Information-Theoretic Privacy Watchdogs

Hsiang Hsu*, Shahab Asoodeh†, and Flavio P. Calmon*
*Harvard University, {hsianghsu, fcalmon}@g.harvard.edu,
†University of Chicago, shahab@uchicago.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 which is a variation of 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. 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 (e.g., erased, randomized) 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 the 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 prove properties of lift as a privacy metric, and show that, by bounding (1), we also bound several information-theoretic privacy measures, including those based on Arimoto's and Sibson's mutual information [3], f-divergences [4], and local differential privacy [5]. 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 [6] (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 [7].

The design of privacy mechanisms is an imminent topic in computer science [8], data mining [9], and information theory [3], [10]–[14] communities. Within the latter, there has been significant effort to characterize fundamental trade-offs between privacy and utility (e.g., [3], [10]), as well as produce privacy metrics with operational significance (e.g., [14]–[16]). We also note that variations of information density were mentioned in [9]–[11] 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.

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 distribution 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 denote $1_{\{\cdot\}}$ to be 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 [17] 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 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 [5], maximal leakage [14], α-leakage [15], and f-information. We note that variations of ε-lift privacy have appeared in the literature under different guises (e.g., [9, Defn. 1] and [10, 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 [10])**.** 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 [5], i.e.*

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

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

Proof. See Appendix A. □

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. We explore further connections between ε-lift privacy and information-theoretic measures for leakage next.

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 [3], [18]. Arimoto's mutual information of order $\alpha\in(1,\infty)$ is given by [18]

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

When $\alpha\to\infty$, $I_\infty^{\mathsf{A}}(S;X)$ characterizes the ability of an adversary to correctly guess S given X [16]. In particular, let the *probability of correctly guessing* S given X be defined as

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

It can be readily verified that $I_\infty^{\mathsf{A}}(S;X)=\log\frac{\mathsf{P_c}(S|X)}{p_S^*}$, where $p_S^*\triangleq\max_{s\in\mathcal{S}}P_S(s)$.

Another operational measure of information leakage recently proposed is Sibson's mutual information [18] of order $\alpha\in(0,1)\cup(1,\infty)$ between S and X, which is given by

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

Here, $D_\alpha(P\|Q)\triangleq\frac{1}{\alpha-1}\log\left(\sum_x P(x)^\alpha Q(x)^{1-\alpha}\right)$ is the Rényi divergence [18]. One can also 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 [14]. More precisely, maximal leakage is equal to the logarithm of the multiplicative gain in guessing *any function* of S given an ovservation 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 $U-S-X$. It is worth mentioning that $I_\alpha^{\mathsf{S}}(S;X)$ and $I_\alpha^{\mathsf{A}}(S;X)$ tend to $I(S;X)$ when $\alpha\to1$.

As shown in Lemma 1, ε-lift privacy controls the mutual information $I(S;X)$. In the following proposition, we further illustrate that ε-lift privacy controls Sibson's and Arimoto's mutual information and f-information as well.

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

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

Proof. See Appendix B. □

The above proposition indicates that an ε-lift privacy guarantee is stronger than those obtained by Arimoto's and Sibson's mutual information and also f-information. Thus, ε-lift privacy inherits the operational interpretation of the well-known privacy measures listed 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 the private feature S.

III. Lift-Based Privacy Watchdog

We define next the *privacy watchdog* 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. Here, we propose to use the lift to generate the privacy score for each sample point. We then show how a privacy mechanism can be designed based on the watchdog.

A. Lift-Based Privacy Watchdog

The lift-based privacy watchdog framework is defined next.

Definition 2. Given a dataset $\mathcal{D}^n = \{(s_i, x_i)\}_{i=1}^n$ drawn i.i.d. from $P_{S,X}$ and the log-lift[1] $i(s_i, x_i)$, the privacy watchdog 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 $\mathcal{X}^c \triangleq \mathcal{X} \backslash \mathcal{X}_\varepsilon$. If $x_i \in \mathcal{X}_\varepsilon$, then the sample x_i can be disclosed as it does not significantly change the belief about any of private features s_i for all i. If $x_i \in \mathcal{X}_\varepsilon^c$, then x_i should be discarded/modified in order to ensure privacy.

The privacy watchdog mechanism described above flags the sample points $x_i \in \mathcal{X}_\varepsilon$ whose privacy scores are below a threshold ε. 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. Perhaps the simplest such mapping is one that generates Y in such a way that $Y = X$ conditioned on the event $X \in \mathcal{X}_\varepsilon$, and draws Y independently from X otherwise. The resulting Y then ensures lift privacy with respect to S. This statement is formalized in the following proposition.

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 privacy watchdog $P_{Y|X}$ 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

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

Proof. See Appendix C. □

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 measure of lift privacy. In light of Lemma 1 and Proposition 1, the guarantees provided by the mapping (9) results in upper bound for measures of information leakage discussed in Section II-B. The mapping (9) is just one example of how the privacy watchdog can be applied to design privacy mechanisms.

[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 [19].

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, denoted by $L(S;Y)$, 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, and $L(S;Y)$ can be any measure of information leakage introduced in Section II-B. Privacy funnel and $\mathsf{F}(P_{S,X}, \lambda)$ are studied in more details in [4], [19].

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 [4] and Gaussian mixture models [19], 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 [19] and (ii) a convex-geometric algorithm devised in [4] 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 [20], [21]. The privacy 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 (thresholded) log-lift from data.

For the mapping given (9), it can be verified that the utility $I(X;Y)$ is given by

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

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 $H_{\mathcal{X}_\varepsilon}$ is the information preserved by the lift privacy, and the second term relates to the size of the set $\mathcal{X}_\varepsilon^c$.

By Proposition 1 and 2, $L(S;Y) \leq \gamma$ for all measures of information leakage introduced in Section II-B, and the objective of the privacy funnel in (10) is upper bounded as

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

where γ was defined in (10). 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 $I(X;Y) = H_{\mathcal{X}_\varepsilon} = H(X)$.

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 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 [6], [22], [23]. In this section, we propose a log-lift estimator based on Donsker-Varadhan representation [6] and then use it to design the privacy watchdog on the ProPublica's COMPAS recidivism dataset [7].

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 unlimited. 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* [24] 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 [24])**.** *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 $\Lambda_0(a)$ in probability over $\mathcal{A}$; (iii) $\Lambda_0(a)$ is continuous and has unique maximum at $a = a_0$, then $\hat{a}$ is a consistent estimator of a_0.*

Using this lemma, together with the universal approximation theorem of neural networks [25], we show in the following proposition that the log-lift estimator in (13) is consistent.

[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.

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 $n > N$,

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

Proof. See Appendix D. □

With the log-lift estimator (13) at hand, the set $\mathcal{X}_\epsilon$ can be determined and hence the proposed privacy 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 in Definition 2 requires only a thresholded version of this function. We will explore the gain (in terms of sample complexity) of this simplification in 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 [7]. 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 Appendix E.

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 Appendix F.

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.

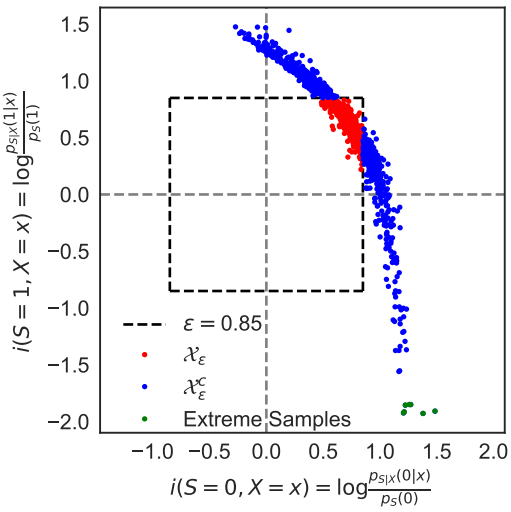

Fig. 1: Estimation of lifts for each samples in the COMPAS dataset. The dashed square contains samples in $\mathcal{X}_\varepsilon$ with $\varepsilon = 0.85$. Green dots show samples with high privacy risk.

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.

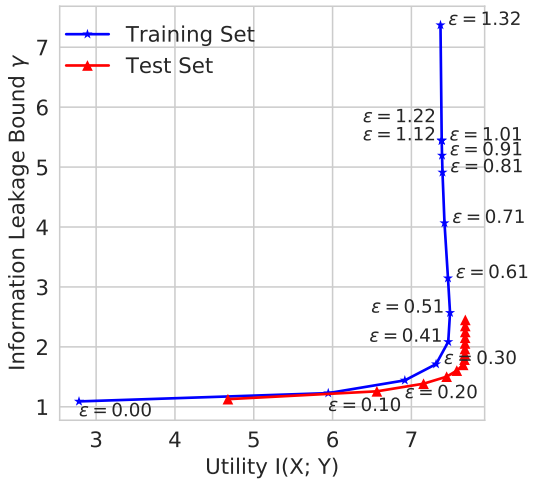

Fig. 2: The trade-off between the utility and information leakage using the privacy watchdog on training and test set in COMPAS. Different ε gives the entire approximation of the privacy funnel. The privacy watchdog reaches a best privacy-utility operation point when ε is around 0.3.

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] 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.

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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

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

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

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

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

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

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

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

[19] 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.

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

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

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

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

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

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

[26] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[27] S. A. van de Geer and S. van de Geer, *Empirical Processes in M-estimation*. Cambridge university press, 2000, vol. 6.

APPENDIX

A. Proof of Lemma 1

(1) holds since $\log \frac{P_{S|X}(s|x)}{P_S(s)} = \log \frac{P_{S,X}(s,x)}{P_S(s)P_X(x)} = \log \frac{P_{X|S}(x|s)}{P_X(x)}$ by Bayes's rules.

(2) holds since for any $x, x' \in \mathcal{X}$ and assuming $P_{S,X}(s,x) > 0$ and $P_{S,X}(s,x') > 0$, by definition 1,

$$\begin{aligned} & \left|\log \frac{P_{S|X}(s|x)}{P_{S|X}(s|x')}\right| \\ = & \left|\log \frac{P_{S|X}(s|x)}{P_S(s)} - \log \frac{P_{S|X}(s|x')}{P_S(s)}\right| \quad (15) \\ \leq & 2\varepsilon, \quad (16) \end{aligned}$$

which is equivalent to the definition of 2ε-local differential private in (5).

(3) holds since

$$\sum_{(s,x)\in\mathcal{S}\times\mathcal{X}} P_{S,X}(s,x) \log \frac{P_{S|X}(s|x)}{P_S(s)} \leq \varepsilon. \quad (17)$$

B. Proof of Proposition 1

For (1), by the assumption, we know $P_{S,X}(s,x)^\alpha \leq e^{\alpha\varepsilon} P_S(s)^\alpha P_X(x)^\alpha$, and

$$\begin{aligned} & D_\alpha(P_{S,X}\|P_S Q_X) \\ = & \frac{1}{\alpha-1}\log\left(\sum_{s,x} \frac{P_{S,X}(s,x)^\alpha}{P_S(s)^{\alpha-1}Q_X(x)^{\alpha-1}}\right) \quad (18) \\ \leq & \frac{1}{\alpha-1}\log\left(e^{\alpha\varepsilon}\sum_{s,x} \frac{P_S(s)^\alpha P_X(x)^\alpha}{P_S(s)^{\alpha-1}Q_X(x)^{\alpha-1}}\right) \quad (19) \\ = & \frac{1}{\alpha-1}\alpha\varepsilon + \frac{1}{\alpha-1}\log\left(\sum_x \frac{P_X(x)^\alpha}{Q_X(x)^{\alpha-1}}\right) \quad (20) \\ = & \frac{\alpha}{\alpha-1}\varepsilon + D_\alpha(P_X\|Q_X). \quad (21) \end{aligned}$$

Therefore, since $\inf_{Q_X} D_\alpha(P_X\|Q_X) = 0$ when $P_X = Q_X$, we have

$$\begin{aligned} I_\alpha^{\mathsf{S}}(S;X) & \leq \frac{\alpha}{\alpha-1}\varepsilon + \inf_{Q_X} D_\alpha(P_X\|Q_X) \quad (22) \\ & = \frac{\alpha}{\alpha-1}\varepsilon. \quad (23) \end{aligned}$$

By taking α to infinity, we have the maximal leakage $I_\infty^S(S;X) \leq \varepsilon$.

For (2), by the assumption, we have

$$P_{S|X}(s|X)^\alpha \leq e^{\alpha\varepsilon} P_S(s)^\alpha \Rightarrow \|P_{S|X}(\cdot|X)\|_\alpha \leq e^\varepsilon \|P_S\|_\alpha. \quad (24)$$

Therefore, we have

$$\begin{aligned} I_\alpha^A(S;X) & = \frac{\alpha}{\alpha-1}\log\frac{\sum_x \|P_{S|X}(\cdot|x)\|_\alpha P_X(x)}{\|P_S\|_\alpha} \quad (25) \\ & \leq \frac{\alpha}{\alpha-1}\log\frac{\sum_x e^\varepsilon \|P_S\|_\alpha P_X(x)}{\|P_S\|_\alpha} \quad (26) \\ & = \frac{\alpha}{\alpha-1}\varepsilon. \quad (27) \end{aligned}$$

By taking α to infinity, it follows that $I_\infty^A(S;X) \leq \varepsilon$ and hence $\mathsf{P_c}(S|X) \leq p_S^* \exp(\varepsilon)$..

For (3), by the assumption and Jensen's inequality [26]

$$\begin{aligned} I_f(S;X) & \triangleq \mathbb{E}_{S\sim P_S, X\sim P_X}[f(l(S,X))] \quad (28) \\ & \geq f(\mathbb{E}_{S\sim P_S, X\sim P_X}[l(S,X)]) \quad (29) \\ & \geq \min_{e^{-\varepsilon}\leq t\leq e^\varepsilon} f(t), \quad (30) \end{aligned}$$

and $I_f(S;X) \leq \max_{e^{-\varepsilon}\leq t\leq e^\varepsilon} f(t)$. In other words, $I_f(S;X) \in [\min_{e^{-\varepsilon}\leq t\leq e^\varepsilon} f(t), \max_{e^{-\varepsilon}\leq t\leq e^\varepsilon} f(t)]$. If the convex function f is specified, the maximization/ minimization in the last inequality can be easily computed.

C. Proof of Proposition 2

Firs note that into two parts

$$\begin{aligned} P_{S|Y}(s|y) & = \sum_{x\in\mathcal{X}_\varepsilon} P_{S|X}(s|x)P_{X|Y}(x|y) \\ & \quad + \sum_{x\notin\mathcal{X}_\varepsilon} P_{S|X}(s|x)P_{X|Y}(x|y). \end{aligned}$$

Also note that by construction we have

$$P_{X|Y}(x|y) = \begin{cases} 1_{\{x=y\}}, & x,y\in\mathcal{X}_\varepsilon, \\ 0, & x\in\mathcal{X}_\varepsilon \text{ and } y\notin\mathcal{X}_\varepsilon, \\ \frac{P_X(x)}{P_X(\mathcal{X}_\varepsilon^c)}, & x,y\notin\mathcal{X}_\varepsilon, \\ 0, & x\notin\mathcal{X}_\varepsilon \text{ and } y\in\mathcal{X}_\varepsilon. \end{cases} \quad (31)$$

Assuming the ε-lift constraint, we can write

$$P_{S|Y}(s|y) \leq e^\varepsilon P_S(s)\sum_{x\in\mathcal{X}_\varepsilon} P_{X|Y}(x|y) + \frac{P_S(s)}{P_X(\mathcal{X}_\varepsilon^c)}\sum_{x\notin\mathcal{X}_\varepsilon} P_{X|S}(x|s),$$

and thus

$$\frac{P_{S|Y}(s|y)}{P_S(s)} \leq \left[e^\varepsilon \sum_{x\in\mathcal{X}_\varepsilon} P_{X|Y}(x|y) + \frac{1}{P_X(\mathcal{X}_\varepsilon^c)}\sum_{x\notin\mathcal{X}_\varepsilon} P_{X|S}(x|s)\right].$$

As a very crude bound, we obtain

$$\log\frac{P_{S|Y}(s|y)}{P_S(s)} \leq \log\left[e^\varepsilon + \frac{1}{P_X(\mathcal{X}_\varepsilon^c)}\right]. \quad (32)$$

Similarly, for the other direction,

$$\begin{aligned} P_{S|Y}(s|y) & \geq \frac{P_S(s)}{P_X(\mathcal{X}_\varepsilon^c)}\sum_{x\notin\mathcal{X}_\varepsilon} P_{X|S}(x|s) \\ & = \frac{P_S(s)}{P_X(\mathcal{X}_\varepsilon^c)}\left[1 - \sum_{x\in\mathcal{X}_\varepsilon} P_{X|S}(x|s)\right] \\ & \geq \frac{P_S(s)}{P_X(\mathcal{X}_\varepsilon^c)}(1 - e^\varepsilon P_X(\mathcal{X}_\varepsilon)). \quad (33) \end{aligned}$$

Combining (32) and (33), we have

$$\log\left[\frac{1-e^\varepsilon P_X(\mathcal{X}_\varepsilon)}{P_X(\mathcal{X}_\varepsilon^c)}\right] \leq \log\frac{P_{S|Y}(s|y)}{P_S(s)} \leq \log\left[\frac{1-e^\varepsilon P_X(\mathcal{X}_\varepsilon)+e^\varepsilon}{P_X(\mathcal{X}_\varepsilon^c)}\right].$$

Consequently, mechanism $P_{Y|X}$ described in (9) satisfies γ-lift privacy with

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

D. Proof of Proposition 3

First, by triangle inequality, for all $s \in \mathcal{S}$ and $x \in \mathcal{X}$

$$\begin{aligned} & |\hat{g}_n(s,x) - g^*(s,x)| \\ \leq \; & |\hat{g}_\theta(s,x) - g^*(s,x)| + |\hat{g}_n(s,x) - \hat{g}_\theta(s,x)|, \end{aligned} \quad (34)$$

where $\hat{g}_\theta(s,x)$ is defined as

$$\hat{g}_\theta = \underset{g \in \mathcal{G}(\Theta)}{\operatorname{argmax}} \mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{P_S P_X}[e^{g(S,X)}]. \quad (35)$$

Since $\mathbb{E}_{P_{S,X}}[g(S,X)]$ is finite, by the universal approximation theorem [25], there exists a set of parameters θ such that with probability one

$$|\hat{g}_\theta(s,x) - g^*(s,x)| \leq \frac{\eta}{2}, \; \forall s \in \mathcal{S}, x \in \mathcal{X}. \quad (36)$$

Moreover, let the objective function of the extremum estimator be

$$\Lambda(g)_n \triangleq \mathbb{E}_{P_{S_n,X_n}}[g(S,X)] - \log \mathbb{E}_{P_{S_n} P_{X_n}}[e^{g(S,X)}]. \quad (37)$$

First, since Θ is compact and the mappings represented by neural networks are continuous, the images $\mathcal{G}(\Theta)$ is also compact.

Second, by triangular inequality, for $g \in \mathcal{G}(\Theta)$, we have

$$\begin{aligned} & |\Lambda(g)_n - (\mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{p_S p_X}[e^{g(S,X)}])| \quad (38) \\ \leq \; & \sup_{g \in \mathcal{G}(\Theta)} |\mathbb{E}_{P_{S,X}}[g(S,X)] - \mathbb{E}_{P_{S_n,X_n}}[g(S,X)]| \\ + \; & \sup_{g \in \mathcal{G}(\Theta)} |\log \mathbb{E}_{P_S P_X}[g(S,X)] - \log \mathbb{E}_{P_{S_n} P_{X_n}}[g(S,X)]|. \end{aligned}$$

Since the function g is given by a neural network, it can be uniformly bounded by some constant M, i.e. $|g| \leq M$ for all θ, s and x. Since logarithm is Lipschitz continuous with constant e^M in the interval $[e^{-M}, e^M]$, we have

$$\begin{aligned} & |\log \mathbb{E}_{P_S P_X}[g(S,X)] - \log \mathbb{E}_{P_{S_n} P_{X_n}}[g(S,X)]| \\ \leq \; & e^M |\mathbb{E}_{P_S P_X}[g(S,X)] - \mathbb{E}_{P_{S_n} P_{X_n}}[g(S,X)]|. \end{aligned} \quad (39)$$

Moreover, since $\mathcal{G}$ is compact and g is continuous, the functions g and e^g satisfy the uniform law of large numbers [27]. Thus, Given $\eta > 0$, there exists an integer N such that for all $n \geq N$ and with probability one,

$$\sup_{g \in \mathcal{G}(\Theta)} |\mathbb{E}_{P_{S,X}}[g(S,X)] - \mathbb{E}_{P_{S_n,X_n}}[g(S,X)]| \leq \frac{\eta}{2}, \text{ and} \quad (40)$$

$$\begin{aligned} & \sup_{g \in \mathcal{G}(\Theta)} |\log \mathbb{E}_{P_S P_X}[g(S,X)] - \log \mathbb{E}_{P_{S_n} P_{X_n}}[g(S,X)]| \\ \leq \; & \frac{\eta}{2} e^{-M}. \end{aligned} \quad (41)$$

Summarizing (38)-(41), we have with probability one

$$|\Lambda(g)_n - (\mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{P_S P_X}[e^{g(S,X)}])| \leq \eta. \quad (42)$$

In other words, there exists a limiting function $\Lambda(g)_0 = \mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{P_S P_X}[e^{g(S,X)}]$ such that $\Lambda(g)_n$ converges to $\Lambda(g)_0$ in probability.

Third, since $\Lambda(g)_0 = \mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{P_S P_X}[e^{g(S,X)}]$ consists of linear combinations (expectations) and continuous mappings (logarithm and exponential) of the continuous function g, $\Lambda(g)_0$ is continuous. Moreover, $\Lambda(g)_0$ has a unique optimizer $g_0 = g^*$. By Lemma 2, we know that with probability one,

$$|\hat{g}_n(s,x) - \hat{g}_\theta(s,x)| \leq \frac{\eta}{2}. \quad (43)$$

Combining (36) and (43), the desired result follows.

E. Experimental Settings

We construct two neural networks: the first is used to estimate the conditional probability $p_{S|X}$, and the second is used as the log-lift estimator (13). For both neural networks, to avoid digressing, we adopt vanilla feed-forward architectures with two fully-connected layers and a readout layer. In the first neural network, all layers have 128 neurons and `ReLU` activation functions, and in the second neural network, all layers have 128 neurons and `tanh` activation functions. In training both neural networks, we use `AdagradOptimizer` with learning rate 0.001, and run 3000 times with the whole training set.

F. Extreme Samples

Gender	Race	Age	Prior Counts	Length of Stay	Recidivism
F	AA	39	0	57	5
M	C	41	5	1	1
F	C	31	2	1	1
M	C	50	3	11	2
M	C	27	2	1	2

TABLE II: Samples with low $i(s=0,x)$ and high $i(s=1,x)$ in Fig. 1. F: Female, M: Male, AA: African American, C: Caucasian.

Gender	Race	Age	Prior Counts	Length of Stay	Recidivism
M	C	29	0	0	2
M	C	45	0	2	2
M	C	25	0	1	2
M	C	42	1	1	2
M	AA	25	3	7	6
F	AA	37	3	28	7
M	AA	19	2	1	10
M	C	56	1	1	1

TABLE III: Samples with $i(s=0,x) \approx 0$ and high $i(s=1,x)$ in Fig. 1. F: Female, M: Male, AA: African American, C: Caucasian.

Gender	Race	Age	Prior Counts	Length of Stay	Recidivism
M	C	25	1	42	3
M	AA	53	11	110	6
M	AA	23	4	31	10
M	C	48	0	68	8

TABLE IV: Samples with $i(s=1,x) \approx 0$ and high $i(s=0,x)$ in Fig. 1. F: Female, M: Male, AA: African American, C: Caucasian.