Feature

Hsiang Hsu | Harvard University, Cambridge, MA 02138 USA | Email: hsianghsu@g.harvard.edu
Natalia Martinez | Duke University, Durham, NC 27708 USA | Email: natalialmg91@gmail.com
Martin Bertran | Duke University, Durham, NC 27708 USA | Email: martin.a.bertran@gmail.com
Guillermo Sapiro, Fellow, IEEE | Duke University, Durham, NC 27708 USA and also Apple, Inc., Durham, NC 27713 USA | Email: guillermo.sapiro@duke.edu
Flavio P. Calmon | Harvard University, Cambridge, MA 02138 USA | Email: flavio@seas.harvard.edu

A Survey on Statistical, Information, and Estimation—Theoretic Views on Privacy

Abstract—Privacy has become an emerging challenge in both information theory and computer science due to massive (centralized) collection of user data. In this article, we overview privacy-preserving mechanisms and metrics from the lenses of information theory, and unify different privacy metrics, including *f*-divergences, Rényi divergences, and differential privacy (DP), in terms of the probability likelihood ratio (and its logarithm). We review recent progress on the design of privacy-preserving mechanisms according to the privacy metrics in computer science, where DP is the standard privacy notion which controls the output shift given small input perturbation, and information theory, where the privacy is guaranteed by minimizing information leakage. In particular, for DP, we include its important variants (e.g., Rényi DP, Pufferfish privacy) and properties, discuss its connections with information-theoretic quantities, and provide the operational interpretations of its additive noise mechanisms. For information-theoretic privacy, we cover notable frameworks, including the privacy funnel, originated from rate-distortion theory and information bottleneck, to privacy guarantee against statistical inference/guessing, and information obfuscation on samples and features. Finally, we discuss the implementations of these privacy-preserving mechanisms in current data-driven machine learning scenarios, including deep learning, information obfuscation, federated learning, and dataset sharing.

Introduction

In 1943, during the peak of World War II, Alan Turing visited Bell Labs to examine the X-system—a secret voice scrambler for private telephone communications between the authorities in London and Washington [1]. At Bell Labs, Turing met with a young Claude Shannon, who was also working on cryptography (among other things). In an interview with R. Price in 1982 [2], Shannon recalled that Turing and him would frequently have lunch together but would avoid discussing cryptography. They preferred topics that were not classified, such as computing machines and the human brain. According to Shannon, he described to the British computer scientist the seminal ideas of what would become information theory. "He was interested." Shannon told Price, "He didn't believe they [my ideas] were in the right direction. I got a fair amount of negative feedback almost."

Since the inception of both fields, privacy has been squarely within the purview of information theory and computer science. As foreshadowed by the interaction between Shannon and Turing, the two communities have developed their own approach to the problem of preventing unauthorized extraction of information from disclosed data,[1] with each community considering their own models and often applying (very) distinct mathematical techniques.

Three decades after Shannon and Turing met, the contrast between the information theory and the computer science approaches to privacy was again made evident in two seminal papers published a year apart. The first paper was authored by Aaron Wyner in October 1975—while also working at Bell Labs—and introduced the wire-tap channel [4]. Wyner considered a model where data are transmitted over a discrete, memoryless channel (DMC) subject to a wire-tap at the receiver. The wire-tap is modeled as a second DMC whose output is observed by a passive eavesdropper. Wyner derived the maximum "error-free" communication rate between the legitimate

2692-4080 © 2021 IEEE
Digital Object Identifier 10.1109/MBITS.2021.3108124
Date of publication 27 Aug. 2021; date of current version 4 Feb. 2022.

[1] The definition of privacy adopted in our manuscript is "the problem of preventing unauthorized extraction of information from communications over an insecure channel." This definition is due to Diffie and Hellman [3].

transmitter and receiver while ensuring perfect secrecy against the wire-tapper—a result later strengthened and generalized in several directions (see [5]). He proved that perfect privacy can be achieved by designing codes that take advantage of the noisier channel observed by the eavesdropper, guaranteeing secrecy *even if the eavesdropper is computationally unbounded*. The goal of ensuring privacy without making assumptions on an adversary's computational capabilities—often referred to as *information-theoretic secrecy*—would become a centerpiece of the security research developed within the information theory community.

Almost exactly a year later, in November 1976, Diffie and Hellman published the ground breaking paper "New Directions in Applied Cryptography" in the IEEE TRANSACTIONS ON INFORMATION THEORY[2] [3]. This paper described the foundations of public key cryptography, as well as public key distribution systems and verifiable digital signatures. Instead of aiming for information-theoretic secrecy, the cryptographic approach outlined by Diffie and Hellman ensures security against a computationally bounded adversary, relying on the computational difficulty in discovering private information without additional knowledge (e.g., of a private key). This assumption makes public key cryptographic systems significantly easier to engineer and deploy, not requiring an *a priori* secret key agreement between communicating parties.

Since the publication of these two papers, public key cryptography has achieved widespread deployment, fundamentally impacting banking, healthcare, public services, and beyond. Algorithms resulting from the computer science approach to cryptography—where an adversary is assumed to be computationally bounded—are used billions of times per day, in applications that range from digital rights management to cryptocurrency. In fact, if you are reading this article on your computer, you likely have used public key cryptography to authenticate the website used to download this article. In contrast, information-theoretic approaches to secrecy have seen far less success in practice: perfect secrecy against a computationally unbounded adversary requires rigid assumptions, leading to elegant mathematical models, but security schemes that are often unwieldy and difficult to engineer in practice.

Today, the information theory and computer science approaches to privacy intersect yet again. The ever-increasing harvesting of individual-level data (i.e., "Big Data") has led to new challenges and opportunities for both fields. On the one hand, society has enjoyed significant utility (in the usual economic sense) from massive data collection and processing, ranging from large datasets for research to personalized services and innovative business models. On the other hand, widespread data collection poses new privacy threats: for example, social media posts can lead to undesirable political targeting [6], parameters and outputs of a machine learning model may reveal sensitive information about the dataset used during training [7], and public databases may be deanonymized with only a few queries [8].

Several of the privacy risks faced in the current data-driven economy cannot be easily addressed via cryptographic techniques. The key challenge is that an adversarial party observes (by design) disclosed data in the clear. For example, a statistician who queries a database that contains sensitive information will receive a numerical value as an output (e.g., number of individuals with income below a threshold in a given population)—simply encrypting the output would provide no utility. By performing multiple queries, the statistician could potentially infer private information. This is exactly the challenge faced by the U.S. Census when disclosing statistics about the American population while being legally bound to preserve individual privacy and anonymity [9]. A similar challenge arises when companies train machine learning models with user data: the consumer receives utility from disclosing personal information (e.g., facial recognition in images taken on a smartphone), yet is subject to privacy risks via undesired inferences (e.g., an adversary may learn individual-level information by probing the model).

In contrast to cryptography and information-theoretic security, the privacy desideratum in many data-driven applications is not to ensure zero information leakage, be it against a computationally bounded or information-theoretic adversary. Instead, the main goal is to ensure a provable level of privacy—albeit not necessarily perfect—while achieving a target level of utility. The privacy threat model is given by an inferring adversary who observes disclosed data and attempts to accurately estimate sensitive information (e.g., a user's political preference or if a target individual is in a database).

As we shall shortly see, recent approaches to privacy against statistical inference introduced by both computer scientists and information theorists do not make computational assumptions on the adversary. The distinguishing factor between privacy metrics is the adversary's inference goal (e.g., probability of correct guessing, small mean-squared reconstruction error) and how private information is modeled. Utility, in turn, is more difficult to quantify since it is inherently application dependent. Delineating and navigating the fundamental privacy-utility tradeoff is a challenge at the heart of recent privacy research.

The goal of this article is to review emerging privacy metrics and models introduced by the computer science and information theory communities in recent years. These models deal in quantities familiar to most information theorists (e.g., mutual information, f-divergences, probability of correct guessing) and are amenable toward mathematical tools developed by the information theory community over the past decades. We also discuss emerging privacy challenges in machine learning,

[2] Diffie and Hellman's paper has arguably among the most ambitious opening sentence ever published in the IEEE TRANSACTIONS ON INFORMATION THEORY, stating that "we stand today on the brink of a revolution in cryptography." In hindsight, they were absolutely correct.

such as federated learning (FL) and dataset obfuscation. These applications provide a unique opportunity for information theory research with real-world impact.

We hope to convince the reader that despite differences in notation and publication venues, the gap between current information-theoretic and computer science approaches to privacy is smaller than ever before, and there is ample opportunity for collaboration and cross-pollination of ideas between both fields. Perhaps if Shannon and Turing were to meet today to discuss privacy—say, in the hallways of Google, Apple, or Microsoft—they would likely find common ground.

Notation

Capital (e.g., X) and calligraphic letters (e.g., $\mathcal{X}$) 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 $X \sim P_X$. For the sake of simplicity, we assume that the probability distributions have finite support; however, the results demonstrated in this article can be generalized to probability distributions of continuous random variables by considering the Radon–Nikodym derivatives. 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 ith entry of $\mathbf{z}$. We denote $\mathbb{R}$ the set of real numbers, $\mathbb{R}^+$ the set of positive real numbers, and $\mathbb{N}$ the set of natural numbers. Finally, for $k \in \mathbb{N}$, $[k]$ denotes $[1, 2, \ldots, k]$ and Δ_k denotes the k-dimensional probability simplex.

Organization

In the remainder of the article, we first introduce information-theoretic quantities that measure privacy leakage using probability distance in section "Quantifying Privacy via Divergences." These privacy leakage measures are both used in differential privacy (section "Prior-Independent Privacy Mechanisms") and information-theoretic privacy (section "Prior-Dependent Privacy Mechanisms"). In section "Quantifying Privacy via Perturbation," we introduce the definition and properties of differential privacy and it variants, and connect DP with information-theoretic quantities. Finally, in section "Applications in Machine Learning," we discuss recent applications of differential privacy (DP) and information-theoretic privacy in the data-driven deep learning scenarios.

Overview of Privacy Mechanisms and Metrics

Privacy is usually ensured via a *privacy-preserving mechanism:* an algorithm that randomizes data (or a function thereof) in order to thwart unwanted statistical inferences. A privacy mechanism may, for example, add noise to the output of a query over a database, or randomize data in order to obfuscate private information prior to release to a third party. The performance of a privacy mechanism is quantified in terms of a *privacy metric*.[3] Naturally, the privacy metric to be attained is a key factor when designing and evaluating privacy-preserving mechanisms. Several works have proposed privacy metrics suitable for different application contexts and adversarial threats. What distinguishes different metrics are the kind of adversary they consider, the data sources they assume to be available to the adversary, and the aspects of privacy they measure [10]. Ideally, privacy metrics should carry operational meaning beyond their mathematical definition. For example, Asoodeh *et al.* [11] proposed privacy metrics based on the probability of an adversary identifying/guessing a given individual in a dataset, Liao *et al.* [12] introduced a tunable measure that can be adapted to specific adversarial actions, and Issa *et al.* [13] quantified maximal leakage where the adversary is capable of guessing any function of the dataset.

In this review, we discuss privacy-preserving mechanisms and privacy metrics through an information-theoretic lens: we formulate metrics in terms of underlying probability distributions and do not account for computational assumptions on an adversary. In particular, there are two families of privacy-preserving mechanisms based on assumptions made on private information: 1) *prior-independent* mechanisms, where minimal assumptions are made on the data distribution and an adversary's side information; and 2) *prior-dependent* mechanisms, where the mechanism designer has (partial) knowledge about private data statistics (e.g., probability distribution) and adversarial capabilities.

Prior-Independent Privacy Mechanisms

The most popular privacy metric is DP. Broadly speaking, DP quantifies how small perturbations at the input of a privacy mechanism affects the probability distribution of the output of the mechanism. A mechanism is said to be ϵ-differentially private if the probability of any output event does not change by more than a multiplicative factor e^{ϵ} for any two *neighboring* inputs. The definition of neighboring inputs depends on a predefined metric on the input space (e.g., two inputs within a Hamming distance of 1).

DP is a prior-independent mechanism (up to the definition of neighboring) since it does not depend on the probability distribution of the data. A motivating example for the definition of DP is in statistical queries over a database: the result of a query should be approximately the same no matter whether a dataset contains an individual's record. The privacy guarantee of DP can usually be achieved by additive noise mechanisms, that is adding a small perturbation/random noise sampled from different distributions to the released data (e.g., Gaussian, Laplacian, or exponential noise [14]).

Since its introduction in 2006, several variants of DP have been proposed. The main distinguishing factors between

[3] The term "metric" is used here not in the usual mathematical sense (i.e., a distance function), but rather as a measure of privacy risk.

these metrics is how the concept of "neighboring" is defined and how the change in output probability distribution induced by two neighboring inputs is quantified. For example, approximate differential privacy relaxes DP by allowing an additional small additive parameter δ [15] in addition to the multiplicative factor e^{ϵ}. Local DP assumes that *all* inputs are neighboring, aiming to model adversaries who have access to individual data points in a dataset [16]. Rényi differential privacy uses Rényi divergence to measure the difference in output distribution from two neighboring inputs [17], and is closely related to zero-concentrated DP [18], [19] and variations inspired by formulations based on hypothesis testing [20].

DP satisfies two desirable properties of a privacy metric: composability and robustness to postprocessing (see section "Composability and the Moments Accountant Approach"). Since privacy leakage may accumulate when an adversary observes multiple responses from a DP mechanism, the composability property guarantees that the aggregate output after multiple observations still satisfies the DP. Quantifying how privacy decays is an important theme in recent DP research [21]–[23]. Moreover, DP is robust to postprocessing in the sense that the outputs differentially private mechanism is still differentially private. In information-theoretic parlance, DP satisfies a form of data-processing inequality—a fact that follows naturally from the connection between DP and f-divergences (see section "Quantifying Privacy via Divergences"). Together, the composability and robustness to postprocessing allow the designer of privacy mechanisms to modularize their construction and analysis for a target privacy leakage budget.

Prior-Dependent Privacy Mechanisms

When statics and/or the probability distribution of the dataset can be (partially) known or estimated, the design of privacy-preserving mechanisms and privacy metrics has been studied in information-theoretic (IT) privacy. IT privacy metrics aim to quantify the amount of information an adversary gains about private features of the data given access to disclosed data. Here, privacy metrics are formulated in terms of divergences between probability distributions such as f-divergences (e.g., mutual information [24], chi-squared divergence [25], etc.), and Rényi divergence [12]. One of the advantages in using divergences and related quantities as privacy metrics is that they can often be equipped with operational meaning in terms of an adversary's ability to infer sensitive data. For instance Asoodeh *et al.* [11] studied a maximum *a posteriori* adversary that can guess specific private features and Issa *et al.* [13] introduced maximal leakage to quantify worst-case privacy threats where the adversary is capable of guessing any function of the dataset (not limited to the private features). Despite having significantly different operational meanings, the privacy measures in both of these examples can be quantified by divergences between probability distributions. In addition to quantifying privacy leakage, divergence measures that are common in information theory can also be used to quantify the utility of the released data (see, e.g. [26], where utility is quantified in terms of mutual information). The tradeoff between allowing a reasonable amount of utility to be drawn from the disclosed data and satisfying a privacy guarantee is analogous to the rate-distortion tradeoff found in lossy compression: at one extreme, no data are released (perfect privacy, no utility), and at the other extreme, data are disclosed as-is (no privacy, maximum utility). By using prior knowledge on the data statistics and tractable assumptions on adversary's inference capability, IT privacy cannot only characterize the fundamental privacy limits, but also help understand how to navigate the privacy-utility tradeoff [27].

Quantifying Privacy via Divergences

Consider two distributions P and Q with support $\mathcal{X}$. The likelihood ratio[4] $l(x)$ for $x \in \mathcal{X}$ is defined as

$$l(x) \triangleq \frac{P(x)}{Q(x)}. \tag{1}$$

When considering two random variables $S, Y \sim P_{S,Y}$, where S represents an individual's private/sensitive features (e.g., political preference), and Y the released data, and setting $P = P_{S,Y}$ and $Q = P_S P_Y$, the likelihood ratio becomes

$$l(s,y) = \frac{P_{S,Y}(s,y)}{P_S(s)P_Y(y)}. \tag{2}$$

This quantity is at the heart of most information-theoretic measures of privacy [13] as well as DP [14], [19]. Of course, the logarithm of this likelihood ratio $i(s,y) \triangleq \log l(s,y)$, termed *information density* [28], plays a central role in spectral methods, finite-blocklength analysis, statistics (binary hypothesis testing) in information theory [29]. Intuitively, the information density captures the change in belief about a sensitive attribute S upon an observation of disclosed information Y.

Measuring Privacy Leakage

The privacy leakage can be understood as the "amount of information" obtained about the private feature S through observing the released data Y. A widely used measure of the mutual dependence between the two variables is Shannon's *mutual information* $I(S;Y)$ between S and Y, given by the expectation of the information density

$$I(S;Y) \triangleq \mathbb{E}_{P_{S,Y}}[i(s,y)] = D(P_{S,Y} \| P_S P_Y) \tag{3}$$

where $D(P\|Q)$ is the Kullback–Leibler (KL) divergence. The equivalence between mutual information and the KL divergence allows us to generalize the mutual information by other information-theoretic divergences based on the likelihood ratio $l(s,y)$ and the information density $i(s,y)$. For example,

[4] This quantity is frequently referred to as *lift* in the data mining, or *pointwise mutual information* in natural language processing.

f-divergences [30] considers the expectation of a convex transformation of the likelihood ratio, and the Rényi divergence [12], [13] is proportional to the cumulant-generating function of the information density. The latter can also be viewed as a parameterized (by α) family of divergences, allowing for more freedom in tuning the metric for different needs in practice. These metrics are closely connected to differential privacy (section "Quantifying Privacy via Perturbation"), and widely used in information and estimation-theoretic privacy (section "Renyi Entropy and Divergence").

f-Divergences

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 set $\mathcal{X}$, and P is absolutely continuous with respect to Q. The f-divergence between P and Q is given by

$$D_f(P\|Q)\triangleq\mathbb{E}_Q\left[f\left(\frac{P(X)}{Q(X)}\right)\right]. \tag{4}$$

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

$$I_f(S;X)\triangleq D_f(P_{S,X}\|P_SP_X). \tag{5}$$

For more properties of f-divergences see, for example, [30]. The KL divergence $D(P\|Q)$ and Shannon's mutual information $I(S;X)$ are special cases of (4) and (5), respectively, when $f(t)=t\log t$. Moreover, other measures that are widely used in statistics and information theory can also be formulated as f-divergences. For example, α-Hellinger divergence $\chi^\alpha(P\|Q)$ of order $\alpha\in(0,1)\bigcup(1,\infty)$ uses $f_\alpha(t)=\frac{t^\alpha-1}{\alpha-1}$. An important f-divergence for privacy is the E_λ-divergence [30] with $f_\lambda(t)=\max\{t-\lambda,0\}$,

$$E_\lambda(P\|Q)=\sup_{X\in\mathcal{X}}[P(X)-\lambda Q(X)]. \tag{6}$$

These information-theoretic divergences are closely related to DP, see section "Quantifying Privacy via Perturbation."

Rényi Entropy and Divergence

The Rényi entropy $H_\alpha(P)$ and divergence $R_\alpha(P\|Q)$ of order $\alpha\in\mathbb{R}^+/\{1\}$ are defined, respectively, as

$$H_\alpha(P)\triangleq\frac{1}{1-\alpha}\log\sum_x P(x)^\alpha=\frac{\alpha}{1-\alpha}\log\|P\|_\alpha$$
$$R_\alpha(P\|Q)\triangleq\frac{1}{\alpha-1}\log\left(\sum_x\left(\frac{P(x)}{Q(x)}\right)^\alpha Q(x)\right). \tag{7}$$

Both of these two quantities are defined by their continuous extensions for $\alpha=1$ and ∞. In particular, for $\alpha=1$, the Rényi entropy and divergence recover Shannon entropy and KL divergence, respectively [12]. For $\alpha=\infty$, $H_\infty=\min_x\log 1/P(x)$ is called the min-entropy and $R_\infty(P\|Q)=\max_x\log P(x)/Q(x)$ is called the max-divergence. Rényi entropy and divergence generalize the usual notion of mutual information. Notably, Arimoto's and Sibson's mutual information have recently been proposed as operational measures for information leakage, see e.g., [11], [13].

Consider two random variables $(X,Y)\sim P_{X,Y}$, Arimoto's mutual information of order $\alpha\in\mathbb{R}^+/\{1\}$ is given by

$$I_\alpha^{\mathsf{A}}(X;Y)\triangleq H_\alpha(X)-H_\alpha(X|Y). \tag{8}$$

It can also be defined (by continuity) for the extreme cases $\alpha=1$ and ∞, respectively, as $\lim_{\alpha\to1}I_\alpha^{\mathsf{A}}(X;Y)=I(X;Y)$ and $I_\infty^{\mathsf{A}}(X;Y)\triangleq\lim_{\alpha\to\infty}I_\alpha^{\mathsf{A}}(X;Y)$. The latter characterizes the ability of an adversary to correctly guess X given Y. In particular, it can be verified [11] that $I_\infty^{\mathsf{A}}(X;Y)=\log\frac{\mathsf{P_c}(X|Y)}{p_X^*}$, where

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

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

Another operational measure of information leakage recently proposed is Sibson's mutual information [13] of order $\alpha\in\mathbb{R}^+/\{1\}$ between X and Y, which is given by

$$I_\alpha^{\mathsf{S}}(X;Y)\triangleq\inf_{Q_Y}R_\alpha(P_{X,Y}\|P_XQ_Y). \tag{9}$$

One can similarly define $I_\infty^{\mathsf{S}}(X;Y)$ as the limit of $I_\alpha^{\mathsf{S}}(X;Y)$ when $\alpha\to\infty$. This quantity, termed *maximal leakage*, was recently shown to bear an important interpretation in terms of worst-case privacy threats [13]. More precisely, maximal leakage is equal to the logarithm of the multiplicative gain in guessing *any function* of X given the observation of Y, that is

$$I_\infty^{\mathsf{S}}(X;Y)=\max_{U-X-Y}\log\frac{\mathsf{P_c}(U|Y)}{p_U^*} \tag{10}$$

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

Information and Estimation-Theoretic Privacy

Several privacy works in information theory considered a scenario where two parties share information over a noiseless channel [12], [13], [24]. Consider a Markov chain

$$S-X-Y \tag{11}$$

where $X\in\mathcal{X}$ is the observed data; $S\in\mathcal{S}$ is the sensitive/private attributes, and $Y\in\mathcal{Y}$ is the released data. The goal of information and estimation-theoretic privacy is to determine a conditional probability (i.e., a channel) $P_{Y|X}$ as a (randomized) privacy-preserving mechanism to publish the data without leaking the private information S. Next, we discuss three setups for designing the privacy mechanisms.

The Privacy Funnel

Assuming that both $P_{S,X}$ is known in (11), the first scenario, termed the privacy funnel (PF), seeks to determine a mapping $P_{Y|X}$ that minimizes the privacy leakage $I(S;Y)$ while preserving useful information (utility) $I(X;Y) \geq x$, where $x \geq 0$ is a controllable parameter, i.e.,

$$\min_{P_{Y|X}} I(S;Y) \text{ such that } I(X;Y) \geq x. \tag{12}$$

Inspired by rate-distortion theory [26], the PF (and its variants) has shown to be an useful information-theoretic framework in designing privacy mechanisms [27]. The PF is also closely related to the famous information bottleneck; in fact, the information bottleneck and the PF jointly determine the achievable set of the mutual information pairs $\{I(S;Y), I(X;Y) : S - X - Y\}$ [25].

Privacy Against Statistical Inference

The second setup [24] considers the worst-case side information the threat model may have about the input under (11). Consider a cost function $c : \mathcal{S} \times \Delta_{|\mathcal{S}|-1} \to \mathbb{R}^+$, and let the solutions $q \in \Delta_{|\mathcal{S}|-1}$ prior to and after observing Y to be c_0^* and c_y^* respectively, i.e.,

$$\begin{aligned} c_0^* &= \min_{q \in \Delta^{|\mathcal{S}|-1}} \mathbb{E}_{S \sim P_S(\cdot)}[c(S,q)] \\ c_y^* &= \min_{q \in \Delta^{|\mathcal{S}|-1}} \mathbb{E}_{S \sim P_{S|Y}(\cdot|Y=y)}[c(S,q)] \end{aligned} \tag{13}$$

the average and maximum gain of an adversary is defined as $\Delta c = c_0^* - \mathbb{E}_Y[c_y^*]$ and $\Delta c^* = c_0^* - \max_{y \in \mathcal{Y}} c_y^*$, respectively. The goal of the channel designer is to find $P_{Y|X}$ to minimize the gain if the adversary while preserving useful information in X by limiting the distortion between X and Y given a distortion function $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$. In summary, the overall optimization problem is

$$\min_{P_{Y|X}} \Delta c \text{ or } \Delta c^* \text{ such that } \mathbb{E}_{X,Y}[d(X,Y)] \leq \Delta. \tag{14}$$

When choosing the log-loss cost function $c(S,q) = -\log q(S)$, the average cost gain and maximum cost gain are $\Delta c = I(S;Y)$ and $\Delta c^* = H(S) - \min_{y \in \mathcal{Y}} H(S \mid Y = y)$, respectively, with the latter being the maximum information leakage, and (14) can be efficiently solved as a convex optimization.

Privacy in Terms of Guessing

The last threat model is expanded even further by assuming that only $P_{Y|X}$ is known (i.e., the prior over the set of observations P_X is unknown), this latter threat model treats privacy as an exclusive function of $P_{Y|X}$.

Maximal Leakage. One potential limitation of the previous threat model is that the channel designer knows the potential function of interest of the adversary (that is, $P_{S|X}$ is known to *a priori*). To overcome this limitation [13] defines the *maximal leakage* from X to Y as

$$\mathcal{L}_{maxL}(X \to Y) \triangleq \sup_{S-X-Y-\hat{S}} \log \frac{P(S=\hat{S})}{\max_{s \in \mathcal{S}} P_S(s)} \tag{15}$$

where the supremum is taken over all distributions $P_{S|X}, P_{\hat{S}|Y}$ with $S, \hat{S}$. Equation (15) measures the (worst case) information gain the adversary obtains from observing Y in guessing the value of a potentially sensitive attribute S. It can be shown that the maximal leakage is the Sibson mutual information of order ∞, i.e., $\mathcal{L}_{maxL}(X \to Y) = I_\infty^{\mathsf{S}}(X;Y)$, which shares many properties with Shannon information, namely nonnegativity and data processing inequality [13].

α-Maximal Leakage: Liao *et al.* [12] proposed two measures of information leakage. For $\alpha \in [1, \infty]$, the α-leakage is defined for a joint distribution $P_{S,Y}$ as

$$\mathcal{L}_\alpha(S \to Y) \triangleq \frac{\alpha}{\alpha - 1} \log \frac{\max_{P_{\hat{S}|Y}} \mathbb{E}[P(\hat{S} = S|S,Y)^{\frac{\alpha}{\alpha-1}}]}{\max_{P_{\hat{S}}} \mathbb{E}[P(\hat{S} = S|S)^{\frac{\alpha}{\alpha-1}}]} \tag{16}$$

and the *α-maximum leakage* as

$$\mathcal{L}_\alpha^{\max}(X \to Y) \triangleq \sup_{S-X-Y} \mathcal{L}_\alpha^{\max}(S \to Y). \tag{17}$$

Similarly to $\mathcal{L}_{maxL}$, α-maximum leakage measures the multiplicative increase in probability of correctly guessing any sensitive attribute S from Y that needs to be sent through the intermediary channel $P_{Y|X}$. By using the equivalence $\mathcal{L}_\alpha(S \to Y) = I_\alpha^A(S;Y)$, it can be shown that

$$\mathcal{L}_\alpha^{\max}(X \to Y) = \begin{cases} \sup_{P_{\hat{X}}} I_\alpha^S(\hat{X};Y), \, \alpha \in (1, \infty] \\ I(X;Y), \, \alpha = 1. \end{cases} \tag{18}$$

In this sense, α-maximal leakage is an intrinsic property of the channel $P_{Y|X}$ (and alphabet size $|\mathcal{X}|$) since it implicitly optimizes over all possible priors P_X.

Quantifying Privacy via Perturbation

DP bounds the statistical difference in the output distribution of a randomized algorithm induced by a small perturbation of its input. Given a privacy-preserving (randomized) mechanism $m : \mathcal{D} \to \mathcal{Y}$ that takes a dataset $D \in \mathcal{D}$ and returns an output $y \in \mathcal{Y}$, the privacy loss [14] measures the statistical difference between a given pair of neighboring datasets[5] $D, D' \in \mathcal{D}$, and output set $\mathcal{S} \subset \mathcal{Y}$ can be defined as the log likelihood ratio

[5] In Dwork *et al.* [14], D and D' are a collection of records from $\mathcal{X}$ that can be expressed with a histogram notation ($D, D' \in \mathbb{N}^{|\mathcal{X}|}$) then D is adjacent to D' if $||D - D'||_1 \leq 1$. Similarly, it is common to define two datasets as neighboring if they differ in at most one record, denoted as $D \sim D'$.

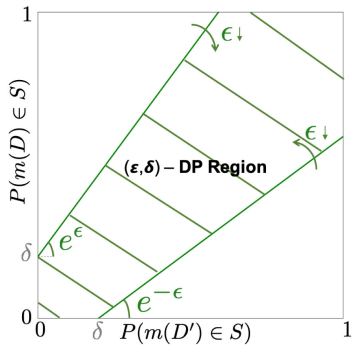

Figure 1

Visualizing the (ϵ, δ)-DP mechanism in terms of outcome probabilities (see (21)). A mechanism is (ϵ, δ)-DP if for every pair of neighboring datasets D, D' and outcome S, the resulting (joint) probability lies in the DP cone. Note that the δ parameter has a large impact on low probability events, while ϵ controls the width of the band around the identity line.

$$L_m(\mathcal{S}; D, D') = \log \frac{\Pr(m(D) \in \mathcal{S})}{\Pr(m(D') \in \mathcal{S})}. \tag{19}$$

The strong guarantees given by DP and its variants are related to bounding this quantity over all possible sets $\mathcal{S}$ and for all possible inputs D, D', as we illustrate next.

DP and Its Variants

Given the leakage parameter $\epsilon \geq 0$, the mechanism m is said to be ϵ-differentially private if for any neighboring datasets $D, D' \in \mathcal{D}$ and $\mathcal{S} \subset \mathcal{Y}$,

$$L_m(\mathcal{S}; D, D') \leq \epsilon. \tag{20}$$

Perfect privacy is assured when $\epsilon = 0$, and when $\epsilon = \infty$, there is no privacy guarantee.

DP can be understood as a bound on the log likelihood ratio between the probability distributions of the output of a mechanism given neighboring inputs. This definition inherently depends on how "neighboring inputs" is defined (e.g., inputs within unit Hamming distance of each other). The most stringent form of DP is when *all* inputs are considered neighboring—a definition referred to as local differential privacy (LDP) [16]. The definition of ϵ-DP can be relaxed to (ϵ, δ)-DP by introducing an additive parameter $\delta \in [0, 1)$ to the constraint (20). A mechanism is said to be (ϵ, δ)-DP if for all neighboring datasets $D, D' \in \mathcal{D}$ and $\mathcal{S} \subset \mathcal{Y}$,

$$\log \frac{\Pr(m(D) \in \mathcal{S}) - \delta}{\Pr(m(D') \in \mathcal{S})} \leq \epsilon. \tag{21}$$

Observe that $(\epsilon, 0)$-DP is equivalent to ϵ-DP. See Figure 1 for a visual representation.

Rényi Differential Privacy

Rényi differential privacy uses Rényi divergence to measure the statistical difference between outputs of a mechanism induced by neighboring inputs [17]. A randomized algorithm satisfies ϵ-Rényi differential privacy (RDP) of order α (or (α, ϵ)-RDP) if for all neighboring datasets $D, D' \in \mathcal{D}$,

$$R_\alpha(P_D \| P_{D'}) \leq \epsilon \tag{22}$$

where $P_D = \Pr(m(D) \in \mathcal{S})$ and $P_{D'} = \Pr(m(D') \in \mathcal{S})$. The relationship between the Rényi divergence with $\alpha = \infty$ and DP is immediate. The definition of ϵ-DP in (20) can be equivalently expressed as $R_\infty(P_D \| P_{D'}) \leq \epsilon$, where $D_\infty(P \| Q)$ is the max-divergence. Comparing to the KL divergence, max-divergence can be viewed as a worst-case analog of KL divergence, similar to the way that min-entropy (section "Rényi Entropy and Divergence") is a worst-case analog of Shannon's entropy.

Pufferfish Privacy

Pufferfish privacy [31], similar to DP, considers a threat model where privacy is defined in terms of decision making (i.e., the adversary is trying to decide between two hypothesis after observing a shared variable). The proposed setting allows the user to tailor privacy definitions based on their needs and assumptions about the attacker's potential beliefs.

Given a set of potential secrets $\mathcal{S}$, a set of discriminative pairs $\mathcal{S}_p \subseteq \mathcal{S} \times \mathcal{S}$, and a collection of data evolution scenarios $\mathcal{E}$, an algorithm $m : \mathcal{D} \to y$ is ϵ-Pufferfish$(\mathcal{S}, \mathcal{S}_p, \mathcal{E})$ private if for all $y \in \mathcal{Y}$, for all $(s_i, s_j) \in \mathcal{S}_p$, and for all $\theta \in \mathcal{E}$ such that $p(s_i|\theta) > 0, p(s_j|\theta) > 0$, the following holds:

$$e^{-\epsilon} \leq \frac{P(s_i \,|\, \theta, m(D) = y)}{P(s_j \,|\, \theta, m(D) = y)} \frac{P(s_j \,|\, \theta)}{P(s_i \,|\, \theta)} \leq e^{\epsilon}, \ \epsilon \geq 0. \tag{23}$$

In other words, Pufferfish privacy requires that the odds prior to observing an output of our mechanism $\frac{P(s_i \,|\, \theta)}{P(s_j \,|\, \theta)}$ are close to the odds after making our observation $\frac{P(s_i \,|\, \theta, m(D)=y)}{P(s_j \,|\, \theta, m(D)=y)}$ for any belief θ the potential adversary may have. With an appropriate choice of $\mathcal{S}, \mathcal{S}_p, \mathcal{E}$, ϵ-Pufferfish privacy can be equivalent to ϵ-DP. Algorithms satisfying particular instantiations of ϵ-Pufferfish $(\mathcal{S}, \mathcal{S}_p, \mathcal{E})$ are shown in [31].

The Additive Noise Mechanism

DP can be achieved by a mechanism where noise is added to a function f computed over data, and the variance of the noise is dependent on the ℓ_p-sensitivity of f, defined as $\Delta_p(f) \triangleq \max_{D \sim D'} \|f(D) - f(D')\|_p$. Intuitively, a greater variance of a noise is needed to randomize a more sensitive function. One example of such mechanism m is of the form

$$m(x) = f(x) + L\left(0, \frac{\Delta_1(f)}{\epsilon}\right) \tag{24}$$

where $L(a, b)$ is the Laplace distribution. This Laplace mechanism is a prototypical ϵ-DP algorithm that releases an approximate

(noisy) answer to the function f. Similarly, a Gaussian mechanism[6] with $N(0,\sigma^2)$ being the Gaussian distribution is defined as

$$m(x) = f(x) + N(0,\sigma^2) \tag{25}$$

can meet (ϵ,δ)-DP for $\epsilon < 1$ and $\sigma > \sqrt{2\log 1.25/\delta}\Delta_2(f)/\epsilon$. We note that there are several other mechanisms that account for different data types, see, for example [32].

Composability and the Moments Accountant Approach

One of the most desirable properties of DP is its *composability*, i.e., the combination of differentially private mechanisms is itself differentially private. To be more precise, the sequential combination of k independent (ϵ_i,δ_i)-DP mechanisms m_i for $i \in [k]$ is $(\sum_{i\in[k]}\epsilon_i, \sum_{i\in[k]}\delta_i)$-DP. Similarly, the parallel combination k independent ϵ_i-DP mechanisms is $\max_{i\in[k]}\epsilon_i$-DP. Another important property of DP is its *robustness to postprocessing*. Formally, if a mechanism m is ϵ_i-DP, then $f\circ m$ is also ϵ_i-DP for any deterministic or randomized function f. For a more interesting *k*-fold adaptive composition, where the independence assumption among the DP mechanisms is dropped, and each individual mechanism can take an auxiliary query parameter, see Dwork *et al.* [14].

The sequential and parallel composabilities are agnostic to the specific mechanism, meaning that they only depend on the type of compositions and the parameters ϵ and δ. To achieve tighter DP composition bounds, the *moments accountant approach* [7] was recently proposed to relate the privacy parameters ϵ and δ with the moments of the privacy loss random variable in (19). The λth moment $\alpha_m(\lambda; D, D')$ is defined as the cumulant moment generating function evaluated at λ,

$$\alpha_m(\lambda; D, D') = \log \mathbb{E}_{Y\sim m(D)}\left[e^{\lambda L_m(Y;D,D')}\right] \tag{26}$$

and the worst-case λ^{th} moment $\alpha_m(\lambda)$ to bound all possible $\alpha_m(\lambda; D, D')$ is defined as

$$\alpha_m(\lambda) = \max_{D\sim D'} \alpha_m(\lambda; D, D'). \tag{27}$$

Abadi *et al.* [7, Th. 2] showed that the tail bound of the privacy loss can translate the moments bound of $\alpha_m(\lambda)$ into (ϵ,δ)-DP guarantees, i.e., for any $\epsilon > 0$ a mechanism m is (ϵ,δ)-DP with $\delta = \min_\lambda \exp(\alpha_m(\lambda) - \lambda\epsilon)$. Moreover, the privacy loss moments $\alpha_m(\lambda)$ of a composition of a sequence of adaptive mechanisms $m_1,\ldots,m_k$, where m_i is the output of the previous $i-1$ mechanisms can be upper bounded as

$$\alpha_m(\lambda) \le \sum_{i=1}^{k} \alpha_{m_i}(\lambda) \quad \forall\lambda. \tag{28}$$

The differentially private stochastic gradient descent (DP-SGD) algorithm, viewed as a sequence of adaptive Gaussian mechanisms, can be analyzed by the moments accountant method [7, Th. 1], leading to a stronger composition theorem than that in Dwork *et al.* [33].

[6] Note that Gaussian mechanism cannot meet ϵ-DP for any ϵ.

Information-Theoretic Divergences and DP

RDP lacks a clear operational interpretation, and thus RDP guarantees are often translated into DP guarantees [34]. Precisely, a mechanism m is (α,γ)-RDP is (ϵ,δ)-DP for any $\epsilon > \gamma$ and $\delta = e^{-(\alpha-1)(\epsilon-\gamma)}$. This translation is extensively used in many recent ML applications [34], [35]; however, it is loose and does not hold for all possible $\epsilon \ge 0$. Asoodeh *et al.* [23] exploited the E_λ-divergence to determine the optimal relationship between DP and RDP. In particular, given a (α,γ)-RDP mechanism m, it is $(\epsilon,\delta_\alpha^\epsilon(\gamma))$-DP for a given $\epsilon \ge 0$, where $\delta_\alpha^\epsilon(\gamma)$ is defined as

$$\delta_\alpha^\epsilon(\gamma) = \sup m \text{ that is } (\alpha,\gamma)\text{-RDP} \sup_{D\sim D'} E_{e^\epsilon}(m(D)\|m(D')). \tag{29}$$

Applications in Machine Learning

We review a few recent results in private machine learning and data science that implement the information-theoretic privacy (section "Information and Estimation-Theoretic Privacy") and DP (section "Quantifying Privacy via Perturbation").

Information Obfuscation

One particular problem that has received significant attention is that of learning data representations from which an adversary cannot reliably estimate a (known) sensitive variable S. The data representations can be viewed as the outputs of a (stochastic) parametric channel distribution $P_{Y|X}$, and can be learnt using neural networks under the framework of the constrained optimization formulation in [14] [36], [37], or by adversarial generative networks [38].

Figure 2 shows one of the applications in Bertran *et al.* [36], where the observations are facial images and the privacy goal is to prevent an adversary from inferring the (labeled) sex of the person in the image; the output variable Y shares the same support than the input variable X (i.e., Y is an image), the distortion metric in this case is preserving subject information as inferred by a model trained on undistorted images. Hsu *et al.* [37] further proposed that an information obfuscation mechanism should ideally target only the features in the data that potentially leak sensitive information, and perturb those information-leaking features in order to achieve better utility. Figure 3 illustrates the detection of information-leaking features that reveal emotion via information density estimation. A Gaussian obfuscation mechanism with provable guarantees is then applied to those information-leaking features.

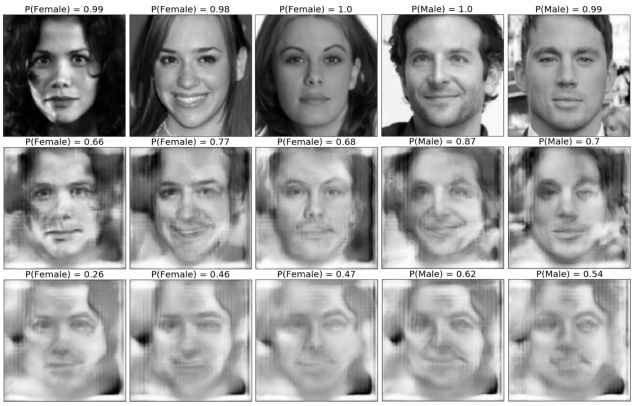

Figure 2

An image to image mapping that obfuscates information about labeled gender while preserving information required to perform subject identification; the mapping is implemented with a neural network. Each row has a different constraint on $I(S;Y)$, the last row being the tightest. The network that performs identity recognition (the desired utility in this example) can be trained using the obfuscated images and is able to achieve a top 5 accuracy larger than 90% from a base of 200 identities. Figure taken from [36].

Variations of (14) with mutual information $I(S;Y)$ as the adversary cost and $I(X;Y)$ as the utility directly translate to the bottleneck formulation [25], [26]. Since many of the aforementioned approaches are purely data driven, it has proven challenging to incorporate some of the stronger notions of estimation-theoretic privacy, partly because optimizing over any data prior P_X (as is required for Sibson's and Arimoto's MI) is a hard task to specify in high-dimensional datasets used in machine learning.

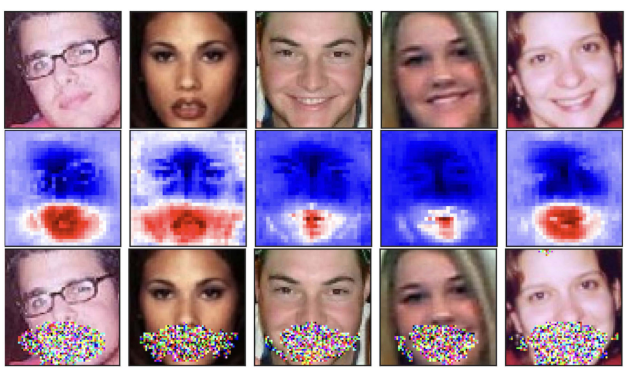

Figure 3

A Gaussian obfuscation mechanism on information-leaking features for emotion. The obfuscated images in the last row cannot be used to perform emotion classification task, but can still be used to infer other useful information, for example, gender. Figure taken from [37].

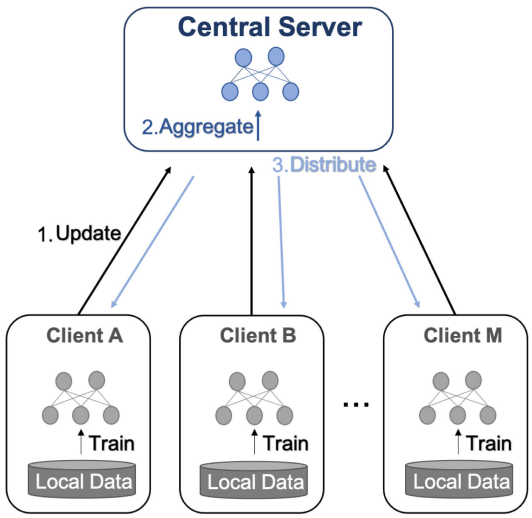

Figure 4

A classical federated learning scenario where a central servers receives updates from a client pool, and distributes the updated model. Clients locally train the model on their data and provide the updated weights. The central server is responsible for aggregating those updates into a unique updated model.

Differential Privacy in Deep Learning

There are several successful developments of differential privacy in modern deep learning methods, which depend heavily on stochastic gradient descent (SGD) to update the hyperparameters in a neural network. The training of neural networks requires large-scale datasets, which may be crowdsourced and contain sensitive information, and thus the neural network models should not expose private information in these datasets [7]. To address this issue, Abadi *et al.* [7] proposed DP-SGD (see section "Composability and the Moments Accountant Approach") to update the weights. DP-SGD enables computationally efficient training of neural networks within a modest privacy budget by adding carefully designed Gaussian noise to the learnt gradients, and has been studied in a wide range of applications [39], [40]. The privacy information may not only be inadvertently leaked during the training of a neural networks, but also from a pretrained model; for example, when performing knowledge distillation from a knowledgeable and pretrained teacher model. To address those concerns, private aggregation of teacher ensembles (PATE) is proposed, which transfers the knowledge of an ensemble of teacher models to a student model, with DP guaranteed by noisy aggregation of teachers' answers [35].

Federated Learning

A common FL scenario is presented in Figure 4, where a central server learns a parametric model by sharing information with different clients that have access to their own local data [41]. The advantage of this setting is that it leads to models with better generalization capabilities, owing to the potentially larger amounts of data the model has access to when

compared to standard, centralized learning. This benefits all clients, especially those that do not have enough data to train a reliable model on their own. Despite this advantage, there are immediate privacy concerns that arise from this collaboration, an example of this is the information the central server (or other clients) can infer from the client's data based on server–client interactions (potentially a major concern in critical domains like healthcare). It is therefore of great interest to provide privacy guarantees on the mechanism that the server uses to aggregate the client's shared information in the process of learning the model.

Variants of DP have been used to address these issues. For example, if the central curator that collects data from all clients is trustworthy, then the curator can be tasked with applying a DP mechanism on the aggregation process, preventing the FL output from leaking information. The need to have a trusted central server can be removed by requiring each client to apply a local DP mechanism to their updates before sharing the information with the server [42], [43]. The main downside of this approach is that the loss in utility from this mechanism needs to be compensated by an even larger pool of clients to maintain good performance [44]. Hybrid DP [45] strikes a compromise between these two approaches where a group of users trust the central server, and opt into the central curator model, with the remaining set of users opting to apply local DP before sharing. Recently, a secure aggregation protocol was proposed in [46], providing strong privacy guarantees on the process of aggregating high-dimensional data (e.g., adding network parameters shared by different clients as in FL). This approach still requires limited trust in the central server, and allows the server to observe the aggregate information at each communication round, which poses a potential information leakage vulnerability, the computation process itself may be computationally inefficient for sparse aggregate vectors. Addressing these limitations remains an open problem [41].

Dataset Sharing

Multiple communities, from machine learning to computer vision and natural language processing, have significantly advanced the research thanks in part to datasets sharing. Datasets are so important to the community that leading conferences such as NeurIPS are starting tracks fully dedicated to datasets.[7] While most of the publicly shared datasets are derived from open data sources, others (e.g., clinical) cannot be openly shared, resulting in slowing the progress in those fields. This challenge opens an opportunity to research on privacy, where, for example, we desire to privately share the data without any sacrifice in its utility. Such no compromise in utility is overall an important topic of study for the privacy community.

[7] [Online]. Available: https://neuripsconf.medium.com/announcing-the-neurips-2021-datasets-and-benchmarks-track-644e27c1e66c

Concluding Remark

We have discussed privacy from the lenses of statistics and information/estimation theory. The more we rely on data-dependent algorithmic-based decisions, the more privacy becomes critical. There are legal aspects of privacy, such as the European General Data Protection Regulation and the right to be forgotten; as well as personal preferences of privacy, all leading to many interesting theoretical and practical questions. One of the exciting areas for privacy research is in healthcare. First, utility cannot be sacrificed at all in most healthcare applications. Second, user's expectations and desires about the privacy of their data needs to be taken into account and not decided by a theory or an algorithm. For example, a user might be willing to sacrifice privacy to further improve utility, or even to help others obtain better utility. Such individual choices bring new directions into the field, and we are excited to see what new developments emerge in the literature.

Acknowledgments

The work of the Harvard team was supported in part by NSF under Grant CIF 1900750 and Grant CIF CAREER 1845852, by an Amazon Research Award, and by a Google Research Faculty Award. The work of the Duke team was supported in part by NSF, NIH, and DoD. These awards and gifts have not influenced the review material here reported.

References

[1] E. M. Guizzo, "The essential message: Claudes Shannon and the making of information theory," Ph.D. dissertation, Massachusetts Inst. Technol., Dept. Humanities, 2003.

[2] R. E. Price, "Claude E. Shannon: An interview conducted by Robert Price," IEEE History Center, Interview, vol. 423, p. 28, 1982.

[3] W. Diffie and M. Hellman, "New directions in cryptography," *IEEE Trans. Inf. Theory*, vol. 22, no. 6, pp. 644–654, Nov. 1976.

[4] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, 1975.

[5] M. Bloch and J. Barros, *Physical-Layer Security: From Information Theory to Security Engineering*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[6] R. Effing, J. Van Hillegersberg, and T. Huibers, "Social media and political participation: Are Facebook, Twitter and Youtube democratizing our political systems?," in *Proc. Int. Conf. Electron. Participation*, 2011, pp. 25–35.

[7] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

[8] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy*, 2008, pp. 111–125.

[9] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," in *Proc. IEEE 24th Int. Conf. Data Eng.*, 2008, pp. 277–286.

[10] I. Wagner and D. Eckhoff, "Technical privacy metrics: A systematic survey," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 1–38, 2018.

[11] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Estimation efficiency under privacy constraints," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1512–1534, Mar. 2019.

[12] J. Liao, O. Kosut, L. Sankar, and F. du Pin Calmon, "Tunable measures for information leakage and applications to privacy-utility tradeoffs," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8043–8066, Dec. 2019.

[13] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1625–1657, 2019.

[14] C. Dwork *et al.*, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3/4, pp. 211–407, 2014.

[15] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. Annu. Int. Conf. Theory Appl. Cryptographic Techn.*, 2006, pp. 486–503.

[16] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. IEEE Found. Comput. Sci.*, 2013, pp. 429–438.

[17] I. Mironov, "Rényi differential privacy," in *Proc. IEEE 30th Comput. Secur. Found. Symp.*, 2017, pp. 263–275.

[18] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," 2016, *arXiv:1603.01887*.

[19] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Proc. Theory Cryptography Conf.*, 2016, pp. 635–658.

[20] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," 2019, *arXiv:1905.02383*.

[21] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," vol. 27, pp. 2879–2887, 2014.

[22] B. Balle and Y.-X. Wang, "Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 394–403.

[23] S. Asoodeh, J. Liao, F. P. Calmon, O. Kosut, and L. Sankar, "A better bound gives a hundred rounds: Enhanced privacy guarantees via f-divergences," in *Proc. IEEE Int. Symp. Inf. Theory*, 2020, pp. 920–925.

[24] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, 2012, pp. 1401–1408.

[25] H. Hsu, S. Asoodeh, S. Salamatian, and F. P. Calmon, "Generalizing bottleneck problems," in *Proc. IEEE Int. Symp. Inf. Theory*, 2018, pp. 531–535.

[26] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *Proc. IEEE Inf. Theory Workshop*, 2014, pp. 501–505.

[27] F. P. Calmon, A. Makhdoumi, and M. Médard, "Fundamental limits of perfect privacy," in *Proc. IEEE Int. Symp. Inf. Theory*, 2015, pp. 1796–1800.

[28] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. Holden-Day, 1964.

[29] H. Koga *et al.*, *Information-Spectrum Methods in Information Theory*, vol. 50. Berlin, Germany: Springer Science & Business Media, 2013.

[30] I. Sason and S. Verdu, "f-divergence inequalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.

[31] D. Kifer and A. Machanavajjhala, "Pufferfish: A framework for mathematical privacy definitions," *ACM Trans. Database Syst.*, vol. 39, no. 1, pp. 1–36, 2014.

[32] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci.*, 2007, pp. 94–103.

[33] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *Proc. IEEE 51st Annu. Symp. Found. Comput. Sci.*, 2010, pp. 51–60.

[34] B. Balle, G. Barthe, M. Gaboardi, J. Hsu, and T. Sato, "Hypothesis testing interpretations and Rényi differential privacy," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2496–2506.

[35] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," 2016, *arXiv:1610.05755*.

[36] M. Bertran *et al.*, "Adversarially learned representations for information obfuscation and inference," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 614–623.

[37] H. Hsu, S. Asoodeh, and F. Calmon, "Obfuscation via information density estimation," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 906–917.

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

[39] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Brief. Bioinf.*, vol. 19, no. 6, pp. 1236–1246, 2018.

[40] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," 2016, *arXiv:1606.09375*.

[41] P. Kairouz *et al.*, "Advances and open problems in federated learning," 2019, *arXiv:1912.04977*.

[42] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately," 2017, *arXiv:1712.01524*.

[43] V. Pihur *et al.*, "Differentially-private "draw and discard" machine learning," 2018, *arXiv:1807.04369*.

[44] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?," *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, 2011.

[45] B. Avent, A. Korolova, D. Zeber, T. Hovden, and B. Livshits, "Blender: Enabling local search with a hybrid differential privacy model," in *Proc. 26th USENIX Secur. Symp.*, 2017, pp. 747–764.

[46] K. Bonawitz *et al.*, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 1175–1191.

Hsiang Hsu received the B.S. degree in electrical engineering and mathematics, and the M.S. degree in communication engineering from National Taiwan University (NTU), Taipei, Taiwan, in 2014 and 2016, respectively. He is currently working toward the Ph.D. degree with the Department of Computer Science, Harvard University, Cambridge, MA, USA.

His research interests include information theory and statistics, with applications to privacy, fairness, representation learning, and continual learning in machine learning. He is a Facebook Fellow.

Natalia Martinez received the B.S. degree in electrical engineering from Universidad de la Republica, Montevideo, Uruguay, in 2015. She is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA.

Her research interests include fairness, privacy, robustness, and explainability of machine learning algorithms.

Martin Bertran received the B.S. degree in electrical engineering from Universidad de la Republica, Montevideo, Uruguay, in 2015. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA.

His research interests include policy discovery in machine learning, with applications to privacy, fairness, and reinforcement learning.

Guillermo Sapiro (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the Technion, Israel in 1993. He did postdoctoral from the Massachusetts Institute of Technology. He is currently a James B. Duke Professor with Duke University, Durham, NC, USA, and is also with Apple, Inc., Cupertino, CA, USA. He was with HP Labs and the University of Minnesota.

He was the recipient of the ONR Young Investigator Award, the Presidential Early Career Awards for Scientist and Engineers, the NSF Career Award, the National Security Science and Engineering Faculty Fellowship, the Test-of-Time Award at ICCV and at ICML.

Dr. Sapiro is a Fellow of SIAM and American Academy of Arts and Sciences. He is the Founding Editor-in-Chief of the *SIAM Journal on Imaging Sciences*.

Flavio P. Calmon received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA in 2015. He is currently an Assistant Professor of electrical engineering with Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA, USA. Before joining Harvard, he was the inaugural Data Science for Social Good Post-Doctoral Fellow with IBM Research, Yorktown Heights, NY, USA.

He was the recipient of the NSF CAREER Award, the Google Research Faculty Award, the Amazon Research Award, the IBM Open Collaborative Research Award, the Harvard Lemann Brazil Research Fund Award, and the Harvard Bias 2 Fund Award. He received the inaugural *Tıtulo de Honra ao Merito* from the University of Brasılia in 2021.