Information-Theoretic Privacy Watchdogs

Hsiang Hsu, Shahab Asoodeh, and Flavio P. Calmon
School of Engineering and Applied Sciences, Harvard University,
{hsianghsu, shahab, fcalmon}@g.harvard.edu,

***Abstract*—Given a dataset comprised of individual-level data, we consider the problem of identifying samples that may be disclosed without incurring a privacy risk. We address this challenge by designing a mapping that assigns a "privacy-risk score" to each sample. This mapping, called the *privacy watchdog*, is based on a sample-wise information leakage measure called the information density, deemed here *lift privacy*. We show that lift privacy is closely related to well-known information-theoretic privacy metrics. Moreover, we demonstrate how the privacy watchdog can be implemented using the Donsker-Varadhan representation of KL-divergence. Finally, we illustrate this approach on a real-world dataset.**

I. Introduction

Consider a data scientist, Alice, who has in hand a dataset $\mathcal{D}^n = \{(s_i, x_i)\}_{i=1}^n$ collected from n individuals. We assume that each entry (s_i, x_i) of the dataset is drawn i.i.d. from $P_{S,X}$, where S represents an individual's private/sensitve features (e.g., political preference) and X the remaining features (e.g., social media posts). Alice wishes to publish the dataset $\{x_i\}_{i=1}^n$, yet knows that doing so may incur a privacy risk: by observing x_i, a malicious party may gain information about the private feature, i.e., $P_{S|X=x_i}$ can be significantly different from P_S. However, not all realizations x_i are equally informative, and certain values could potentially be disclosed with minimal privacy risk, i.e., $P_{S|X=x_i} \approx P_S$ for some x_i. How can Alice identify the entries of $\{x_i\}_{i=1}^n$ that pose the highest (or lowest) privacy threat?

We address this challenge by designing a *privacy watchdog*: a mapping that assigns a privacy-risk score to each sample in the dataset $\mathcal{D}^n$. Ideally, the watchdog should flag samples that must be perturbed in order to ensure privacy, while indicating which samples can be perfectly disclosed without excessive harm. Moreover, the watchdog should be data-driven, learning from the dataset which outcomes of X pose a privacy risk.

To construct a privacy watchdog, we adopt a sample-wise information leakage measure. A natural choice is the ratio

$$l(s,x) \triangleq \frac{P_{S,X}(s,x)}{P_S(s)P_X(x)} = \frac{P_{S|X}(s|x)}{P_S(s)}, \ \forall (s,x) \in \mathcal{S} \times \mathcal{X}, \quad (1)$$

referred to as the *lift* [1] in the data mining literature. The logarithm of the lift (log-lift) $i(s,x) \triangleq \log l(s,x)$ is, of course, the *information density*, and plays a central role in spectral methods in information theory and finite-blocklength analysis [2]. The lift is at the heart of most information-theoretic measures of privacy.

In this paper, we derive properties of lift as a privacy metric, and show that an upper bound on lift (1) leads to upper bounds on several other information-theoretic privacy measures, including those based on Arimoto's [3] and Sibson's [4] mutual information, f-divergences [5], and local differential privacy [6]. Moreover, we demonstrate how a privacy-assuring mapping that merely perturbs the samples with large (absolute) log-lift has favorable performance guarantees in terms of privacy and utility. Of greater practical interest, we use variational representations of divergence metrics [7] (and the Donsker-Varadhan representation in particular) to build lift-based privacy watchdogs using neural networks. We illustrate this approach on ProPublica's COMPAS recidivism dataset [8].

The design of privacy mechanisms is an imminent topic in computer science [9], data mining [10], and information theory [4], [11]–[15] communities. Within the latter, there has been significant effort to characterize fundamental trade-offs between privacy and utility (e.g., [11], [15]), as well as produce privacy metrics with operational significance (e.g., [3], [4], [16]). We also note that variations of information density were mentioned in [10]–[12] as a measure of privacy. Here, we widen our focus beyond the analysis of privacy mechanisms and associated trade-offs to consider the practical challenge faced by Alice. The privacy watchdog proposed here can be applied to real-world datasets (as illustrated in Section IV), and naturally serves as a building block for other privacy mechsmisms (e.g., distorting data in accordance to the risk scores given by the watchdog). Our ultimate goal is to create a richer information-theoretic toolset for addressing privacy challenges commonly found in data science.

The remainder of the paper is organized as follows. We introduce notation and preliminaries next, and examine the properties of lift as a privacy metric in Section II. We formulate the privacy watchdog and explore its application in Section III and finally consider implementation and evaluation with data in Section IV. We omit the proofs of all results due to space limitations. The proofs are available in [17].

A. Notation

Capital and calligraphic letters are used to denote random variables and sets, respectively. We also use boldface lowercase letter to denote vectors. We use $P_{S,X}$, for joint probability distribution of S and X, $P_{S|X}$ for conditional probability distribution of S given X, and P_S and P_X for marginal probability distributions of S and X, respectively. When X is distributed according to P_X, we write $X \sim P_X$. We denote ℓ_p-norm of an n-length vector $\mathbf{z}$ by $\|\mathbf{z}\|_p = (\sum_{i=1}^n \mathbf{z}_i^p)^{\frac{1}{p}}$, where $\mathbf{z}_i$ is the i^{th} entry of $\mathbf{z}$. We write $1_{\{\cdot\}}$ for the indicator function which returns 1 if the condition in the parentheses is satisfied and 0 otherwise.

Let $f:(0,\infty)\to\mathbb{R}$ be a convex function satisfying $f(1)=0$. Assume that P and Q are two probability distributions over a finite set $\mathcal{X}$ and that $P\ll Q$. The f-divergence [18] between P and Q is given by

$$D_f(P\|Q)\triangleq\mathbb{E}_Q\left[f\left(\frac{P(X)}{Q(X)}\right)\right], \tag{2}$$

where $\mathbb{E}_Q$ denotes expectation with respect to distribution Q. This definition can be used to generalize Shannon's mutual information. Replacing P and Q by $P_{S,X}$ and P_SP_X, one can define f-information between S and X as

$$I_f(S;X)\triangleq D_f(P_{S,X}\|P_SP_X). \tag{3}$$

Kullback-Leibler (KL) divergence $D(P\|Q)$ and Shannon's mutual information $I(S;X)$ are special cases of (2) and (3), respectively, when $f(t)=t\log t$.

II. Lift-Based Measure of Information Leakage

In this section, we first overview the privacy definition used to design the watchdog, called ε-lift privacy, and then derive some of its properties. In particular, we show ε-lift privacy is closely related to other existing measures of information leakages such as local differential privacy [6], maximal leakage [4], α-leakage [16], and f-information [5]. We note that variations of ε-lift privacy have appeared in the literature under different guises (e.g., [10, Defn. 1] and [11, Defn. 6]).

A. ε-Lift Privacy

The value of log-lift $i(s,x)$ indicates whether the sample x carries significant information about private feature s. This intuition naturally leads to the following definition.

Definition 1 (ε-lift privacy [11])**.** For $(S,X)\sim P_{S,X}$, we say X is an ε-lift private version of S if

$$-\varepsilon\le i(s,x)\le\varepsilon,\qquad\forall(s,x)\in\mathcal{S}\times\mathcal{X}. \tag{4}$$

In the following lemma, we demonstrate several properties of ε-lift privacy.

Lemma 1. *If X is an ε-lift private version of S, then*

1) *S is an ε-lift private version of X.*
2) *$P_{S|X}$ is 2ε-locally differentially private [6], i.e.*

$$\sup_{\forall s\in\mathcal{S},x,x'\in\mathcal{X}}\frac{P_{S|X}(s|x)}{P_{S|X}(s|x')}\le e^{2\varepsilon}. \tag{5}$$

3) *The mutual information $I(S;X)$ is upper bounded by ε.*

Lemma 1 sheds light on the privacy guarantees that an ε-lift privacy constraint can provide. In particular, if X is an ε-lift private version of S, then X cannot reveal more than ε nats of information (on average) about S. Next, we explore further connections between ε-lift privacy and information-theoretic measures of leakage.

B. Other Information Leakage Measures

Arimoto's and Sibson's mutual information and f-information have recently been proposed as operational measures for information leakage, see e.g, [3] and [4]. Arimoto's mutual information of order $\alpha\in(1,\infty)$ is given by [19]

$$I_\alpha^{\mathsf{A}}(S;X)\triangleq\frac{\alpha}{\alpha-1}\log\frac{\mathbb{E}_X[\|P_{S|X}(\cdot|X)\|_\alpha]}{\|P_S\|_\alpha}. \tag{6}$$

It can also be defined (by continuity) for the extreme cases $\alpha=1$ and ∞, respectively, as $\lim_{\alpha\to1}I_\alpha^{\mathsf{A}}(S;X)=I(S;X)$ and $I_\infty^{\mathsf{A}}(S;X)\triangleq\lim_{\alpha\to\infty}I_\alpha^{\mathsf{A}}(S;X)$, and the latter characterizes the ability of an adversary to correctly guess S given X. In particular, it can be verified [3] that $I_\infty^{\mathsf{A}}(S;X)=\log\frac{\mathsf{P_c}(S|X)}{p_S^*}$, where

$$\mathsf{P_c}(S|X)\triangleq\max_{g\mathcal{X}\to\mathcal{S}}\Pr(S=g(X))=\sum_{x\in\mathcal{X}}\max_{s\in\mathcal{S}}P_{S,X}(s,x),$$

denotes the *probability of correctly guessing* S given X and $p_S^*\triangleq\max_{s\in\mathcal{S}}P_S(s)$, thus providing an operational meaning for $I_\infty^{\mathsf{A}}(S;X)$.

Another operational measure of information leakage recently proposed is Sibson's mutual information [19] of order $\alpha\in(1,\infty)$ between S and X, which is given by

$$I_\alpha^{\mathsf{S}}(S;X)\triangleq\inf_{Q_X}D_\alpha(P_{S,X}\|P_SQ_X), \tag{7}$$

where $D_\alpha(P\|Q)\triangleq\frac{1}{\alpha-1}\log\left(\sum_xP(x)^\alpha Q(x)^{1-\alpha}\right)$ is the Rényi divergence. One can similarly define $I_\infty^{\mathsf{S}}(S;X)$ as the limit of $I_\alpha^{\mathsf{S}}(S;X)$ when $\alpha\to\infty$. This quantity, termed *maximal leakage*, was recently shown to bear an interesting interpretation in terms of worst-case privacy threats [4]. More precisely, maximal leakage is equal to the logarithm of the multiplicative gain in guessing *any function* of S given the observation of X, that is

$$I_\infty^{\mathsf{S}}(S;X)=\max_{U-S-X}\log\frac{\mathsf{P_c}(U|X)}{p_U^*}, \tag{8}$$

where the maximization is taken over random variable U forming the Markov chain $U-S-X$.

Recall that Lemma 1 established a connection between ε-lift privacy and $I(S;X)$. In the following proposition, we generalize this connection to Sibson's and Arimoto's mutual information as well as f-information.

Proposition 1. *If X is an ε-lift private version of S, then*

1) *We have $I_\alpha^{\mathsf{S}}(S;X)\le\frac{\alpha}{\alpha-1}\varepsilon$, $\forall\alpha\in(1,\infty)$. Moreover, the maximal leakage is upper bounded by ε.*
2) *We have $I_\alpha^{\mathsf{A}}(S;X)\le\frac{\alpha}{\alpha-1}\varepsilon$, $\forall\alpha\in(1,\infty)$. Moreover, $\mathsf{P_c}(S|X)\le p_S^*e^\varepsilon$.*
3) *We have $I_f(S;X)\le L(\varepsilon)$ where $L(\varepsilon)\triangleq\sup_{e^{-\varepsilon}\le t\le e^\varepsilon}f(t)$.*

In light of this proposition, the ε-lift privacy guarantee is stronger than those obtained by Arimoto's and Sibson's mutual information and thus ε-lift privacy inherits the operational interpretations described above. In particular, if X is an ε-lift private version of S, then no adversary in possession of observation X can efficiently guess *any function* of S.

III. Lift-Based Privacy Watchdog

We define next the privacy watchdog framework as a simple, yet powerful, privacy technique that acts directly on the sample points. Unlike typical information-theoretic privacy-assuring mechanisms, the privacy watchdog directly assigns a risk score to each sample point from which it determines whether or not a sample can be disclosed unperturbed. Here, we propose to use the lift to generate the risk score for each sample point. We then show how to incorporate such risk scores into designing privacy-assuring mechanisms.

A. Lift-Based Privacy Watchdog

The privacy watchdog framework is described as follows.

Definition 2 (Privacy Watchdog)**.** Given a dataset $\mathcal{D}^n = \{(s_i, x_i)\}_{i=1}^n$ drawn i.i.d. from $P_{S,X}$, the privacy watchdog estimates[1] $i(s,x)$ for each pair $(s,x) \in \mathcal{S} \times \mathcal{X}$ and then decomposes $\mathcal{X}$ into two subsets $\mathcal{X}_\varepsilon \triangleq \{x \in \mathcal{X} |\, |i(x,s)| \leq \varepsilon, \forall s \in \mathcal{S}\}$ and $\mathcal{X}^c \triangleq \mathcal{X} \backslash \mathcal{X}_\varepsilon$. It then then labels all sample points $x_i \in \mathcal{X}_\varepsilon$ as no privacy risk (as they do not significantly change the belief about any of private features) and flags $x_i \in \mathcal{X}_\varepsilon^c$ as potential privacy risks.

Based on the output of the watchdog, we can design a privacy mapping $P_{Y|X}$ that perturbs each sample flagged as posing a privacy risk. A simple such mapping can be constructed as follows: if $x \in \mathcal{X}_\varepsilon$, then it can be perfectly disclosed, i.e., $Y = x$, and if $x \in \mathcal{X}_\varepsilon^c$, then Y can be arbitrary generated on $\mathcal{X}_\varepsilon^c$. The following proposition shows that the output of such mechanism ensures lift privacy with respect to S.

Proposition 2. *Let R_Y be any distribution on a finite set $\mathcal{Y} = \mathcal{X}$ satisfying $R_Y(y) = 0$, $\forall y \in \mathcal{X}_\varepsilon$, and let the $P_{Y|X}$ be given by*

$$P_{Y|X}(y|x) = \begin{cases} 1_{\{x=y\}}, & x \in \mathcal{X}_\varepsilon, \\ R_Y(y), & x \in \mathcal{X}_\varepsilon^c. \end{cases} \tag{9}$$

Then Y is an γ-lift private version of S with

$$\gamma = \max\left\{ \log\left[\frac{1 - e^\varepsilon P_X(\mathcal{X}_\varepsilon) + e^\varepsilon}{P_X(\mathcal{X}_\varepsilon^c)}\right], - \log\left[\frac{1 - e^\varepsilon P_X(\mathcal{X}_\varepsilon)}{P_X(\mathcal{X}_\varepsilon^c)}\right] \right\}.$$

This proposition shows that by disclosing sample points in $\mathcal{X}_\varepsilon$, and regardless of the randomization R_Y used for $\mathcal{X}_\varepsilon^c$, the resulting Y is guaranteed to satisfy the lift privacy constraint. In light of Lemma 1 and Proposition 1, the guarantees provided by the mapping (9) results in upper bounds for the measures of information leakage discussed in Section II-B.

Remark 1. It is important to mention that the mechanism in (9) is perhaps the simplest lift-based privacy-assuring mechanism and is used to merely illustrate the significance of the lift-based watchdog framework. A possibly better mechanism would choose a randomization R_Y that is supported on entire $\mathcal{X}$ (as opposed to $\mathcal{X}_\varepsilon^c$).

[1]We describe in Section IV one method for estimating the log-lift from data.

B. Privacy Funnel

In order to quantify the trade-off between the information leakage incurred by (9) and the utility (information shared between X and Y), we borrow ideas from *privacy funnel* framework [20].

Given a pair of correlated random variables $(S, X) \sim P_{S,X}$, the goal of the *privacy funnel* is to determine a privacy-assuring mapping $P_{Y|X}$ that generates a representation Y of X such that (i) $S - X - Y$ and (ii) a given information leakage metric $L(S;Y)$ (e.g., one of the measures defined in previous section) is minimized while maximizing $I(X;Y)$ (utility preserved). This trade-off can be quantified by the Lagrangian functional

$$\mathsf{F}(P_{S,X}, \lambda) \triangleq \min_{p_{Y|X}} L(S;Y) - \lambda I(X;Y), \tag{10}$$

where larger $\lambda \geq 0$ corresponds to higher utility. Privacy funnel and $\mathsf{F}(P_{S,X}, \lambda)$ are studied in more details in [20]. In general, solving the minimization problem (10) is computationally challenging due to its non-convexity. Although the privacy funnel was derived in closed form expression in simple cases such as binary symmetric channel [5] and Gaussian mixture models [20], it is still unclear how to solve (even algorithmically) the optimization problem in general. There are two algorithms proposed for finding a local minimizer of (10): (i) a greedy algorithm proposed in [20] and (ii) a convex-geometric algorithm devised in [5] which works best when $|\mathcal{S}|$ and $|\mathcal{X}|$ are small. However, these two algorithms are not scalable to high-dimensional settings. To circumvent this issue, algorithms based on neural network architectures have recently been proposed, see e.g., [21] and [22]. The watchdog-based mapping in (9) provides a new direction for designing privacy-assuring mappings with (much) less computational effort, translating the burden to solving the problem of estimating the threshold log-lift from data.

It can be easily verified that for the mechanism given in (9) $I(X;Y)$ is

$$I(X;Y) = H_{\mathcal{X}_\varepsilon} - P_X(\mathcal{X}_\varepsilon^c) \log P_X(\mathcal{X}_\varepsilon^c), \tag{11}$$

where $H_{\mathcal{X}_\varepsilon} \triangleq -\sum_{x \in \mathcal{X}_\varepsilon} p_X(x) \log p_X(x)$ is the entropy of X conditioned on $\mathcal{X}_\varepsilon$. Thus, the utility consists of two parts: the first term is somehow the information preserved by the lift privacy, and the second term relates to the size of the set $\mathcal{X}_\varepsilon^c$. In particular, in low privacy regime, i.e., when $\varepsilon \to \infty$, we have $\mathcal{X}_\varepsilon = \mathcal{X}$ and thus $Y = X$ which leads to the utility $I(X;Y) = H_{\mathcal{X}_\varepsilon} = H(X)$.

By Proposition 1 and 2, there exists a function $\zeta : \mathbb{R}_+ \to \mathbb{R}_+$ of γ such that $L(S;Y) \leq \zeta(\gamma)$ for each measure of information leakage introduced in Section II-B. Thus the objective function of the privacy funnel in (10) is upper bounded as

$$\begin{aligned} \mathsf{F}(P_{S,X}, \lambda) &\leq L(S;Y) - \lambda I(X;Y) \\ &\leq \zeta(\gamma) - \lambda \left[H_{\mathcal{X}_\varepsilon} - P_X(\mathcal{X}_\varepsilon^c) \log P_X(\mathcal{X}_\varepsilon^c)\right], \end{aligned}$$

where γ was defined in Proposition 2. It gives an upper bound of the Lagrangian defined in (10) which can be used to further determine the solution of the privacy funnel.

IV. Privacy Watchdog from Data

We showed in Sections II and III that the ε-lift privacy leads to bounds on various information leakage measures and also can be used to design a privacy watchdog. However, estimating the log-lift from the data is somewhat challenging and has been an active research problem in information theory and computer science communities, see e.g., [7], [23] and [24]. In this section, we propose a log-lift estimator based on Donsker-Varadhan representation [7] and then use it to design a privacy watchdog on the ProPublica's COMPAS recidivism dataset [8].

A. The Log-Lift Estimator

The log-lift estimator takes advantage of the variational representation of KL divergence[2], called Donsker-Varadhan (DV) representation, i.e.

$$I(S;X) = D(P_{S,X}\|P_SP_X) \\ \sup_{g:\mathcal{S}\times\mathcal{X}\to\mathbb{R}} \mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{P_SP_X}[e^{g(S,X)}]. \quad (12)$$

It can be shown that the log-lift is in fact a maximizer of the above optimization problem, i.e., $g^*(s,x) \triangleq \log \frac{P_{S,X}(s,x)}{P_S(s)P_X(x)}$. As such, finding the optimal function $g^*(s,x)$ is equivalent to estimating the log-lift.

In (12), the search space for the function g is unconstrained. A more practical, yet useful assumption, is to restrict the search space to a family $\mathcal{G}(\Theta)$ of bounded functions representable by a neural network with parameters θ in a compact domain $\Theta \subset \mathbb{R}^m$, where m is the number of parameters. The parameters of the neural network can be fit by approximating (e.g., via backpropagation) the solution of the following maximization problem:

$$\hat{g}_n \triangleq \operatorname*{argmax}_{g\in\mathcal{G}(\Theta)} \mathbb{E}_{P_{S_n,X_n}}[g(S,X)] - \log \mathbb{E}_{P_{S_n}P_{X_n}}[e^{g(S,X)}], \quad (13)$$

where P_{S_n,X_n} and $P_{S_n}P_{X_n}$ are the empirical distributions of $P_{S,X}$ and P_SP_X respectively. The estimator in (13) belongs to a broader class of *extremum estimators* [25] which consists of estimators of the form $\hat{a} = \operatorname*{argmax}_{a\in\mathcal{A}} \Lambda_n(a)$, where $\Lambda_n(a)$ is an objective function and $\mathcal{A}$ is a parameter space. The consistency of such estimators is guaranteed according to the following lemma.

Lemma 2 (Consistency of Extremum Estimators [25])**.** *Given the extremum estimator $\hat{a} = \operatorname*{argmax}_{a\in\mathcal{A}} \Lambda_n(a)$, if (i) $\mathcal{A}$ is compact; (ii) there exists a limiting function $\Lambda_0(a)$ such that $\Lambda_n(a)$ converges to $\Lambda_0(a)$ over $\mathcal{A}$ in probability; (iii) $\Lambda_0(a)$ is continuous and has unique maximum at $a = a_0$, then $\hat{a}$ is a consistent estimator of a_0.*

[2]In fact, the variational representation of f-divergences, $D_f(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))]$, where $f^*(y) \triangleq \sup_{x\in\mathbb{R}}[xy - f(x)]$ is the Fenchel conjugate of f, can be used in the log-lift estimator.

Using this lemma, together with the universal approximation theorem of neural networks [26], we show in the following proposition that the log-lift estimator in (13) is consistent.

Proposition 3. *Assume $\mathbb{E}_{P_{S,X}}[g(S,X)]$ and $\mathbb{E}_{P_SP_X}[e^{g(S,X)}]$ are finite. The log-lift estimator*

$$\hat{g}_n = \operatorname*{argmax}_{g\in\mathcal{G}(\Theta)} \mathbb{E}_{P_{S_n,X_n}}[g(S,X)] - \log \mathbb{E}_{P_{S_n}P_{X_n}}[e^{g(S,X)}]$$

is consistent, i.e., for any $\eta > 0$, there exist $N > 0$ such that for all $n > N$,

$$Pr\{|\hat{g}_n(s,x) - g^*(s,x)| \leq \eta\} = 1, \ \forall s \in \mathcal{S}, x \in \mathcal{X}. \quad (14)$$

With the log-lift estimator (13) at hand, the set $\mathcal{X}_\epsilon$ can be determined and hence the proposed watchdog-based privacy mechanism in (9) can be implemented on real-world data, as illustrated next.

As a final remark, we note that the approach outlined above seeks to estimate the value $g^*(s,x)$ across the entire domain $\mathcal{S}\times\mathcal{X}$, whereas the watchdog framework requires only a threshold version of this function. We will explore the gain (in terms of sample complexity) of this simplification in a future work.

B. Numerical Experiments

In order to validate our privacy watchdog mechanism, we implement it on the ProPublica's COMPAS recidivism racial bias dataset [8]. This dataset contains the criminal history and demographic makeup of prisoners in Brower County, Florida from 2013-2014. We set race as the private attribute S, and restrict the dataset to entries with race marked as African American ($S = 0$) and Caucasian ($S = 1$). Moreover, we select *gender*, *age*, *number of prior crimes*, *length of custody* and *likelihood of recidivism* to be the observation X. We pre-process the dataset by dropping missing/incomplete records, convert categorical variables by one-hot encoding, and finally take 5278 samples with $70\% - 30\%$ training-test split. For details about experimental settings, see [17].

In Fig. 1, we demonstrate the estimate of log-lifts $i(S=0,x)$ and $i(S=1,x)$ for all samples, and the boundary of $\mathcal{X}_\varepsilon$ with $\varepsilon = 0.85$. Interestingly, based on the value of the lift, we may be able to provide some interpretation on why a given sample may or may not compromise privacy if released. For instance, in Table I, we select samples (green dots in Fig. 1) with high $i(S=0,x)$ and low $i(S=1,x)$. Observe that young males with a high prior count and high recidivism risk score are flagged as leaking significant information about the private attribute. For other examples of extreme samples, see [17].

In Fig. 2, using the privacy watchdog-based privacy mechanism, we show the trade-off between the utility $I(X;Y)$ (11) and the bounds γ (Propoisiton 2) on information leakage (measured by any metric in Section II-B). When ε is around 0.3, the privacy watchdog chooses to release samples that give best utility and little information leakage. As ε becomes larger, the utility remains unchanged, but the information leakage increases. This kind of numerical analysis could be used to tune the value of ε in the privacy watchdog.

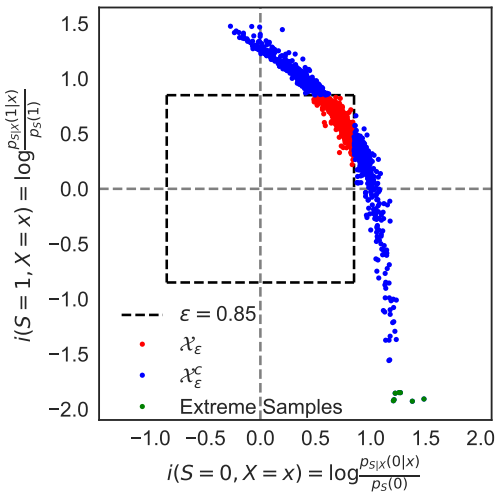

Fig. 1: Estimation of lifts for each samples in the COMPAS dataset. The dashed square contains samples in $\mathcal{X}_\varepsilon$ with $\varepsilon = 0.85$. Green dots show samples with high privacy risk.

Gender	Race	Age	Prior Counts	Length of Stay	Recidivism
M	AA	21	1	1	9
M	AA	33	5	0	5
M	C	43	0	2	1
M	AA	27	13	0	10
M	AA	59	8	8	8
M	AA	29	5	5	7
M	AA	25	1	0	3

TABLE I: Extreme samples in the COMPAS dataset with high $i(s=0,x)$ and low $i(s=1,x)$ in Fig. 1. M: Male, AA: African American, C: Caucasian.

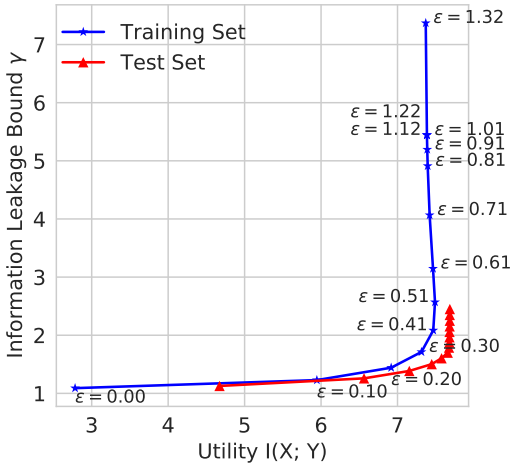

Fig. 2: The trade-off between the utility and information leakage (in privacy funnel with $L(S;Y) = I(S;Y)$ in Section III) on training and test set in COMPAS. Different ε gives the entire approximation of the privacy funnel. The privacy watchdog reaches a best privacy-utility operation point when ε is around 0.3.

REFERENCES

[1] S. Tufféry, *Data mining and statistics for decision making*. Wiley Chichester, 2011, vol. 2.

[2] T. S. Han, *Information-spectrum methods in information theory*. Tokyo, Japan: Baifukan, 1998.

[3] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Estimation efficiency under privacy constraints," *IEEE Trans. Inf. Theory*, pp. 1–1, 2018.

[4] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *arXiv preprint arXiv:1807.07878*, 2018.

[5] H. Hsu, S. Asoodeh, S. Salamatian, and F. P. Calmon, "Generalizing bottleneck problems," in *Proc. of IEEE Int. Symp. Inf. Theory (ISIT)*, 2018.

[6] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. of IEEE Foundations of Computer Science (FOCS)*, 2013.

[7] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[8] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias. propublica, may 23, 2016," 2016.

[9] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[10] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2003, pp. 211–222.

[11] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 2012, pp. 1401–1408.

[12] A. Makhdoumi and N. Fawaz, "Privacy-utility tradeoff under statistical uncertainty." in *Allerton*, 2013, pp. 1627–1634.

[13] B. Rassouli and D. Gunduz, "On perfect privacy and maximal correlation," *arXiv preprint arXiv:1712.08500*, 2017.

[14] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Privacy of dependent users against statistical matching," *arXiv preprint arXiv:1806.11108*, 2018.

[15] M. Diaz, H. Wang, F. P. Calmon, and L. Sankar, "On the robustness of information-theoretic privacy measures and mechanisms," *arXiv preprint arXiv:1811.06057*, 2018.

[16] J. Liao, O. Kosut, L. Sankar, and F. P. Calmon, "A tunable measure for information leakage," *arXiv preprint arXiv:1806.03332*, 2018.

[17] H. Hsu, S. Asoodeh, and F. P. Calmon. Information-theoretic privacy watchdog - extended version. [Online]. Available: https://github.com/HsiangHsu/ISIT-19-Extended-Version

[18] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[19] S. Verdú, "α-mutual information," in *Information Theory and Applications Workshop (ITA), 2015*. IEEE, 2015, pp. 1–6.

[20] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *Proc. of IEEE Information Theory Workshop (ITW)*, 2014.

[21] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Generative adversarial privacy," *arXiv preprint arXiv:1807.05306*, 2018.

[22] A. Tripathy, Y. Wang, and P. Ishwar, "Privacy-preserving adversarial networks," *arXiv preprint arXiv:1712.07008*, 2017.

[23] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[24] I. Belghazi, S. Rajeswar, A. Baratin, R. D. Hjelm, and A. Courville, "Mine: mutual information neural estimation," *arXiv preprint arXiv:1801.04062*, 2018.

[25] T. Amemiya, "Asymptotic properties of extremum estimators," *Advanced econometrics, Harvard university press*, 1985.

[26] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.