf{P}_{\widehat{Y}|Y})^{-1}\mathbf{T}_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}} \\
&= (\mathbf{P}^{-1}_{\widehat{X}|X} \otimes \mathbf{P}^{-1}_{\widehat{Y}|Y})(\mathbf{P}^{-1}_{\widetilde{X}|\widehat{X}} \otimes \mathbf{P}^{-1}_{\widetilde{Y}|\widehat{Y}})\mathbf{T}_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}} \\
&= (\mathbf{P}^{-1}_{\widehat{X}|X} \otimes \mathbf{P}^{-1}_{\widehat{Y}|Y})\dot{\mathbf{T}}_{\widehat{\mathbf{X}},\widehat{\mathbf{Y}}}.
\end{aligned}$$

The estimation error in $\dot{\mathbf{T}}_{\widehat{\mathbf{X}},\widehat{\mathbf{Y}}}$ can be computed using the techniques in [19], [18]. Note that the helper information $\mathbf{T}_{\widetilde{\mathbf{X}}|\widehat{\mathbf{X}}}$ and $\mathbf{T}_{\widetilde{\mathbf{Y}}|\widehat{\mathbf{Y}}}$ can be used to compute $\mathbf{T}_{\widehat{\mathbf{X}}}$ and $\mathbf{T}_{\widehat{\mathbf{Y}}}$ exactly according to Equation 4 and 5. Simply replacing $\mathbf{P}_{\widetilde{X}|\widehat{X}}$ and $\mathbf{P}_{\widetilde{Y}|\widehat{Y}}$ with $\mathbf{T}_{\widetilde{\mathbf{X}}|\widehat{\mathbf{X}}}$ and $\mathbf{T}_{\widetilde{\mathbf{Y}}|\widehat{\mathbf{Y}}}$ in the above equations has negligible effect when n is large. Intuitively, this is because $\mathbf{T}_{\widetilde{\mathbf{X}}|\widehat{\mathbf{X}}}$ and $\mathbf{T}_{\widetilde{\mathbf{Y}}|\widehat{\mathbf{Y}}}$ can be viewed as sampled versions of $\mathbf{P}_{\widetilde{X}|\widehat{X}}$ and $\mathbf{P}_{\widetilde{Y}|\widehat{Y}}$, so that the types converge in the mean-squared sense to the distributions as n grows large.

However, unlike with the marginal types, the joint type $\mathbf{T}_{\widehat{\mathbf{X}},\widehat{\mathbf{Y}}}$ cannot be recovered exactly, even given the helper information $\mathbf{T}_{\widetilde{\mathbf{X}}|\widehat{\mathbf{X}}}$ and $\mathbf{T}_{\widetilde{\mathbf{Y}}|\widehat{\mathbf{Y}}}$. The reason for this is that, even though $P_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}|\widehat{\mathbf{X}},\widehat{\mathbf{Y}}}(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}|\widehat{\mathbf{x}},\widehat{\mathbf{y}}) = \prod_{i=1}^{n} P_{\widetilde{X}|\widehat{X}}(\widetilde{x}_i|\widehat{x}_i)P_{\widetilde{Y}|\widehat{Y}}(\widetilde{y}_i|\widehat{y}_i)$ holds, the identity $\mathbf{T}_{\widetilde{\mathbf{X}}|\widehat{\mathbf{X}}} \otimes \mathbf{T}_{\widetilde{\mathbf{Y}}|\widehat{\mathbf{Y}}} = \mathbf{T}_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}|\widehat{\mathbf{X}},\widehat{\mathbf{Y}}}$ would not necessarily be satisfied, which would be necessary for exact recovery of $\mathbf{T}_{\widehat{\mathbf{X}},\widehat{\mathbf{Y}}}$.

[3] One can view the helper information as obtained from postprocessing $(\widehat{\mathbf{X}}, \widehat{\mathbf{Y}})$. The second implication holds due to properties of the privacy axioms of [13], [14].

[4] Proof is obvious and omitted. It is not necessary that $R_{A,2}$ and $R_{B,2}$ are $\frac{\epsilon}{2n}$-DP as long as R_A and R_B are $\frac{\epsilon}{2n}$-DP. However, conditioning on $R_{A,1}$, $R_{B,1}$, $R_{A,2}$ and $R_{B,2}$ preserving as much as information as it can (i.e., they are in the form of the optimal PRAM suggested by [18]) and $R_{A,1}$ and $R_{B,1}$ are exactly $\frac{\epsilon}{2}$-DP (i.e. for any c and $\frac{\epsilon}{2} > c > 0$, $R_{A,1}$ and $R_{B,1}$ are not $(\frac{\epsilon}{2} - c)$-DP), one can show that $R_{A,2}$ and $R_{B,2}$ has to be $\approx \frac{\epsilon}{2n}$-DP. We suspect that if $R_{A,1}$, $R_{B,1}$ are optimized in any meaningful sense, $R_{A,2}$ and $R_{B,2}$ must be $\approx \frac{\epsilon}{2n}$-DP.

[5] The invertibility can be guaranteed via adding at most $|\mathcal{X}|$ dummy respondents $\widehat{x}_1, \ldots, \widehat{x}_{|\mathcal{X}|}$. After perturbing $\widehat{\mathbf{X}}$, Alice can choose $\widehat{x}_1, \ldots, \widehat{x}_{|\mathcal{X}|}$ and $\widetilde{x}_1, \ldots, \widetilde{x}_{|\mathcal{X}|}$ such that $\mathbf{T}_{\widetilde{\mathbf{X}}|\widehat{\mathbf{X}}}$ is invertible. Alice can pass this helper information to researchers without violating respondent privacy.

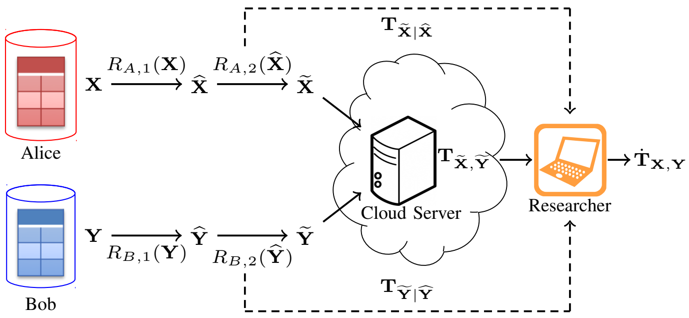

Alice
Bob
$\mathbf{X}$
$R_{A,1}(\mathbf{X})$
$\widehat{\mathbf{X}}$
$R_{A,2}(\widehat{\mathbf{X}})$
$\widetilde{\mathbf{X}}$
$\mathbf{T}_{\widetilde{\mathbf{X}}|\widehat{\mathbf{X}}}$
$\mathbf{Y}$
$R_{B,1}(\mathbf{Y})$
$\widehat{\mathbf{Y}}$
$R_{B,2}(\widehat{\mathbf{Y}})$
$\widetilde{\mathbf{Y}}$
$\mathbf{T}_{\widetilde{\mathbf{Y}}|\widehat{\mathbf{Y}}}$
Cloud Server
$\mathbf{T}_{\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}}}$
Researcher
$\dot{\mathbf{T}}_{\mathbf{X},\mathbf{Y}}$

Fig. 3. An example mechanism in which Alice and Bob (database curators) implement randomized responses in 2 passes, the first providing ϵ-DP to the database respondents, and the second providing δ-distributional ϵ-differential privacy to the empirical distribution of the data.

VI. Discussion

In this paper, we have presented a framework for distributed statistical analysis, in which authorized researchers can obtain, via a cloud server providing computational and storage resources, the empirical statistics of distributed databases, while preserving the privacy of the individual respondents. In addition to a conventional DP requirement for individual privacy (see Definition III.1), we also consider a distribution privacy requirement (see Definition III.2) protecting the aggregate information of the database (i.e., its statistics) from the cloud server. A key result given in (Lemma IV.4), demonstrating the stringency of distributional privacy, states that a mechanism providing distribution privacy requires that – and is implied given that – the mechanism satisfies individual privacy with a much stronger privacy parameter. A consequence of this implication is that the overall privacy of a mechanism must be very strong to provide distributional privacy. To avoid simultaneously reducing the corresponding utility, our proposed system adopts a multilevel approach, where a weaker first round of PRAM that provides individual privacy is followed by a stronger second round that provides distributional privacy against the cloud server. Helper information generated from the second round is provided to the researchers to allow them to overcome the second round of PRAM and recover database statistics.

Several interesting open problems and avenues of further exploration still remain toward completing and extending this privacy framework. For ease of exposition, we presented our framework with two curators and one researcher. However, our framework and results could be extended to any number of curators and researchers, and with flexible authorization (e.g., a researcher that is authorized by Alice but not by Bob). In some preliminary simulations with synthetic data, we have observed that the helper information enables accurate reconstruction of the marginal types, however the accuracy of the joint type reconstructed by the authorized researcher is poor even with the helper information. Our conjecture is that this is due to the fact that the matrices $\mathbf{P}_{\widetilde{X}|\widehat{X}}$ and $\mathbf{P}_{\widetilde{Y}|\widehat{Y}}$ in the Kronecker Product of equation (6) are ill-conditioned. Thus, it is still an open problem is to determine whether the low-rate helper information can allow accurate reconstruction of the joint type, and if not, what alternative helper information (if any) would enable accurate joint type reconstruction, while retaining the privacy of the respondents.

References

[1] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, Mar. 1965.

[2] ——, "The linear randomized response model." *Journal of American Statistical Association*, vol. 66, no. 336, pp. 884–888, Dec. 1971.

[3] C. Dwork, "Differential privacy: a survey of results," in *Proceedings of the 5th international conference on Theory and applications of models of computation*, ser. TAMC'08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 1–19.

[4] C. Dwork and A. Smith, "Differential privacy for statistics: What we know and what we want to learn," *Journal of Privacy and Confidentiality*, vol. 1, no. 2, 2009.

[5] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" in *Foundations of Computer Science, IEEE Annual Symposium on*, vol. 0. Los Alamitos, CA, USA: IEEE Computer Society, 2008, pp. 531–540.

[6] K. Chaudhuri, C. Monteleoni, and A. Sarwate, "Differentially private empirical risk minimization," in *J. Mach. Learn. Res.*, vol. 12. JMLR.org, Jul. 2011, pp. 1069–1109.

[7] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Trans. on Knowl. and Data Eng.*, vol. 13, no. 6, pp. 1010–1027, Nov. 2001.

[8] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," CMU, SRI, Tech. Rep., 1998.

[9] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, ser. SP '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 173–187.

[10] B.-C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala, "Privacy-preserving data publishing," *Foundations and Trends in Databases*, vol. 2, no. 1-2, pp. 1–167, 2009.

[11] C. Dwork, "Differential privacy," in *International Colloquium on Automata, Languages and Programming*, 2006, pp. 1–12.

[12] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, ser. SIGMOD '11. New York, NY, USA: ACM, 2011, pp. 193–204.

[13] D. Kifer and B.-R. Lin, "Towards an axiomatization of statistical privacy and utility," in *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ser. PODS '10. New York, NY, USA: ACM, 2010, pp. 147–158.

[14] ——, "An axiomatic view of statistical privacy and utility," To appear in Journal of Privacy and Confidentiality.

[15] S. Zhou, K. Ligett, and L. Wasserman, "Differential privacy with compression," in *Proc. IEEE International Symposium on Information Theory (ISIT)*. Piscataway, NJ, USA: IEEE Press, 2009, pp. 2718–2722.

[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis." in *Proceedings of the Third conference on Theory of Cryptography*, ser. TCC'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 265–284.

[17] X. Xiao, Y. Tao, and M. Chen, "Optimal random perturbation at multiple privacy levels," in *Proc. VLDB Endow.*, vol. 2, no. 1. VLDB Endowment, Aug. 2009, pp. 814–825.

[18] S. Agrawal and J. R. Haritsa, "A framework for high-accuracy privacy-preserving mining," in *Proceedings of the 21st International Conference on Data Engineering*, ser. ICDE '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 193–204.

[19] A. van den Hout and P. G. M. van der Heijden, "Randomized response, statistical disclosure control and misclassification: a review," *International Statistical Review*, vol. 70, pp. 269–288, Aug. 2002.