Privacy Against Statistical Inference

Flávio du Pin Calmon
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: flavio@mit.edu

Nadia Fawaz
Technicolor
Palo Alto, CA 94301
Email: nadia.fawaz@technicolor.com

***Abstract*—We propose a general statistical inference framework to capture the privacy threat incurred by a user that releases data to a passive but curious adversary, given utility constraints. We show that applying this general framework to the setting where the adversary uses the self-information cost function naturally leads to a non-asymptotic information-theoretic approach for characterizing the best achievable privacy subject to utility constraints. Based on these results we introduce two privacy metrics, namely average information leakage and maximum information leakage. We prove that under both metrics the resulting design problem of finding the optimal mapping from the user's data to a privacy-preserving output can be cast as a modified rate-distortion problem which, in turn, can be formulated as a convex program. Finally, we compare our framework with differential privacy.**

I. Introduction

A. Motivation

Increasing volumes of user data are being collected over wired and wireless networks, by a large number of companies who mine this data to provide personalized services or targeted advertising to users. As a consequence, privacy is gaining ground as a major topic in the social, legal, and business realms. This trend has spurred recent research in the area of theoretical models for privacy, and their application to the design of privacy-preserving services. Most privacy-preserving techniques, such as anonymization, k-anonymity [1] and differential privacy [2], are based on some form of perturbation of the data, either before or after the data is used in some computation. These perturbation techniques provide privacy guarantees at the expense of a loss of accuracy in the computation result, which leads to a privacy-accuracy trade-off.

In this paper, we consider the general setting where a user wishes to release a set of measurements to an analyst who provides a service (e.g. a recommendation system), while keeping data that are correlated with these measurements private. On one hand, the analyst is a legitimate receiver for these measurements, from which he expects to derive some utility. On the other hand, the correlation of these measurements with the user's private data gives the analyst the ability to illegitimately infer private information. The tension between the privacy requirements of the user and the utility expectations of the analyst gives rise to the problems of privacy-utility trade-off modeling, and the design of release schemes minimizing the privacy risks incurred by the user, while satisfying the utility constraints of the analyst.

B. Contributions

Our contributions are three-fold. First, we propose a general statistical inference framework to capture the privacy threat incurred by a user who releases information given certain utility constraints. The privacy risk is modeled as an inference cost gain by a passive but curious adversary upon observing the information released by the user. In broad terms, this cost gain represents the "amount of knowledge" learned by an adversary about the private data after observing the user's output. The design problem of finding the optimal mapping from the user's information to a privacy-preserving output is formulated as an optimization problem where the cost gain of the adversary is minimized for a given set of utility constraints. This formulation is general and given in terms of minimizing both the average and the maximum cost gain, being applicable to different cost functions.

Second, we apply this general framework to the case when the adversary uses the self-information cost function. We show how this naturally leads to a non-asymptotic information-theoretic framework to characterize the information leakage subject to utility constraints. Based on these results we introduce two privacy metrics, namely *average information leakage* and *maximum information leakage*. We also demonstrate that the problem of designing a privacy preserving mechanism that achieves the optimal privacy-accuracy tradeoff both for the average and maximum information leakage can be cast as modified rate-distortion problems. We then prove that these problems, in turn, can be expressed as convex programs. As a consequence, the privacy preserving mapping that achieves the optimal privacy-utility tradeoff can be efficiently found using convex minimization algorithms or widely available convex solvers.

Finally, we compare the average information leakage and maximum information leakage metrics with differential privacy. We show that differential privacy does not provide in general *any* privacy guarantees in terms of average or maximum information leakage. Furthermore, we introduce the definition of *information privacy*, and prove that information privacy implies both differential privacy and privacy in terms of (average or maximum) information leakage.

C. Related Work

In the privacy research community, a prevalent and strong notion of privacy is that of differential privacy [2], [3]. Differential privacy bounds the variation of the distribution of the released output given the input database, when the input database varies slightly, e.g. by a single entry. Intuitively, released outputs satisfying differential privacy render the distinction between ”neighboring” databases difficult. distinguish between. However, differential privacy neither provides guarantees, nor an intuition, on the amount of information leaked when a differentially private release occurs. Moreover, user data usually presents correlations. Differential privacy does not factor in correlations in user data, as the distribution of user data is not taken into account in this model. A natural question is how the notion of privacy proposed in this paper compares to that of differential privacy. We cover this question in more details in Section V.

Several approaches rely on information-theoretic tools to model privacy-accuracy trade-offs, such as [4]–[7]. Indeed, information theory, and more specifically rate-distortion theory, appear as natural frameworks to analyze the privacy-accuracy trade-off resulting from the distortion of correlated data. Although the approach we introduce in this paper involves information theoretic metrics, it is fundamentally different from previous information theoretic privacy models. Indeed, traditional information theoretic privacy models, such as [5], [7], focus on collective privacy for all or subsets of the entries of a database, and provide asymptotic guarantees on the average remaining uncertainty per database entry – or equivocation per input variable – after the output release. More precisely, the average equivocation per entry is modeled as the conditional entropy of the input variables given the released output, normalized by the number of input variables. In contrast, the general framework introduced in this paper provides privacy guarantees in terms of bounds on the inference cost gain that an adversary achieves by observing the released output. The use of a self-information cost yields a non-asymptotic information theoretic framework modeling the privacy risk in terms of information leakage. This framework, in turn, can be used to design practical privacy preserving mappings. Finally, we would like to point out that the formulation in [4], differs from previously mentioned information theoretic models, and addresses a particular case of the general framework introduced in this paper.

The paper is organized as follows. We describe the set-up and the threat model in Section II, and formulate the privacy-accuracy trade-off in Section III. Our main results and their proofs are presented in Section IV. Finally, in Section V we draw a comparison between the privacy notion proposed in this paper, and other existing privacy models, leading to the concluding remarks in Section VI.

II. General Setup and threat model

In this section we outline the general setup considered in this paper and the corresponding threat model.

A. General setup

We assume that there are two parties that communicate over a noiseless channel, namely Alice and Bob. Alice has access to a set of measurement points, represented by the variable $Y \in \mathcal{Y}$, that she wishes to transmit to Bob. At the same time, Alice requires that a set of variables $S \in \mathcal{S}$ should remain private, where S is jointly distributed with Y according to the distribution $(Y, S) \sim p_{Y,S}(y, s)$, $(y, s) \in \mathcal{Y} \times \mathcal{S}$. Depending on the considered setting, the variable S can be either directly accessible to Alice or inferred from Y. If no privacy mechanism was in place, Alice would simply transmit Y to Bob.

Bob has a utility requirement for the information sent by Alice. Furthermore, Bob is honest but curious, and will try to learn S from Alice’s transmission. Alice’s goal is to find and transmit a distorted version of Y, denoted by $U \in \mathcal{U}$, such that U satisfies a target utility constraint for Bob, but “protects” (in a sense made more precise later) the private variable S. We assume that Bob is passive but computationally unbounded, and will try to infer S based on U.

We consider, without loss of generality, that $S \to Y \to U$. Note that this model can capture the case where S is directly accessible by Alice by appropriately adjusting the alphabet $\mathcal{Y}$. For example, this can be done by representing $S \to Y$ as an injective mapping or allowing $\mathcal{S} \subset \mathcal{Y}$. In other words, even though the privacy mechanism is designed as a mapping from $\mathcal{Y}$ to $\mathcal{U}$, it is not limited to an output perturbation, and it encompasses input perturbation settings.

Definition 1. A privacy preserving mapping is a probabilistic mapping $g : \mathcal{Y} \to \mathcal{U}$ characterized by a transition probability $p_{U|Y}(u|y),\ y \in \mathcal{Y},\ u \in \mathcal{U}$.

Since the framework developed here results in formulations that are similar to the ones found in rate-distortion theory, we will use the term distortion to indicate a measure of utility. Furthermore, we will use the terms utility and accuracy interchangeably throughout the paper.

Definition 2. Let $d : \mathcal{Y} \times \mathcal{U} \to \mathbb{R}^+$ be a given distortion metric. We say that a privacy preserving mapping has distortion Δ if $\mathbb{E}_{Y,U}[d(Y,U)] \leq \Delta$.

We make the following assumptions:

1) Alice and Bob know the prior distribution of $p_{Y,S}(\cdot)$. This represents the side information that an adversary has.
2) Bob has complete knowledge of the privacy preserving mapping, i.e., g and $p_{U|Y}(\cdot)$ are known.

Note that this represents the *worst-case* statistical side information that an adversary can have about the input.

B. Threat model

We assume that Bob selects a revised distribution $q \in \mathcal{P}_S$, where $\mathcal{P}_S$ is the set of all probability distributions over $\mathcal{S}$, in order to minimize an expected cost $C(S,q)$. In other words, the adversary chooses q as the solution of the minimization

$$c_0^* = \min_{q \in \mathcal{P}_S} \mathbb{E}_S[C(S,q)] \tag{1}$$

prior to observing U, and

$$c_u^* = \min_{q \in \mathcal{P}_S} \mathbb{E}_{S|U}[C(S,q)|U=u] \tag{2}$$

after observing the output U. Note that this restriction on Bob models a very broad class of adversaries that perform statistical inference, capturing how an adversary acts in order to infer a revised belief distribution over the private variables S when observing U. After choosing this distribution, the adversary can perform an estimate of the input distribution (e.g. using a MAP estimator). However, the quality of the inference is inherently tied to the revised distribution q.

The average cost gain by an adversary after observing the output is

$$\Delta C = c_0^* - \mathbb{E}_U[c_u^*]. \tag{3}$$

The maximum cost gain by an adversary is measured in terms of the most informative output (i.e. the output that give the largest gain in cost), given by

$$\Delta C^* = c_0^* - \min_{u \in \mathcal{U}} c_u^*. \tag{4}$$

In the next section we present a formulation for the privacy-accuracy tradeoff based on this general setting.

III. A General Formulation for the Privacy-Accuracy Tradeoff

A. The privacy-accuracy tradeoff as an optimization problem

Our goal is to design privacy preserving mappings that minimize ΔC or ΔC^* for a given distortion level Δ, characterizing the fundamental privacy-utility tradeoff. More precisely, our focus is to solve optimization problems over $p_{U|Y} \in \mathcal{P}_{U|Y}$ of the form

$$\min\ \Delta C \text{ or } \Delta C^* \tag{5}$$

$$\text{s. t. } \mathbb{E}_{Y,U}[d(Y,U)] \leq \Delta\ , \tag{6}$$

where $\mathcal{P}_{U|Y}$ is the set of all conditional probability distributions of U given Y.

Remark 1. In the remainder of the paper we consider only one distortion constraint. However, it is straightforward to generalize the formulation and the subsequent optimization problems to multiple distinct distortion constraints $\mathbb{E}_{Y,U}[d_1(Y,U)] \leq \Delta_1, \ldots, \mathbb{E}_{Y,U}[d_n(Y,U)] \leq \Delta_n$. This can be done by simply adding an additional linear constraint to the convex program.

B. Application examples

We illustrate next how the proposed model can be cast in terms of privacy preserving queries and hiding features within data sets.

1) Privacy-preserving queries to a database: The framework described above can be applied to database privacy problems, such as those considered in differential privacy. In this case we denote the private variable as a vector $\mathbf{S} = S_1, \ldots, S_n$, where $S_j \in \mathcal{S}$, $1 \leq j \leq n$ and $S_1, \ldots, S_n$ are discrete entries of a database that represent, for example, the entries of n users. A (not necessarily deterministic) function $f : \mathcal{S}^n \to \mathcal{Y}$ is calculated over the database with output Y such that $Y = f(S_1, \ldots, S_n)$. The goal of the privacy preserving mapping is to present a query output U such that the individual entries $S_1, \ldots, S_n$ are "hidden", i.e. the estimation cost gain of an adversary is minimized according to the previous discussion, while still preserving the utility of the query in terms of the target distortion constraint. We illustrate this case with the counting query, which will be a recurring example throughout the rest of this paper.

Example 1 (Counting query)**.** Let $S_1, \ldots, S_n$ be entries in a database, and define:

$$Y = f(S_1, \ldots, S_n) = \sum_{i=1}^{n} \mathbb{1}_A(S_i), \tag{7}$$

where

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \text{ has property } A, \\ 0 & \text{otherwise.} \end{cases}$$

In this case there are two possible approaches: (i) output perturbation, where Y is distorted directly to produce U, and (ii) input perturbation, where each individual entry S_i is distorted directly, resulting in a new query output U.

2) Hiding dataset features: Another important particularization of the proposed framework is the obfuscation of a set of features S by distorting the entries of a data set Y. In this case $|\mathcal{S}| \ll |\mathcal{Y}|$, and S represents a set of features that might be inferred from the data Y, such as age group or salary. The distortion can be defined according to the the utility of a given statistical learning algorithm (e.g. a recommendation system) used by Bob.

IV. Privacy-accuracy tradeoff results

The formulation introduced in the previous section is general and can be applied to different cost functions. In this section we particularize the formulation to the case where the adversary uses the self-information cost function, as discussed below.

A. The self-information cost function

The *self information* (or *log-loss*) cost function is given by

$$C(S, q) = -\log q(S). \tag{8}$$

There are several motivations for using such a cost function. For an overview of the central role of the self-information cost function in prediction, we refer the reader to [8]. Briefly, the self-information cost function is the only local, proper and smooth cost function for an alphabet of size at least three. Furthermore, since the minimum self-information loss probability assignments are essentially ML estimates, this cost function is consistent with a "rational" adversary. In addition, the average cost-gain when using the self-information cost can be related to the cost gain when using any other bounded cost function [8]. Finally, as we will see below, this minimization implies a "closeness" constraint between the prior and a posteriori probability distributions in terms of KL-divergence. In Section V we compare the resulting privacy measure with that of differential privacy and *information-privacy*.

In the next sections we show how the cost minimization problems in (5) used with the self-information cost function can be cast as convex programs and, therefore, can be efficiently solved using interior point methods or widely available convex solvers.

B. Average information leakage

It is straightforward to show that for the log-loss function $c_0^* = H(S)$ and, consequently, $c_u^* = H(S|U=u)$, and, therefore

$$\Delta C = I(S; U) = \mathbb{E}_U[D(p_{S|U}||p_S)], \tag{9}$$

where $D(\cdot||\cdot)$ is the KL-divergence. The minimization (5) can the be rewritten according to the following definition.

Definition 3. The *average information leakage* of a set of features S given a privacy preserving output U is given by $I(S;U)$. A privacy-preserving mapping $p_{U|Y}(\cdot)$ is said to provide the *minimum average information leakage* for a distortion constraint Δ if it is the solution of the minimization

$$\begin{aligned} \min_{p_{U|Y}} \quad & I(S;U) && (10) \\ \text{s.t.} \quad & \mathbb{E}_{Y,U}[d(Y,U)] \leq \Delta\ . && (11) \end{aligned}$$

Observe that finding the mapping $p_{U|Y}(u|y)$ that provides the minimum information leakage is a modified rate-distortion problem. Alternatively, we can rewrite this optimization as

$$\begin{aligned} \min_{p_{U|Y}} \quad & \mathbb{E}_U[D(p_{S|U}||p_S)] && (12) \\ \text{s.t.} \quad & \mathbb{E}_{Y,U}[d(Y,U)] \leq \Delta\ . && (13) \end{aligned}$$

The minimization (12) has an interesting and intuitive interpretation. If we consider KL-divergence as a metric for the distance between two distributions, (12) states that the revised distribution after observing U should be as close as possible to the a priori distribution in terms of KL-divergence.

The following theorem shows how the the optimization in the previous definition can be expressed as a convex optimization problem. We note that this optimization is solved in terms of the unknowns $p_{U|Y}(\cdot|\cdot)$ and $p_{U|S}(\cdot|\cdot)$, which are coupled together through a linear equality constraint.

Theorem 1. *Given $p_{S,Y}(\cdot,\cdot)$, a distortion function $d(\cdot,\cdot)$ and a distortion constraint Δ, the mapping $p_{U|Y}(\cdot|\cdot)$ that minimizes the average information leakage can be found by solving the following convex optimization (assuming the usual simplex constraints on the probability distri-*

butions):

$$\min_{p_{U|Y}, p_{U|S}} \sum_{u\in\mathcal{U}}\sum_{s\in\mathcal{S}} p_{U|S}(u|s)p_S(s)\log\frac{p_{U|S}(u|s)}{p_U(u)} \tag{14}$$

$$\text{s.t.} \quad \sum_{u\in\mathcal{U}}\sum_{y\in\mathcal{Y}} p_{U|Y}(u|y)p_Y(y)d(u,y) \leq \Delta, \tag{15}$$

$$\sum_{y\in\mathcal{Y}} p_{Y|S}(y|s)p_{U|Y}(u|y) = p_{U|S}(u|s) \ \forall u,s, \tag{16}$$

$$\sum_{s\in\mathcal{S}} p_{U|S}(u|s)p_S(s) = p_U(u) \ \forall u. \tag{17}$$

Proof: Clearly the previous optimization is the same as (10). To prove the convexity of the objective function, note that $h(x,a) = ax\log x$ is convex for a fixed $a \geq 0$ and $x \geq 0$, and, therefore, the perspective of $g_1(x,z,a) = ax\log(x/z)$ is also convex in x and z for $z > 0, a \geq 0$ [9]. Since the objective function (14) can be written as

$$\sum_{u\in\mathcal{U}}\sum_{s\in\mathcal{S}} g(p_{U|S}(u|s), p_U(u), p_S(s)),$$

it follows the optimization is convex. In addition, since $p(u) \to 0 \Leftrightarrow p(u|s) \to 0 \ \forall u$, the minimization is well defined over the probability simplex. ∎

Remark 2. Note that the previous optimization can also be solved using a dual minimization procedure analogous to the Arimoto-Blahut algorithm [10] by starting at a fixed marginal probability $p_U(u)$, solving a convex minimization at each step (with an added linear constraint compared to the original algorithm) and updating the marginal distribution. However, the above formulation allows the use of efficient algorithms for solving convex problems, such as interior-point methods. In fact, the previous minimization can be simplified to formulate the traditional rate-distortion problem as a single convex program, not requiring the use of the Arimoto-Blahut algorithm.

Remark 3. The formulation in Theorem 1 can be easily extended to the case when U is determined directly from S, i.e. when Alice has access to S and the privacy preserving mapping is given by $p_{U|S}(\cdot|\cdot)$ directly. For this, constraint (16) should be substituted by

$$\sum_{y\in\mathcal{Y}} p_{Y|S}(y|s)p_{U|Y,S}(u|y,s) = p_{U|S}(u|s) \ \forall u,s, \tag{18}$$

and the following linear constraint added

$$\sum_{s\in\mathcal{S}} p_{S|Y}(s|y)p_{U|Y,S}(u|y,s) = p_{U|Y}(u|y) \ \forall u,y, \tag{19}$$

with the minimization being performed over the variables $p_{U|Y,S}(u|y,s), p_{U|Y}(u|y)$ and $p_{U|S}(u|s)$, with the usual simplex constraints on the probabilities.

We now particularize the previous result for the case where Y is a deterministic function of S.

Corollary 1. *If Y is a deterministic function of S and $S \to Y \to U$ then the minimization in* (10) *can be simplified to a rate-distortion problem:*

$$\min_{p_{U|Y}} \ I(Y;U) \tag{20}$$

$$\text{s. t.} \ \ \mathbb{E}_{Y,U}[d(Y,U)] \leq D \ . \tag{21}$$

Furthermore, by restricting $U = Y + Z$ and $d(Y,U) = d(Y-U)$, the optimization reduces to

$$\max_{p_Z} \ H(Z) \tag{22}$$

$$\textit{s. t.} \ \ \mathbb{E}_Z[d(Z)] \leq \Delta \ . \tag{23}$$

Proof: Since Y s a deterministic function of S and $S \to Y \to U$, then

$$\begin{aligned} I(S;U) &= I(S,Y;U) - I(Y;U|S) \qquad (24)\\ &= I(Y;U) + I(S;U|Y) - I(Y;U|S) \qquad (25)\\ &= I(Y;U), \qquad (26) \end{aligned}$$

where (26) follows from the fact that Y is a deterministic function of S ($I(Y;U|S) = 0$) and $S \to Y \to U$ ($I(S;U|Y) = 0$). For the additive noise case, the result follows by observing that $H(Y|U) = H(Z)$. ∎

C. Maximum information leakage

The minimum over all possible maximum cost gains of an adversary that uses a log-loss function in (4) is given by

$$C^* = \max_{u\in\mathcal{U}} H(S) - H(S|U=u).$$

The previous expression motivates the definition of *maximum information leakage,* presented below.

Definition 4. The *maximum information leakage* of a set of features S is defined as the maximum cost gain, given in terms of the log-loss function, that an adversary obtains by observing a single output, and is given by $\max_{u\in\mathcal{U}} H(S) - H(S|U=u)$. A privacy-preserving mapping $p_{U|Y}(\cdot)$ is said to achieve the *minmax information leakage* for a distortion constraint Δ if it is a solution of the minimization

$$\min_{p_{U|Y}} \max_{u\in\mathcal{U}} \ H(S) - H(S|U=u) \tag{27}$$

$$\text{s. t.} \ \mathbb{E}[d(U,Y)] \leq \Delta \tag{28}$$

The following theorem demonstrates how the mapping that achieves the minmax information leakage can be determined as the solution of a related convex program that finds the minimum distortion given a constraint on the maximum information leakage.

Theorem 2. *Given $p_{S,Y}(\cdot,\cdot)$, a distortion function $d(\cdot,\cdot)$ and a constraint ϵ on the maximum information leakage, the minimum achievable distortion and the mapping that achieves the minmax information leakage can be found by solving the following convex optimization (assuming the implicit simplex constraints on the probability distributions):*

$$\min_{p_{U|Y}, p_{U|S}} \sum_{u\in\mathcal{U}}\sum_{s\in\mathcal{S}} p_{U|Y}(u|y)p_Y(y)d(u,y) \tag{29}$$

$$\text{s.t.} \sum_{y\in\mathcal{Y}} p_{Y|S}(y|s)p_{U|Y}(u|y) = p_{U|S}(u|s)\ \forall u,s, \tag{30}$$

$$\sum_{s\in\mathcal{S}} p_{U|S}(u|s)p_S(s) = p_U(u)\ \forall u, \tag{31}$$

$$\delta p_U(u) + \sum_{s\in\mathcal{S}} p_{U,S}(u,s)\log\frac{p_{U,S}(u,s)}{p_U(u)} \leq 0\ \forall u, \tag{32}$$

where $\delta = H(S) - \epsilon$. Therefore, for a given value of Δ, the optimization problem in (27) *can be efficiently solved with arbitrarily large precision by performing a line-search over $\epsilon \in [0, H(S)]$ and solving the previous convex program at each step of the search.*

Proof: The convex program in (27) can be reformulated to return the minimum distortion for a given constraint ϵ on the minmax information leakage as

$$\min_{p_{U|Y}} \ \mathbb{E}[d(U,Y)] \tag{33}$$

$$\text{s.t. } H(S|U=u) \geq \delta\ . \tag{34}$$

It is straightforward to verify that constraint (32) can be written as (34). Following the same steps as the proof of Theorem 1 and noting that the function $g_2(x,z,a) = ax\log(ax/z)$ is convex for $a,x \geq 0$, $z > 0$, it follows that (34) and, consequently, (32), is a convex constraint. Finally, since the optimal distortion value in the previous program is a decreasing function of ϵ, it follows that the solution of (27) can be found through a line-search in ϵ. ■

Remark 4. Analogously to the average information leakage case, the convex program presented in Theorem (2) can be extended to the setting where the privacy preserving mapping is given by $p_{U|S}(\cdot|\cdot)$ directly. This can be done by substituting (31) by (18) and adding the linear constraint (19).

Even though the convex program presented in Theorem 2 holds in general, it does not provide much insight on the structure of the privacy mapping that minimizes the maximum information leakage for a given distortion constraint. In order to shed light on the nature of the optimal solution, we present the following result for the particular case when Y is a deterministic function of S and $S \to Y \to U$.

Corollary 2. *For $Y = f(S)$, where $f : \mathcal{S} \to \mathcal{Y}$ is a deterministic function, $S \to Y \to U$ and a fixed prior $p_{Y,S}(\cdot,\cdot)$, the privacy preserving mapping that minimizes the maximum information leakage is given by*

$$p^*_{U|Y} = \arg\min_{p_{U|Y}} \max_{u\in\mathcal{U}} D(p_{Y|U}||\zeta) \tag{35}$$

$$\text{s.t. } \mathbb{E}[d(U,Y)] \leq \Delta,$$

where $\zeta(y) = \frac{2^{H(S|Y=y)}}{\sum_{y'\in\mathcal{Y}} 2^{H(S|Y=y')}}$.

Proof: Under the assumptions of the corollary, note that for a given $u \in \mathcal{U}$ (and assuming that the logarithms are in base 2)

$$\begin{aligned}
&H(S|U=u) = \\
&-\sum_{s\in\mathcal{S}} p_{S|U}(s|u)\log p_{S|U}(s|u) \\
&= -\sum_{s\in\mathcal{S}}\left(\sum_{y\in\mathcal{Y}} p_{S|Y}(s|y)p_{Y|U}(y|u)\right) \\
&\times\left(\log\sum_{y'\in\mathcal{Y}} p_{S|Y}(s|y')p_{Y|U}(y'|u)\right) \\
&= -\sum_{s\in\mathcal{S}} p_{S|Y}(s|f(s))p_{Y|U}(f(s)|u) \\
&\quad \times \log p_{S|Y}(s|f(s))p_{Y|U}(f(s)|u) && (36) \\
&= -\sum_{s\in\mathcal{S},y\in\mathcal{Y}} p_{S|Y}(s|y)p_{Y|U}(y|u)\log p_{S|Y}(s|y)p_{Y|U}(y|u) && (37) \\
&= H(Y|U=u) + \sum_{y\in\mathcal{Y}} p_{Y|U}(y|u)H(S|Y=y) && (38) \\
&= \sum_{y\in\mathcal{Y}} p_{Y|U}(y|u)\log\frac{2^{H(S|Y=y)}}{p_{Y|U}(y|u)} && (39) \\
&= -D(p_{Y|U}||\zeta) + \log\left(\sum_{y\in\mathcal{Y}} 2^{H(S|Y=y)}\right), && (40)
\end{aligned}$$

where (36) and (37) follows by noting that $p_{S|Y}(s|y) = 0$ if $y \neq f(s)$. The result follows directly by substituting (40) in (27). ■

For Y a deterministic function of S, the optimal privacy preserving mechanism is the one that approximates (in terms of KL-divergence) the posterior distribution of Y given U to $\zeta(\cdot)$. Note that the distribution $\zeta(\cdot)$ captures the inherent uncertainty that exists in the function f for different outputs $y \in \mathcal{Y}$. The purpose of the privacy preserving mapping is then to augment this uncertainty, while still satisfying the distortion constraint. In particular, the larger the uncertainty $H(S|Y=y)$, the larger the probability of $p_{Y|U}(y|u)$ for all u. Consequently, the optimal privacy mapping (exponentially) reinforces the posterior probability of the values of y for which there is a large uncertainty regarding the features S. This fact is illustrated in the next example, where we revisit the counting query presented in Example 1.

Example 2 (Counting query continued)**.** Assume that each database input S_i, $1 \leq i \leq n$ satisfies $\Pr(\mathbb{1}_A(S_i) = 1) = p$ and are independent and identically distributed. Then Y is a binomial random variable with parameter (n,p). It follows that $H(\mathbf{S}|Y=y) = \log \binom{n}{y}$. Consequently, the optimal privacy preserving mapping will be the one that results in a posterior probability $p_{Y|U}(y|u)$ that is proportional to the size of the pre-image of y, i.e. $p_{Y|U}(y|u) \propto |f^{-1}(y)| = \binom{n}{y}$.

V. Comparison of privacy metrics

We now compare average information leakage and maximum information leakage with differential privacy and *information privacy*, the latter being a new metric introduced in this section. We first recall the definition of differential privacy, presenting it in terms of the model discussed in Section II and assuming that the set of features $\mathbf{S}$ is a vector given by $\mathbf{S} = (S_1, \ldots, S_n)$, where $S_i \in \mathcal{S}$.

Definition 5 ([3])**.** A privacy preserving mapping $p_{U|\mathbf{S}}(\cdot|\cdot)$ provides ϵ-differential privacy if for all inputs $\mathbf{s}_1$ and $\mathbf{s}_2$ differing in at most one entry and all $B \subseteq \mathcal{U}$,

$$\Pr(U \in B|\mathbf{S} = \mathbf{s}_1) \leq \exp(\epsilon) \times \Pr(U \in B|\mathbf{S} = \mathbf{s}_2) \ . \tag{41}$$

An alternative (and much stronger) definition of privacy, related to the one presented in [6] is given below. We note that this definition is unwieldy, but explicitly captures the ultimate goal in privacy: the posterior and prior probabilities of the features S do not change significantly given the output.

Definition 6. A privacy preserving mapping $p_{U|\mathbf{S}}(\cdot|\cdot)$ provides *ϵ-information privacy* if for all $\mathbf{s} \subseteq \mathcal{S}^n$:

$$\exp(-\epsilon) \leq \frac{p_{\mathbf{S}|U}(\mathbf{s}|u)}{p_{\mathbf{S}}(\mathbf{s})} \leq \exp(\epsilon) \ \forall u \in \mathcal{U} : p_U(u) > 0. \tag{42}$$

Note that ϵ-information privacy implies directly 2ϵ-differential privacy and maximum information leakage of at most $\epsilon/\ln 2$ bits, as shown below.

Theorem 3. *If a privacy preserving mapping $p_{U|\mathbf{S}}(\cdot|\cdot)$ is ϵ-information private for some input distribution such that supp$(p_U) = \mathcal{U}$, then it is at least 2ϵ-differentially private and leaks at most $\epsilon/\ln 2$ bits on average.*

Proof: Note that for a given $B \subseteq \mathcal{U}$

$$\begin{aligned}\frac{\Pr(U \in B|\mathbf{S} = \mathbf{s}_1)}{\Pr(U \in B|\mathbf{S} = \mathbf{s}_2)} &= \frac{\Pr(\mathbf{S} = \mathbf{s}_1|U \in B)\Pr(\mathbf{S} = \mathbf{s}_2)}{\Pr(\mathbf{S} = \mathbf{s}_2|U \in B)\Pr(\mathbf{S} = \mathbf{s}_1)} \\ &\leq \exp(2\epsilon),\end{aligned}$$

where the last step follows from (41). Clearly if $\mathbf{s}_1$ and $\mathbf{s}_2$ are neighboring vectors (i.e. differ by only one entry), then 2ϵ-differential privacy is satisfied. Furthermore

$$\begin{aligned}H(\mathbf{S}) - H(\mathbf{S}|U=u) &= \sum_{\mathbf{s} \in \mathcal{S}^n} p_{\mathbf{S}|U}(\mathbf{s}|u) p_U(u) \log \frac{p_{\mathbf{S}|U}(\mathbf{s}|u)}{p_{\mathbf{S}}(\mathbf{s})} \\ &\leq \sum_{\mathbf{s} \in \mathcal{S}^n, u \in \mathcal{U}} p_{\mathbf{S}|U}(\mathbf{s}|u) p_U(u) \frac{\epsilon}{\ln 2} \\ &= \frac{\epsilon}{\ln 2}\end{aligned}$$

■

We show in the next theorem that differential privacy *does not guarantee* privacy in terms of average information leakage *in general* and, consequently in terms of maximum information leakage and information privacy. More specifically, guaranteeing that a mechanism is ϵ-differentially private *does not* provide *any* guarantee on the information leakage.

Theorem 4. *For every $\epsilon > 0$ and $\delta \geq 0$, there exists an $n \in \mathbb{Z}_+$, sets $\mathcal{S}^n$ and $\mathcal{U}$, a prior $p_{\mathbf{S}}(\cdot)$ over $\mathcal{S}^n$ and a privacy mapping $p_{U|S}(\cdot|\cdot)$ that is ϵ-differentially private but leaks at least δ bits on average.*

Proof: We prove the statement by explicitly constructing an example that is ϵ-differentially private, but an arbitrarily large amount of information can leak on average from the system. For this, we return to the counting query discussed in examples 1 and 2 with, the sets $\mathcal{S}$ and $\mathcal{Y}$ being defined accordingly, and letting $\mathcal{U} = \mathcal{Y}$. We do not assume independence of the inputs.

For the counting query and for any given prior, adding Laplacian noise to the output provides ϵ-differential privacy [3]. More precisely, for the output of the query given in (7), denoted as $Y \sim p_Y(y), 0 \leq y \leq n$, the mapping

$$U = Y + N, \quad N \sim \text{Lap}(1/\epsilon), \tag{43}$$

where the pdf of the additive noise N given by

$$p_N(r;\epsilon) = \frac{\epsilon}{2}\exp(-|r|\epsilon), \tag{44}$$

is ϵ-differentially private. Now assume that ϵ is given, and denote $\mathbf{S} = (X_1, \ldots, X_n)$. Set k and n such that $n \mod k = 0$, and let $p_{\mathbf{S}}(\cdot)$ be such that

$$p_Y(y) = \begin{cases} \frac{1}{1+n/k} & \text{if } y \mod k = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{45}$$

With the goal of lower-bounding the information leakage, assume that Bob, after observing U, maps it to the nearest value of y such that $p_Y(y) > 0$, i.e. does a maximum a posteriori estimation of Y. The probability that Bob makes a correct estimation (and neglecting edge effects), denoted by $\alpha_{k,n}(\epsilon)$, is given by:

$$\alpha_{k,n}(\epsilon) = \int_{\frac{-k}{2}}^{\frac{k}{2}} \frac{\epsilon}{2}\exp(-|x|\epsilon)dx = 1 - \exp\left(-\frac{k\epsilon}{2}\right). \tag{46}$$

Let E be a binary random variable that indicates the event that Bobs makes a wrong estimation of Y given U. Then

$$\begin{aligned} I(Y;U) &\geq I(E,Y;U) - 1 \\ &\geq I(Y;U|E) - 1 \\ &\geq \Pr\{E=0\}I(Y;U|E=0) - 1 \\ &= \left(1 - e^{-\frac{k\epsilon}{2}}\right)\log\left(1 + \frac{n}{k}\right) - 1, \end{aligned}$$

which can be made arbitrarily larger than δ by appropriately choosing the values of n and k. Since Y is a deterministic function of $\mathbf{S}$, $I(Y;U) = I(\mathbf{S};U)$, as shown in the proof of Corollary 1, and the result follows. ■

The counterexample used in the proof of the previous theorem can be extended to allow the adversary to recover *exactly* the inputs generated the ouput U. This can be done by assuming that the inputs are ordered and correlated in such a way that $Y = y$ if and only if $S_1 = 1, \ldots, S_y = 1$. In this case, for n and k sufficiently large, the adversary can exploit the input correlation to correctly learn the values of $S_1, \ldots, S_n$ with arbitrarily high probability.

Differential privacy does not necessarily guarantee low leakage of information – in fact, an arbitrarily large amount of information can be leaking from a differentially private system, as shown in Theorem 4. This is a serious issue when using solely the differential privacy definition as a privacy metric. In addition, it follows as a simple extension of [11, Prop. 4.3] that $I(S;U) \leq O(\epsilon n)$, corroborating that differential privacy does not bound above the average information leakage when n is sufficiently large.

Nevertheless, differential privacy does have an operational advantage since it does not require any prior information. However, by neglecting the prior and requiring differential privacy, the resulting mapping might not be *de facto* private, being suboptimal under the information leakage measure. We note that the presented formulations can be made prior independent maximizing the minimum information leakage over a set of possible priors. This problem is closely related to universal coding [10].

VI. Conclusions

In this paper we presented a general statistical inference framework to capture the privacy threat incurred by a user that releases data to a passive but curious adversary given utility constraints. We demonstrated how under certain assumptions this framework naturally leads to an information-theoretic approach to privacy. The design problem of finding privacy-preserving mappings for minimizing the information leakage from a user's data with utility constraints was formulated as a convex program. This approach can lead to practical and deployable privacy-preserving mechanisms. Finally, we compared our approach with differential privacy, and showed that the differential privacy requirement does not necessarily constrain the information leakage from a data set.

References

[1] L. Sweeney, "K-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.

[2] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *TCC*, 2006. [Online]. Available: http://www.cs.bgu.ac.il/~kobbi/papers/sensitivity-tcc-final.pdf

[3] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*. Springer, 2006, vol. 4052, pp. 1–12.

[4] I. S. Reed, "Information Theory and Privacy in Data Banks," in *Proceedings of the June 4-8, 1973, national computer conference and exposition*, ser. AFIPS '73. ACM, 1973, pp. 581–587.

[5] H. Yamamoto, "A source coding problem for sources with additional outputs to keep secret from the receiver of wiretappers," *IEEE Trans. Inf. Theory*, vol. 29, no. 6, 1983.

[6] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proceedings of the twenty-second ACM Symposium on Principles of Database Systems*, New York, NY, USA, 2003, pp. 211–222.

[7] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "A Theory of Privacy and Utility in Databases," *ArXiv e-prints*, Feb. 2011. [Online]. Available: http://arxiv.org/abs/1102.3751

[8] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. on Inform. Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.

[9] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Mar. 2004.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition*, 2nd ed. Wiley-Interscience, Jul. 2006.

[11] A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. P. Vadhan, "The limits of two-party differential privacy," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 18, no. 106, 2011.