Information-Theoretic Approaches to Differential Privacy

AYŞE ÜNSAL and MELEK ÖNEN, Digital Security Department, EURECOM, France

This tutorial studies relations between differential privacy and various information-theoretic measures by using several selective articles. In particular, we present how these connections can provide new interpretations for the privacy guarantee in systems that deploy differential privacy in an information-theoretic framework. Accordingly, the tutorial delivers an extensive summary on the existing literature that makes use of information-theoretic measures and tools such as mutual information, min-entropy, Kullback-Leibler divergence, and rate-distortion function for quantification and characterization of differential privacy in various settings.

CCS Concepts: • **Security and privacy → Information-theoretic techniques**; **Privacy-preserving protocols;**

Additional Key Words and Phrases: Differential privacy, mutual information, relative entropy, rate-distortion theory, min-entropy, leakage

ACM Reference format:
Ayşe Ünsal and Melek Önen. 2023. Information-Theoretic Approaches to Differential Privacy. *ACM Comput. Surv.* 56, 3, Article 76 (October 2023), 18 pages.
https://doi.org/10.1145/3604904

1 INTRODUCTION

Over the past decade, **machine learning** (**ML**) algorithms have found application in a vast and rapidly growing number of systems for analyzing and classifying large amounts of data. Despite the improvement and comfort that was brought to our daily lives by applications that employ these algorithms, they also gave cause for concern in terms of security and data privacy due to their undesired consequences. The increasing popularity of ML techniques opened the door for attackers, especially when these techniques were deployed to be used in critical areas such as intrusion detection, autonomous driving, or healthcare. In particular, an adversary may look for means to modify the model, misclassify some inputs, and consequently succeed in unauthorized cyber access, car accidents, or even health problems. It is not unrealistic to imagine the scenario, where a self-driving car causes an accident due to ignoring a stop sign, which through tampering by an adversary was made to look like a parking sign.

In addition to the security aspect of such an attack, user-data privacy is also prone to violations in this problem. Such data is considered as highly sensitive, since it contains information on location that could lead to the discovery of personal habits and may enable vehicle identification. In

Authors' address: A. Ünsal (corresponding author) and M. Önen, Digital Security Department, EURECOM, Campus SophiaTech, 450 Route des Chappes, Biot 06410, France; emails: {ayse.unsal, melek.onen}@eurecom.fr.

0360-0300/2023/10-ART76 $15.00
https://doi.org/10.1145/3604904

general, the high quality and high accuracy of ML predictions strongly depend on the collection of large datasets. Such a large-scale data collection gives cause for privacy concerns and makes users vulnerable to fraudulent use of personal information. When individuals willingly share some of their personal data with an Internet service, statistical independence of the representation of the data and the actual individual is a desired quality of the underlying system. At least from a conceptual perspective, a measure of this independence relates to the amount of privacy an individual can expect from the system. However, it is possible to successfully de-anonymize or re-identify the owner of the data as proven by a number of studies as follows [19, 33, 40, 45]. For instance, Facebook and Cambridge Analytica are real-life examples of massively used online services, which were proven to be a threat to privacy of individuals back in 2010, when Cambridge Analytica acquired a great number of Facebook users' data for the purpose of using the right political advertisement. More recently, it was discovered that Pegasus spyware has been used for reading text messages, tracking calls and locations, and accessing the targeted device's camera and microphone in many versions of Apple's iOS and Android [8]. These few examples of privacy rights' violations make it clear that protecting the privacy of personal data is a major concern in today's world.

In order to address data-privacy requirements in such contexts, two application methods are used in current systems, namely, local and global privacy. In local privacy methods, individuals publish a private version of their own information, as is the case of a social networking website. Global privacy methods make use of a trusted (central) server or curator that publishes private query responses related to a group of individuals. A common characteristic of both approaches is that data is typically coded using some randomizing function prior to its publication. **Differential privacy** (**DP**) [13] is a stochastic measure of privacy that is now used in conjunction with ML algorithms while managing large datasets to ensure data privacy of individual users. It has furthermore been used to develop practical methods for protecting private user data when they provide information to the ML system. In these cases, the use of a DP measure aims to preserve the accuracy of the ML model without incurring a cost of the privacy of individual participants. An embedded application in Google's Chrome Web Browser [20], a Census Bureau project called OntheMAP [29], LinkedIn, and Apple's iOS 11 are only a few examples of real-life applications that have already deployed DP to address and overcome this vulnerability of users in terms of privacy of personal information.

A mechanism or a randomized function of a dataset is called *differentially private* if the absence or presence of any participant's data has a negligible impact on the output of the mechanism when any of the participants decides to submit or equivalently remove their data from a statistical dataset. This idea is roughly depicted in Figure 1. In some sense, DP is a notion of robustness against such changes in the dataset. The degree of this change is measured and determined by an adjustable privacy parameter (or the privacy budget) and the amount of the change that any single argument to the system reflects on its output is called the sensitivity of the system. The major challenge is to offset the accuracy of the output of a statistical dataset against the level of the privacy protection guaranteed to the participants. Indeed, noisier data results in a stronger level of privacy due to increased randomness and this reflects as a reduction in accuracy of the output.

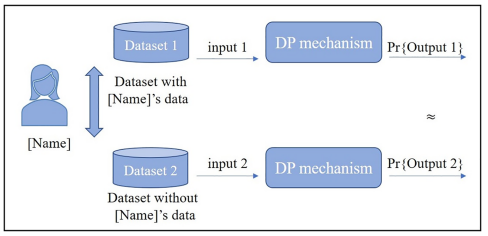

Fig. 1. Differential privacy.

Timeliness and necessity of the tutorial: DP raises great interest among researchers, particularly from computer science and statistics circles, who contributed to what we already know about

this strong mathematical formulation of privacy. There are several detailed surveys of what is known today regarding DP from the perspective of computer scientists and statisticians [14, 16, 18]. More recently, also researchers in information theory/electrical engineering circles contributed to the literature on the subject. However, a full information-theoretic understanding of DP and its information-theoretic connections with other trustworthy features are still lacking. This tutorial provides a *selective* summary of what we know regarding the relationship between DP and information theory to enable information theorists, primarily, to build up on that to produce fundamental formulations and limits of privacy in various settings.

On the relevance of information-theoretic connections with DP: Originally, the use of the mutual information functional as a privacy metric dates back to [28] for studying the domain of genome privacy prior to the existence of DP. Even though there are different opinions on the form of the exact relation, a number of studies relate the (conditional [9] or unconditional [32, 43]) mutual information between the entries of the dataset and the query response to DP, which could be interpreted as a measure of utility as well as of privacy. Under certain conditions, differential privacy and the *mutual information DP*, have proven to be equal in [9] where the authors redefine well-known information-theoretic quantities as privacy constraint. Overall, a mutual information-based approach to DP will allow many rules and properties that apply to the mutual information functional to be carried on to DP leaving no room for ambiguity regarding the essence of the privacy guarantee. Furthermore, in [9], the mutual information-based DP removes the requirement for neighborhood among datasets and strengthens the original definition. Hereafter, we enlist possible directions of research where the information-theoretic connections with DP is pertinent. The reader should note that the following list is exemplary and non-exhaustive. Some items will be studied in detail within the content of this tutorial in further sections.

- **Cryptography:** A major example is the connection with *semantic security* via an information-theoretic approach. [7] proves an equivalence between a mutual information-based DP constraint and semantic security where a maximization is taken over database distributions. Additionally, [44] introduces a new data-privacy protection model that aims to achieve *Dalenius' goal* as well as to have better utility. The privacy channel capacity results are obtained through direct translations of well-known information-theoretic approaches to DP. In particular, the parallel drawn between the information privacy model and the multiple-access channel makes a great promise for the use of an information-theoretic framework to quantify the privacy guarantee that a differentially private system can provide to its users.
- **Security:** [41] presented an application of the so-called Kullback-Leibler DP [9] (to be defined later) for detecting misclassification attacks in differentially private Laplace mechanisms. Accordingly, the corresponding distributions of relative entropy are considered as the differentially private noise with and without the adversary's advantage in order to establish the relationship between the impact of the attack and the detection of the adversary as a function of the sensitivity and the privacy budget of the mechanism. Besides adversarial classification, information-theoretic approaches for bounding the *communication complexity* of computing a function, which originally uses combinatorial measures [35], can also be applied to DP. Information complexity [4] is a lower bound on communication complexity that is obtained using Shannon's mutual information and refers to the minimum amount of information that a communication protocol leaks about its users' inputs. [30] introduces an upper bound on the information cost of a two-party differentially private protocol using the same approach that will be studied in detail in Section 4. [27], on the other hand, covers the privacy of physical layer for a two-receiver broadcast channel through analyzing connections between a differential privacy-based metric to physical layer secrecy. Accordingly, the

authors show that for the privacy of anonymous communication networks in the case of a degraded two-user broadcast channel, differentially private receiver-message unlinkability is equivalent up to a constant to several secrecy metrics. Finally, [27] presents the rate region of the (ϵ, δ)-differentially private receiver-message unlinkability satisfying strong secrecy.

- **ML: Probably approximately correct (PAC)** learning theory, which composes the mathematical framework of ML, is related to differentially private learning by using the mutual information function in [31]. Accordingly, the author establishes an information-theoretic connection between the Gibbs estimator, which gives the minimum of PAC-Bayesian bounds, and the exponential mechanisms to show that the Gibbs estimator minimizes the expected empirical risk and the mutual information between the sample and the predictor.
- **Quantum computation:** There also has been a serious effort toward building connections between quantum computation and DP [1, 24, 39, 46]. Some works build the bridge between the two via *quantum information theory* that draws Shannon information theory, quantum mechanics, and computer science together. Quantum DP is originally defined in [46] for adaptation of DP to quantum information processing. [24] focuses on quantum DPy using an information-theoretic framework, which is translated into quantum divergence.

Outline: Section 2 provides necessary preliminaries from the literature on DP. Introductory preliminaries are followed by novel metrics derived through information-theoretic measures for quantifying privacy guarantee of differentially private mechanisms in Section 3 along with their ordering and comparisons. Section 4 presents upper bounds on information cost and maximal leakage based on Shannon entropy as well as min-entropy in differentially private mechanisms. In Section 5, we discuss the connections between DP and source-coding theory, in addition to an exemplary result on adversarial classification in differentially private mechanisms from a rate-distortion perspective. To conclude, in Section 6, we point out possible research directions on information-theoretic approaches to DP for future work.

2 PRELIMINARIES

This section is reserved for a review of some important preliminaries from the DPy literature. We begin with defining the notion of neighborhood of datasets and the sensitivity of DP.

Definition 2.1. Two datasets x and $\tilde{x}$ are called neighbors, if the following equality holds:

$$d(x, \tilde{x}) = 1, \tag{1}$$

where $d(.,.)$ denotes the Hamming or l_1 distance between the datasets [16].

Definition 2.1 considers symmetry among neighbors in terms of the size of the dataset as depicted in Figure 2. This is further relaxed to include the datasets, where neighborhood is due to the addition or removal of a record as shown in Figure 1. In both cases, neighbors differ in a single row.

Definition 2.2. Global sensitivity, denoted by s, of a function (or a query) $q : D \rightarrow \mathbb{R}^k$ is the smallest possible upper bound on the distance between the images of q when applied to two neighboring datasets x and $\tilde{x}$. This means that the l_1 distance is bounded as follows: $\|q(x) - q(\tilde{x})\|_1 \leq s$ [15] .

Basically, sensitivity of a differentially private mechanism of Definition 2.2 is the tightest upper bound on the images of a query (a mapping function) for neighbors. It is a function of the type of the query having an opposite relationship with the privacy, since higher sensitivity of the query refers to a stronger requirement for privacy guarantee; consequently more noise is needed to achieve that guarantee.

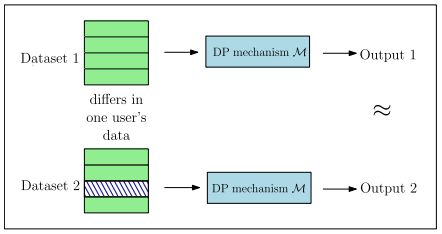

Fig. 2. Symmetric neighborhood of differential privacy.

The original definition of DP makes use of this notion of neighborhood between datasets. An informal definition is depicted in Figure 2. Accordingly, a mechanism $\mathcal{M}$ is said to be differentially private if for any two neighboring datasets, corresponding outputs of the mechanism, Outputs 1 and 2, are indistinguishable. In other words, the output of a differentially private mechanism is expected to behave in the same way whether or not one contributes to the dataset with their data. The following formal definition of DP introduced and studied by Dwork et al. in various publications [13, 14, 16] clarifies the mathematical meaning of indistinguishability of the outputs corresponding to neighboring datasets.

Definition 2.3 ((ϵ, δ)-DP). A randomized algorithm $\mathcal{M}$ is (ϵ, δ)-differentially private if $\forall S \subseteq Range(\mathcal{M})$ and $\forall x, \tilde{x}$ that are neighbors within the domain of $\mathcal{M}$, the following inequality holds.

$$\Pr[\mathcal{M}(x) \in S] \leq \Pr[\mathcal{M}(\tilde{x}) \in S]\, e^{\epsilon} + \delta. \tag{2}$$

For two DP measures ϵ_1-DP and ϵ_2-DP where $\epsilon_1, \epsilon_2 > 0$, ϵ_1-DP $\geq \epsilon_2$-DP denotes that ϵ_1-DP is a stronger privacy metric than ϵ_2-DP. Analogous to Definition 2.3, there are two other cases of DP where either of the privacy parameters, ϵ or δ, equals zero. The ordering of these three cases from the strongest to the weakest privacy metric is as follows:

$$\epsilon\text{-DP} \geq (\epsilon, \delta)\text{-DP} \geq \delta\text{-DP}. \tag{3}$$

Dwork's original definition of DP in Definition 2.3 emanates from a notion of statistical indistinguishability of two different probability distributions given by the next definition.

Definition 2.4 (Statistical Closeness). Two probability distributions P_1 and P_2 are said to be (ϵ, δ)-close denoted by $P_1 \overset{(\epsilon,\delta)}{\approx} P_2$ over the measurable space $(\Omega, \mathcal{F})$ iff the following inequalities hold.

$$P_1(A) \leq e^{\epsilon} P_2(A) + \delta, \quad \forall A \in \mathcal{F}, \tag{4}$$

$$P_2(A) \leq e^{\epsilon} P_1(A) + \delta, \quad \forall A \in \mathcal{F}. \tag{5}$$

Some important properties of statistical closeness are recalled here that will be used in Section 3 to prove equality between mutual information functional and DP.

— Property 1: Statistical closeness has the following relation with KL divergence.

$$P_1 \overset{(\epsilon,0)}{\approx} P_2 \implies \begin{matrix} D(P_1||P_2) \leq \min\{\epsilon, \epsilon^2\} \\ D(P_2||P_1) \leq \min\{\epsilon, \epsilon^2\}. \end{matrix} \tag{6}$$

Note that, the right-hand sides of the inequalities are given in nats.

— Property 2: Due to Pinsker's inequality, we also have

$$D(P_1||P_2) \leq \epsilon \text{ nats} \implies P_1 \overset{(0,\sqrt{\epsilon/2})}{\approx} P_2. \tag{7}$$

— Property 3: For any $\epsilon' < \epsilon$ and $\delta' = 1 - \frac{(e^{\epsilon'}+1)(1-\delta)}{e^{\epsilon}+1}$, we have the following relation.

$$P_1 \overset{(\epsilon,\delta)}{\approx} P_2 \implies P_1 \overset{(\epsilon',\delta')}{\approx} P_2. \tag{8}$$

2.1 How to Obtain ϵ- and (ϵ, δ)-DP?

A differentially private mechanism is named after the probability distribution of the perturbation applied onto the query output, in the global setting. In the following, we remind the reader of the Laplace distribution and introduce Laplace and Gaussian mechanisms. The Laplace distribution, also known as the double exponential distribution, with location parameter μ and scale parameter b is defined by

$$Lap(x; \mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}, \tag{9}$$

where its mean equals its location parameter μ and its variance is $2b^2$.

Definition 2.5. Laplace mechanism [15] for a function (or a query) $q : D \rightarrow \mathbb{R}^k$ is defined by

$$\mathcal{M}(x, q(.), \epsilon) = q(x) + (Z_1, \ldots, Z_k), \tag{10}$$

where $Z_i \sim Lap(b = s/\epsilon)$, $i = 1, \ldots, k$ denote i.i.d. Laplace random variables.

Definition 2.6. Gaussian mechanism [15] is defined for a function (or a query) $q : D \rightarrow \mathbb{R}^k$ as follows:

$$\mathcal{M}(x, q(.), \epsilon, \delta) = q(x) + (Z_1, \ldots, Z_k), \tag{11}$$

where $Z_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, k$ denote i.i.d. Gaussian random variables with the variance $\sigma^2 = \frac{2s^2 \log(1.25/\delta)}{\epsilon^2}$.

Theorem 2.7 ([16]). *For any $\epsilon, \delta \in (0, 1)$, the Gaussian mechanism satisfies (ε, δ)-DP.*

Remark. As an alternative to Laplacian perturbation applied on the query output that results in $(\epsilon, 0)$-DP, Gaussian noise provides a more relaxed privacy guarantee, that is, (ϵ, δ)-DP. However, in some cases, the application of Gaussian noise becomes more useful. Vector-valued Laplace mechanisms require the use of l_1 sensitivity whereas the vector-valued Gaussian mechanism allows l_1 or l_2 sensitivity, where l_2 sensitivity is defined as $\max_{x,\tilde{x}} \|q(x) - q(\tilde{x})\|_2 \leq s$, for neighboring x and $\tilde{x}$. Dependent on the query function, when the l_2 sensitivity is significantly lower than the l_1 sensitivity, the Gaussian mechanism requires much less noise.

Remark (The Optimal ϵ-differentially Private Mechanism). A natural question that comes to mind is if we can do better than the Laplace mechanism. The work in [23] improves the Laplace mechanism of [15] by characterizing the fundamental tradeoff between the differentially private mechanism's privacy and utility to define an *optimal* ϵ mechanism. Accordingly, [23, Theorem 1] shows that such a mechanism is obtained by applying a staircase-shaped probability distribution as the perturbation on real and integer-valued query functions in the low-privacy regime (i.e., when ϵ is large). The Laplace mechanism outperforms the optimal $(\epsilon, 0)$ mechanism in the high-privacy regime.

3 SHANNON INFORMATION AND RELATIVE ENTROPY AS A PRIVACY CONSTRAINT

This first main part of the tutorial is dedicated to the presentation of information-theoretic quantities adapted to be used as privacy constraint in systems that deploy (ϵ, δ)-differential privacy.

Definition 3.1 (ϵ–Mutual Information–DP [9]). For a dataset $X^n = (X_1, \ldots, X_n)$ with the corresponding ML output Y according to the randomized mechanism represented by $\mathcal{M} = P_{Y|X^n}$, ***mutual information differential privacy*** (**MI-DP**) is defined as

$$\sup_{i, P_{X^n}} I(X_i; Y|X^{-i}) \leq \epsilon \text{ nats}, \tag{12}$$

where $X^{-i} = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$ denotes the dataset entries excluding X_i.

The ϵ–MI-DP definition of Cuff et al. in [9] combines the Shannon information with the notion of *identifiability*, which is defined using the Bayesian approach on indistinguishability of the neighboring datasets. Accordingly, a mechanism $\mathcal{M}$ satisfies ϵ **identifiability** for some positive and real ϵ if the following inequality holds for any neighboring entries $x, \tilde{x} \in \mathcal{D}^n$ and any output $y \in \mathcal{D}^n$.

$$P_{X|Y}(x|y) \leq \mathrm{e}^{\epsilon} P_{X|Y}(\tilde{x}|y). \tag{13}$$

Both ϵ–MI-DP and ϵ identifiability are subject to the implicit strong adversary assumption [9] (also called the *informed adversary* in [15]) where the adversary has the knowledge of all but a single entry in a dataset and aims to discover the last one. The condition in Equation (13) suggests that for small values of ϵ, neighboring datasets are indistinguishable based on the posterior probabilities of the output. This is what makes it hard to associate the representation of the data and the data owner, which translates to re-identification. Another line of work in [43] defines the MI-based differential privacy as a lossy source-coding problem without the maximization taken over all possible dataset distributions. Definition 3.1 differs from the information-theoretic definitions of original DP by incorporating that no assumptions are made on prior dataset distributions. Maximization over all possible input distributions in Equation (12) assures that the DP is a property of the mechanism resembling the well-known formula of the Shannon capacity.

Next, we remind the reader of the so-called *KL DP*.

Definition 3.2 (ϵ-KL DP [9]). A randomized mechanism $P_{Y|X}$ guarantees ϵ–KL DP, if the following inequality holds for all its neighboring datasets x and $\tilde{x}$,

$$D(P_{Y|X=x}||P_{Y|X=\tilde{x}}) \leq \mathrm{e}^{\epsilon}. \tag{14}$$

3.1 Main Result

Using information-theoretic quantities to study privacy may not be a brand new approach, nonetheless, the following result draws the strongest link between the two areas. Ordering and equivalence of ϵ–MI-DP and DP is given by Theorem 3.3.

Theorem 3.3 ([9]). *The following chain of inequalities holds:*

$$\epsilon\text{-DP} \geq \epsilon\text{–MI-DP} \geq (\epsilon, \delta)\text{-DP}. \tag{15}$$

Conditioned on the cardinality of the input $\mathcal{X}_i$ or the output $\mathcal{Y}$ of the differentially private mechanisms, an equivalence is achieved between ϵ–MI-DP and (ϵ, δ)-DP. Then, we have

$$\epsilon\text{–MI-DP} = (\epsilon, \delta)\text{-DP}. \tag{16}$$

The case (ϵ, δ)-DP $\geq \epsilon$–MI-DP depends on the cardinality bound $\min\{|\mathcal{Y}|, \max_i |\mathcal{X}_i|\}$.

The sketch of the proof of Theorem 3.3 [9]. As is well known, (un/conditional) MI can be represented as a function of relative entropy. The proof of Theorem 3.3 starts off by proving an even more powerful chain of in/equalities among all three variations ϵ-, (ϵ, δ)-, and δ-DP, MI-DP, and the KL DP. The chain of inequalities in Equation (15) is expanded out as follows.

$$\epsilon\text{-DP} \overset{(a)}{\geq} \text{KL-DP} \overset{(b)}{\geq} \epsilon\text{-MI-DP} \overset{(c)}{\geq} \delta\text{-DP} \overset{(d)}{=} (\epsilon, \delta)\text{-DP}. \tag{17}$$

Equation (17) shows that an ϵ-DP mechanism also guarantees ϵ–MI-DP. Relations (a) and (b) in Equation (17) are the results of Property 1 of statistical closeness given by Definition 2.4. Ordering

in (b) is achieved as follows:

$$D(P_{Y|X^n=x^n}||P_{Y|X^{-i}=x^{-i}}) = D\left(P_{Y|X^n=x^n}||\mathbb{E}\left[P_{Y|X_i=\tilde{X},X^{-i}=x^{-i}}\right]\right) \tag{18}$$

$$\leq \mathbb{E}\left[D(P_{Y|X^n=x^n}||P_{Y|X^{-i}=\tilde{X},X^{-i}=x^{-i}})\right] \tag{19}$$

$$\leq \epsilon \text{ nats} \tag{20}$$

for $\tilde{X} \sim P_{X_i|X^{-i}=x^{-i}}$ and x^{-i} denotes an instance of X^{-i}. Thus, in Equation (18) we use $P_{Y|X^{-i}=x^{-i}} = \mathbb{E}[P_{Y|X_i=\tilde{X},X^{-i}=x^{-i}}]$. The steps in Equations (19) and (20), respectively follow due to Jensen's inequality and the definition of MI based on relative entropy, that is,

$$I(X_i;Y|X^{-i}) = \mathbb{E}\left[D(P_{Y|X^n=\bar{X}^n}||P_{Y|X^{-i}=\bar{X}^{-i}})\right] \tag{21}$$

where $\bar{X}^n \sim P_{X^n}$. Ordering (c) that states ϵ–MI-DP $\geq$ δ-DP is a consequence of Lemma 3.4.

Lemma 3.4 ([9]). *The following statement is satisfied with respect to the relation between ϵ–MI-DP and (δ)-DP.*

$$\epsilon\text{–MI-DP} \implies (0, \sqrt{2\epsilon})\text{-DP}. \tag{22}$$

Equation (22) is tightened as ϵ–MI-DP $\implies (0,\delta')$-DP for $\epsilon \in [0, \ln 2]$ for $\delta' = 1 - 2h^{-1}(\ln 2 - \epsilon)$ and h^{-1} denotes the inverse of the binary entropy function.

Lastly, ordering (d) in Equation (17) is due to Property 3 given by Definition 2.4. The reader is referred to [9, Section 3.3] for the full proof.

Remark. The major strength of ϵ–MI-DP over other alternative MI-based definitions of DP lies in **the maximization taken over all possible input distributions to capture the fact that differential privacy does not require a particular distribution of the input.** Moreover, from a stochastic perspective, conditional MI reflects the strong adversary assumption of DP and establishes another major strength of ϵ–MI-DP that is based on Dwork's standard definition of differential privacy that originally stems from this assumption. Conditioning on the remaining entries of the dataset in ϵ–MI-DP demonstrates that the adversary has the knowledge of the entire dataset except for one entry, which was transmitted implicitly by using the notion of neighboring datasets in the original stochastic definition of DP. From a practical point of view, another major strength of Definition 3.1 lies in the ability to transfer information-theoretic rules and properties defined for Shannon information and related measures onto DP.

Next part provides some of the well-known information-theoretic rules that also apply to DP as a consequence of MI-DP and the ordering in Equation (17).

3.2 Composability of ϵ–MI-DP via Information-Theoretic Rules

This part is dedicated to some of the well-known properties of MI that are now directly applicable on ϵ–MI-DP.

— Bounding the conditional MI: If X is independent of Z, then the following inequality holds.

$$I(X;Y|Z) \geq I(X;Y). \tag{23}$$

— Consequence of data processing inequality: If $X \rightarrow Y \rightarrow Z$ form a Markov chain in that order, that is, X and Z are conditionally independent given Y, then the following inequality holds.

$$I(X;Y|Z) \leq I(X;Y). \tag{24}$$

— Chain rule:

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z). \tag{25}$$

— Independence: If the differentially private mechanism $\mathcal{M} = P_{Y|X^n}$ satisfies ϵ–MI-DP where $\{X_i\}_{i=1}^n$ are mutually independent, then the following chain of inequalities hold.

$$\sup_{i,P_{X^n}} I(X_i;Y) \leq \sup_{i,P_{X^n}} I(X_i;Y|X^{-i}) \leq \epsilon. \tag{26}$$

Some of the fundamental rules of MI enlisted above are transferred onto DP as a result of Theorem 3.3. Several important properties of DP are straightforward to prove in this MI-based approach. Next, we prove the composition theorem of DP with the aid of these properties. Originally, composability—an important property of $(\epsilon, 0)$-DP—states that a number of queries under DP also collectively satisfies DP where the privacy budget of the collection is scaled proportionally to the number of queries [17, 25]. Corollary 3.5 is a reflection of the composition theorem for $(\epsilon, 0)$-DP onto ϵ–MI-DP, which shows that the composability can be defined and proven using information-theoretic quantities and their corresponding properties.

Corollary 3.5 (Composition of ϵ–MI-DP [9]). *For randomized mechanisms $\mathcal{M}_j = P_{Y_j|X^n}$ that individually satisfy ϵ–MI-DP with k conditionally independent outputs $\{Y_1, \ldots, Y_k\}$ given the input $\{X_1, \ldots, X_n\}$, the collection of k mechanisms $\mathcal{M}_k = P_{Y_k|X^n}$ also satisfies ϵ–MI-DP with the privacy parameter $\sum_j^k \epsilon_j$.*

Proof. For any P_{X^n} and i, the collection of $P_{Y_k|X^n}$ satisfies ϵ–MI-DP, which is bounded as follows:

$$I(X_i;Y|X^{-i}) = \sum_{l=1}^{m} I(X_i;Y_l|X^{-i},Y^{l-1}) \tag{27}$$

$$\leq \sum_{l=1}^{m} I(X_i;Y_l|X^{-i}). \tag{28}$$

Equation (27) follows due to the chain rule given by Property 3 in Section 3.2. The step in Equation (28) uses a property of the data-processing inequality (Property 2 in Section 3.2) due to the conditional independence between X_i and Y^{l-1} given Y_l. Finally, Equation (29) substitutes Definition 3.1 as given below.

$$I(X_i;Y|X^{-i}) \leq \sum_{l=1}^{m} \epsilon_j \text{ nats}. \tag{29}$$

□

This result completes the first main part of the tutorial.

4 INFORMATION-THEORETIC BOUNDS ON DP

In this section, we review four selective publications [2, 6, 12, 30] that present upper bounds on the performance of differentially private mechanisms using different metrics. We begin with the two-party differential privacy in the distributed setting in the upcoming part.

4.1 Bounding the Information Cost

Contrarily to the common client-server setting where the server answers queries of clients based on its access policy, in the two-party distributed setting parties execute their analysis on joint data where the aim is to provide a two-sided privacy guarantee for each party's data. In such a setting, each side sees the protocol/mechanism as a differentially private version of the other side's input data. Information cost of a two-user DP model in such a setting refers to the amount of information gathered from each party's inputs using the exchanged messages. In order to prove the usefulness

and practicality of DP, McGregor et al. characterize in [30] a fundamental connection between the information cost and DP. Accordingly, the authors present an upper bound on the information cost of such a mechanism by defining the cost as the MI between the inputs and the random transcript of the mechanism denoted $\Pi(.,.)$, which simply is the sequences of exchanged messages between the two parties.

Definition 4.1 (Information Cost). For two inputs X and Y of a two-party mechanism $\mathcal{M}$ with probability distribution P, the information cost of the mechanism is defined as

$$Icost_P(\mathcal{M}) = I(X, Y; \Pi(X, Y)). \tag{30}$$

For a finite alphabet Σ, the two-party ϵ-DP mechanism $\mathcal{M}(x, y)$ with $x, y \in \Sigma^n$ and every distribution P defined on $\Sigma^n \times \Sigma^n$, the information cost of this mechanism satisfies the upper bound

$$Icost_P(\mathcal{M}) \leq 3\epsilon n. \tag{31}$$

For the special case of $\Sigma = \{0, 1\}$ and P is the uniform distribution, the bound in Equation (31) is improved to $1.5\epsilon^2 n$ [30, Proposition 4.3].

Derivation of the upper bounds. For the two-party random input denoted by $T = (X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ and independent sample T' from the uniform distribution P, we have

$$\begin{aligned} I(\Pi(T); T) &= H(\Pi) - H(\Pi|T) \\ &= \mathbb{E}_{(t,\pi)\leftarrow(T,\Pi(T))} \log \frac{\Pr[\Pi[T] = \pi | T = t]}{\Pr[\Pi[T] = \pi]} && (32) \\ &\leq 2(\log_2 \mathrm{e})\epsilon n. && (33) \end{aligned}$$

Equation (33) is equivalent to the right-hand side of Equation (31) and obtained using the following interval for any t and t'.

$$\mathrm{e}^{(-2\epsilon n)} \leq \frac{\Pr[\Pi(t) = \pi]}{\Pr[\Pi(t') = \pi]} \leq \mathrm{e}^{(2\epsilon n)}. \tag{34}$$

The improvement is achieved by setting $\Sigma = \{0, 1\}$ for a uniform distribution P as follows.

$$\begin{aligned} I(T; \Pi(T)) &= \sum_{i\in[2n]} I(T_i; \Pi(T)|T_1 \cdots T_{i-1}) && (35) \\ &= \sum_{i\in[2n]} H(T_i|T_1 \cdots T_{i-1}) - H(T_i|\Pi(T)T_1 \cdots T_{i-1}) && (36) \\ &\leq \sum_{i\in[2n]} (1 - H(\mathrm{e}^{\epsilon}/2)) && (37) \\ &\leq \sum_{i\in[2n]} \frac{\epsilon^2}{2\ln 2}. && (38) \end{aligned}$$

The first term in Equation (36) equals 1 since each T_i is independent and uniform in P. Due to the DP property and the Bayes rule, we have $\forall t_1, \ldots, t_{i-1}, \pi$ the ratio confined in the interval $(\mathrm{e}^{-\epsilon}, \mathrm{e}^{\epsilon})$ as given by

$$\mathrm{e}^{-\epsilon} \leq \frac{\Pr[T_i = 0|T_1, \ldots, T_{i-1} = t_1, \ldots, t_{i-1}, \Pi[T] = \pi]}{\Pr[T_i = 1|T_1, \ldots, T_{i-1} = t_1, \ldots, t_{i-1}, \Pi[T] = \pi]} \leq \mathrm{e}^{\epsilon}. \tag{39}$$

Accordingly, the second term in Equation (36) is bounded by the entropy in Equation (37). Finally, in Equation (38), the base of the logarithm is changed and summed over $2n$ terms to get $\log_2(\mathrm{e})\epsilon^2 n$. [10] presents an adaptation of the upper bound in Equation (31) to the MI between the distribution over the inputs of an ϵ-differentially private mechanism and the mechanism's output by replacing

the second party's input with a constant to obtain the same behavior of $3\epsilon n$. Accordingly, for a query $q : (\mathbb{Z}^+)^d \to \mathbb{R}^k$, an ϵ-differentially private mechanism $\mathcal{M} : (\mathbb{Z}^+)^d \to P(\mathbb{R}^k)$ and a dataset size of n, the MI $I(X; \mathcal{M}(X))$ is upper bounded by $3\epsilon n$. Bounding the size of the dataset by n, allows the input distribution to be narrowed down to $X \in [n]^d$ for $[n] = \{0, 1, \ldots, n\}$. This results in the direct application of the upper bound (31) by McGregor et al. when the second party's input is set to be a constant.

Remark. Equation (31) bounds the information cost as a function of the privacy budget of a DP mechanism and combined with [5], the result signifies that any mechanism that satisfies DP can be compressed. Additionally, well-known bounds for the information cost in various settings can be employed to characterize the gap between the optimal and computational DP mechanisms.

4.2 Upper Bound on Maximal Leakage

[12] is one of the first examples of the line of work that modeled the problem of defining the optimal mapping of the input data to a privatized output in order to determine the privacy-utility tradeoff by using rate-distortion theory. Additionally, the authors compare DP with the maximum information leakage to prove that DP does not grant privacy with regard to average and maximal leakage. Their model is designed as a noiseless communication channel between two parties to transmit a number of measurements denoted $Y \in \mathcal{Y}$ to the receiving end, as well as a set of variables $X \in \mathcal{X}$ that is required to remain private to the sender. X and Y follow the joint distribution $(Y, X) \sim p_{Y,X}(y, x),\ (y, x) \in \mathcal{Y} \times \mathcal{X}$.

ϵ-information privacy is defined as follows in the sense of a differentially private mechanism as a stronger alternative to Dwork's original definition. Accordingly, ϵ-information privacy captures the fundamental aim of privacy of resisting to notable change in the conditional prior and posterior probabilities of the features given the output.

Definition 4.2 ([21]). A privacy-preserving mapping defined by the transition probability $p_{Y|\mathbf{X}}(.|.)$ for a set of features $\mathbf{X} = (X_1, \ldots, X_n)$ where $X_i \in \mathcal{X},\ y \in \mathcal{Y}$ provides ϵ-DP

$$e^{-\epsilon} \leq \frac{p_{\mathbf{X}|Y}(\mathbf{x}|y)}{p_{\mathbf{X}}(\mathbf{x})} \leq e^{\epsilon} \tag{40}$$

for all $y \in \mathcal{Y} : p_Y(y) > 0$ if $\forall x \subseteq \mathcal{X}^n$.

Definition 4.2 is used for bounding the *maximal (information) leakage* defined by

$$\max_{y \in \mathcal{Y}} H(X) - H(X|Y = y). \tag{41}$$

Maximal leakage refers to the maximum cost gain achieved by the adversary using a single output. The main result of [12] connecting ϵ-information privacy to DP is given by the next theorem.

Theorem 4.3 (Upper Bound on Maximal Leakage of Differential Privacy [12]). *If a privacy-preserving mapping $p_{Y|\mathbf{X}}(.|.)$ is ϵ-information private for some supp$(p_Y) = \mathcal{Y}$, then it provides at least 2ϵ-DP and the maximal leakage is at most $\frac{\epsilon}{\ln 2}$.*

Proof. For neighbors $\mathbf{x}_1$ and $\mathbf{x}_2$, we have for $p_{Y|\mathbf{X}}(.|.)$ and a subset $B \subseteq \mathcal{Y}$

$$\frac{\Pr[Y \in B|\mathbf{X} = \mathbf{x}_1]}{\Pr[Y \in B|\mathbf{X} = \mathbf{x}_2]} = \frac{\Pr[\mathbf{X} = \mathbf{x}_1|Y \in B]\Pr[\mathbf{X} = \mathbf{x}_2]}{\Pr[\mathbf{X} = \mathbf{x}_2|Y \in B]\Pr[\mathbf{X} = \mathbf{x}_1]} \tag{42}$$

$$\leq e^{2\epsilon}. \tag{43}$$

The bounding step in Equation (43) is a result of Definition 2.3. The maximum amount of information that is leaked from ϵ-information private mapping Equation (41) is bounded as given below.

$$H(X) - H(X|Y=y) = \sum_{\mathbf{x}\in\mathcal{X}^n} p_{\mathbf{X}|Y}(\mathbf{x}|y)p_Y(y)\log\left(\frac{p_{\mathbf{X}|Y}(\mathbf{x}|y)}{p_{\mathbf{X}}(\mathbf{x})}\right) \tag{44}$$

$$\overset{(i)}{\leq} \sum_{\mathbf{x}\in\mathcal{X}^n, y\in\mathcal{Y}} p_{\mathbf{X}|Y}(\mathbf{x}|y)p_Y(y)\log \mathrm{e}^{\epsilon} \tag{45}$$

$$\overset{(ii)}{=} \frac{\ln \mathrm{e}^{\epsilon}}{\ln 2}. \tag{46}$$

Step (i) results from applying the upper bound of Definition 4.2 and from changing the range of the sum. In step (ii), the base of the logarithmic function is changed and the summation equals to 1, thus we get $\epsilon/\ln 2$. □

4.3 Upper Bound on Maximal Leakage Based on Min-Entropy

This part presents the review of an upper bound on the maximal leakage of ϵ-DP by [6]. The distinction of the work stems from using *min-entropy* rather than Shannon entropy. The ultimate goal of [6] is to compare and formally characterize connections between DP and information-theoretic leakage. The main contribution is establishing such a connection by upper bounding the information leakage in terms of DPy as a function of the privacy budget.

[6] justifies the use of min-entropy by its association to strong security guarantees. For X and Y, respectively denoting the input and output to a probabilistic program and the conditional distribution, $P_{Y|X}$ is characterized by the program's semantics and composes an information-theoretic channel between X and Y. In this setting, the adversary aims to infer the value of X upon reception of the output Y. The unconditional min-entropy $H_\infty(X)$ of X is defined by

$$H_\infty(X) = -\log \max_x P_X(x), \tag{47}$$

whereas the conditional min-entropy $H_\infty(Y|X)$ of $P_{Y|X}$ yields

$$H_\infty(Y|X) = -\log \sum_y P_Y(y) \max_x P_{X|Y}(x,y). \tag{48}$$

The min-entropy–based leakage denoted by L is the difference between $H_\infty(X)$ and $H_\infty(Y|X)$ depending on both the channel $P_{Y|X}$ and the input distribution P_X. Min-entropy–based maximal leakage $ML(P_{Y|X})$ is given by

$$ML(P_{Y|X}) = \max_{P_X}(H_\infty(X) - H_\infty(Y|X)). \tag{49}$$

For channels of a single bit of range, that is, when $Range(X) = Range(Y) = \{0,1\}$, [6, Theorem 3] states that for an ϵ-differentially private channel $P_{Y|X}$, the maximal leakage is upper bounded by

$$ML(P_{Y|X}) \leq \log \frac{2\mathrm{e}^{\epsilon}}{1+\mathrm{e}^{\epsilon}}. \tag{50}$$

The bound in Equation (50) is proven to apply to channels of arbitrary finite range in [6, Corollary 1]. Accordingly, the channel $P_{Y|X}$ is summarized in Table 1 for $\sum_i^n p_i = \sum_i^n q_i = 1$.

For an ϵ-differentially private channel $P_{\bar{Y}|X}$, where the output $\bar{Y}$ is defined over the range $\{0,1\}$, the leakage of $P_{Y|X}$ and that of $P_{\bar{Y}|X}$ coincide. Similarly, for the channel $P_{\bar{Y}|X}$ we have the following matrix of probabilities for $I = \{i|p_i \leq q_i\}$. In Table 2, $\bar{p}$ and $\bar{q}$ respectively denote the sums over I as $\sum_{i\notin I} p_i$ and $\sum_{i\in I} q_i$. Hence, their respective complements yield $1-\bar{p} = \sum_{i\in I} p_i$ and

Table 1. The Channel $P_{Y|X}$ with $X = \{0, 1\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$

$P_{Y\|X}$	$Y = y_1$	$\cdots$	$Y = y_n$
$X = 0$	p_1	$\cdots$	p_n
$X = 1$	q_1	$\cdots$	q_n

Table 2. The Channel $P_{\bar{Y}|X}$ with $X, \bar{Y} = \{0, 1\}$

$P_{\bar{Y}\|X}$	$\bar{Y} = 0$	$\bar{Y} = 1$
$X = 0$	$\bar{p}$	$1 - \bar{p}$
$X = 1$	$\bar{q}$	$1 - \bar{q}$

$1 - \bar{q} = \sum_{i \notin I} q_i$. Plugging in [6, Theorem 3] with the definition of min-entropy–based maximal leakage, the equivalence of $ML(P_{Y|X})$ and $ML(P_{\bar{Y}|X})$ is proven by

$$ML(P_{Y|X}) = \log \sum_y \max_x P_{Y|X}(y, x) \tag{51}$$

$$= \log(\bar{p} + \bar{q}) \tag{52}$$

since Equation (52) is $ML(P_{\bar{Y}|X})$.

Additionally, ϵ-DP of the channel $P_{Y|X}$ guarantees that $q_i \leq \mathrm{e}^{\epsilon}$ for every $i \in I$ and thus, $\bar{q} \leq \mathrm{e}^{\epsilon}\bar{p}$. The same applies to $p_i \leq \mathrm{e}^{\epsilon}$ for every $i \notin I$.

4.4 Information-Theoretic Post-processing of Differential Privacy

The post-processing property is one of the important features of DP and ensures that the privacy protection of a differentially private mechanism is not affected by arbitrary computations applied on the mechanism's output [16]. In other words, it is impossible to *undo* the privacy guarantee of DP by post-processing the data. More formally, if the mechanism $\mathcal{M} : \mathbb{N}^{|\mathcal{X}|} \to R$ satisfies (ϵ, δ)-DP, for any arbitrary mapping $f : R \to R'$, $f \circ \mathcal{M} : \mathbb{N}^{|\mathcal{X}|} \to R'$ also satisfies (ϵ, δ)-DP.

A simpler version of the problem of investigating the connection between DP and min-entropy leakage in [6] is initiated by [2, 3] for an individual rather than the entire universe of databases. In [2, 3], the authors consider a model where information leakage is used to measure the amount of information that an attacker can learn about the database that also allows one to quantify the utility of the query via min-entropy. Applying Bayesian post-processing on the differentially private output of the mechanism, it is shown that the utility function is closely related to conditional min-entropy and to the min-entropy leakage.

5 DP AS A SOURCE-CODING PROBLEM

Several works study the connection between DP and (lossy) source coding from various aspects [12, 32, 34, 36, 43, 47] and some tailored the rate-distortion theory to identify a tradeoff between privacy and distortion. [12] is one of the first examples that model DP using a rate-distortion perspective establishing a tradeoff between privacy and utility. The authors set the amount of information obtained by the adversary (i.e., the leakage) as the cost gain and minimize it subject to a set of utility constraints, which reflect the role of the distortion function in the original setting of the rate-distortion theory. On the other hand, in [43], the distortion between the input and output of the mechanism is used to determine the number of rows that differ and it is minimized subject to three different privacy metrics, in order to establish how many rows need to be modified to

preserve the privacy guarantee. Accordingly, the distortion is defined as the Hamming distance d between the input and output of a dataset as $d : \mathcal{D}^n \times \mathcal{D}^n \rightarrow \mathbb{N}$. The contribution of [43] is to demonstrate a connection between identifiability, DP, and *the* ***mutual-information privacy*** (**MIP**) that is defined by $I(X;Y)$ for the input X and output Y. The privacy-distortion problem of [43] is defined as follows.

$$\min_{P_{Y|X}} I(X;Y), \tag{53}$$

$$\text{s.t. } \mathbb{E}[d(X,Y)] \leq D, \tag{54}$$

$$\sum_{y \in \mathcal{D}^n} p_{Y|X}(y|x) = 1, \forall x \in \mathcal{D}^n, \tag{55}$$

$$p_{Y|X}(y|x) \geq 0, \forall x, y \in \mathcal{D}^n. \tag{56}$$

The main objective of [43] is to investigate and explain the relation between identifiability, DP, and MIP in order to compare them. The authors show that there exists a privacy mechanism that minimizes both $I(X;Y)$ and the identifiability. The **privacy-distortion function** denoted as $\epsilon^*(D)$ refers to the smallest DP level for a given maximum allowable distortion D. The MI-based privacy level is bounded as follows:

$$\epsilon^*(D) \leq \epsilon \leq \epsilon^*(D) + 2\epsilon_X, \tag{57}$$

where the maximal prior probability difference is

$$\epsilon_X = \max_{x, \tilde{x} \in \mathcal{D}^n : x \sim \tilde{x}} \ln \frac{p_X(x)}{p_X(\tilde{x})} \tag{58}$$

for neighboring datasets x and $\tilde{x}$. This MI-based mechanism satisfies ϵ-DP. In light of [9, Theorem 1], which is visited in Section 3, the exact relation and ordering between conditional MI and DP are today known.

[36] studies the convergence of the source distribution estimate to the actual distribution based on the output from a locally differentially private mechanism. The fundamental difference in this setting stems from the fact that the DP noise that is applied on each user's data locally, removes the requirement for a notion of neighborhood between datasets. In the model of [36], the source $\{X_i\}$ follows a discrete distribution P and the mechanism $\mathcal{M}$ refers to the application of local DP noise on n i.i.d. source symbols that outputs the privatized observations $\{Y_i\}$ following the distribution Q, that is, $P\mathcal{M}$. The goal of the legitimate observer is to estimate the source distribution P using the noisy outputs $\{Y_i\}$ subject to either of f-divergence, **mean-squared error** (**MSE**), or total variation as the fidelity criteria. At the same time, an adversary aims to discover some source samples X_i. The authors present upper and lower bounds on their formulation of the tradeoff between DP level and fidelity loss based on the aforementioned three loss functions.

5.1 An Adaptation to Adversarial Classification

Introducing adversarial examples to ML systems is a specific type of sophisticated and powerful attack, whereby additional (sometimes specially crafted) or modified inputs are provided to the system with the intent of being misclassified by the model as legitimate. Adversarial classification is one possible defense proposed to correctly detect adversarial examples that aim to fool the classifier that detects outliers. In [41, 42], DP is weaponized by the adversary in order to ensure to remain undetected. In addition to the statistical approach using hypothesis testing to establish a threshold of detection for the adversary as a function of the privacy budget, [42] also introduces an original adaptation of lossy source coding to upper bound the impact of the attack.

In this setting, the adversary not only wants to discover the data but also aims to harm the differentially private mechanism by modifying the released information without being detected.

This tradeoff between two conflicting goals of adversary is remodeled via the rate-distortion theory balancing the adversary's advantage and the security of the Gaussian DP mechanism. Accordingly, the MI between the input and output of a communication channel in the original rate-distortion problem is now replaced by the datasets before and after the alteration applied by the adversary that are considered as neighbors, where the absolute difference between the two corresponds to the impact of the attack. Neighboring input vectors $X^n = \{X_1, \ldots, X_n\}$ and $\tilde{X}^n = \{X_1, \ldots, X_i, \ldots, X_n + X_{adv}\}$ are assumed to be i.i.d. following the Gaussian distribution with the parameters $\mathcal{N}(0, \sigma^2_{X_i})$ with the difference of a single record denoted $X_{adv} \sim \mathcal{N}(0, \sigma^2_{adv})$. The query function takes the aggregation of this dataset as $q(\mathbf{X}) = \sum_i^n X_i$ and the DP mechanism adds Gaussian noise Z on the query output leading to the noisy output in the following form, $\mathcal{M}(\mathbf{X}, q(.), \epsilon, \delta) = Y = \sum_i^n X_i + Z$. An adversary adds a single record denoted X_{adv} to this dataset. The modified output of the DP mechanism becomes $\sum_i^n X_i + X_{adv} + Z$.

Theorem 5.1. *The privacy-distortion function for a dataset X^n and Gaussian mechanism as defined by Definition (2.6) is*

$$P(s) = \frac{1}{2} \log \left(f_n \left(1 + \prod_i^n \sigma^2_{X_i} / s^2 \right) \right), \tag{59}$$

for $s \in [0, \prod_i^n \sigma^2_{X_i}]$ and zero elsewhere. σ_{X_i} denotes the standard deviation of X_i for $i = 1, \ldots, n$, and f_n is some constant dependent on the size of the dataset n.

The sketch of the proof proceeds as follows. The MI between the datasets before and after the attack is derived as follows:

$$I(X^n; \tilde{X}^n) = h(\tilde{X}^n) - h(\tilde{X}^n | X^n) \tag{60}$$

$$\geq \frac{1}{2} \sum_{i=1}^{n} \log \left((2\pi \mathrm{e}) \, \sigma^2_{X_i} \right) - \frac{1}{2} \log \left(2\pi \mathrm{e} s^2 \right) \tag{61}$$

$$= \frac{1}{2} \log \left((2\pi \mathrm{e})^{n-1} \prod_i^n \sigma^2_{X_i} / s^2 \right). \tag{62}$$

Corollary 5.2. *The second-order statistics of the additional data inserted into the dataset by the adversary is upper bounded as follows:*

$$\sigma^2_{X_{adv}} \leq \frac{1}{(2\pi \mathrm{e})^{n-1}} \left[\frac{s^2}{1 - s^2 / \sigma^2_{X_n}} \right] \tag{63}$$

for $s^2 = \frac{\sigma_z^2 \epsilon^2}{2 \log(1.25/\delta)}$ and $n \geq 2$.

We have the following considering the neighbor that includes X_{adv} has now $(n+1)$ entries over n rows as $\tilde{X}^n = \{X_1, X_2, \ldots, X_n + X_{adv}\}$. Accordingly, the second expansion is derived on X^n as

$$I(X^n; \tilde{X}^n) = h(X^n) - h(X^n | \tilde{X}^n) \tag{64}$$

$$\leq \sum_{i=1}^{n} \frac{1}{2} \log \left(2\pi \mathrm{e} \right)^n \sigma^2_{X_i} - \frac{1}{2} \log \left((2\pi \mathrm{e})^n \sigma^2_{X_{adv}} \right) \tag{65}$$

$$\leq \frac{1}{2} \log \prod_{i=1}^{n-1} \sigma^2_{X_i} \left(1 + \frac{\sigma^2_{X_n}}{\sigma^2_{X_{adv}}} \right) \tag{66}$$

leading to the upper bound in Equation (63). Due to the adversary's attack, in the first term of Equation (65), we add up the variances of $(n + 1)$ X_i's including X_{adv}. Since Equation (62) $\geq$

Equation (66), we obtain the upper bound in Corollary 5.2 For the detailed derivation of Equations (62) and (66), the reader is referred to [42].

Remark. The second expansion of the MI between neighboring datasets derived in Equation (66), can be related to the well-known **rate-distortion function of the Gaussian source** that, originally, provides the minimum possible transmission rate for a given distortion balancing (mostly for the Gaussian case) the squared-error distortion with the source variance. Combining Equation (62) with Equation (66) characterizes the privacy-distortion tradeoff of the Gaussian mechanism and bounds the impact of the adversary's modification on the original data in order to avoid detection in some sense *calibrating* the adversary's attack to the sensitivity of the differentially private mechanism.

6 WHAT *ELSE* DO WE WANT TO LEARN?

As convenient and practical as it is, using information-theoretic quantities as privacy constraint is not fully exploited. This final part is reserved for concluding the tutorial by pointing out possible research directions on information-theoretic approaches to DP for the future.

In particular, for classification of adversarial examples in differentially private mechanisms where adversaries may seek for ways to harm the systems via modifying the ML model and misclassifying to model inputs, the source-coding theory could provide new insights in the DP measure itself. A great majority of the existing information theory literature benefits from source-coding theory for quantifying the privacy guarantee or for determining the leakage as already mentioned in Section 5. [36] stands out in the way the rate-distortion perspective is translated for DP where various fidelity criteria is set to determine how fast the empirical distribution converges to the actual source distribution. This approach could be extended for detection of adversarial examples attacking differentially private mechanisms beyond the work [41], where the authors presented an application of the KL DP for detecting misclassification attacks in Laplace mechanisms. The corresponding distributions of relative entropy are considered as the differentially private noise with and without the adversary's advantage. The essential distinction that has to be made as relating DP to MI is that the mutual information requires an input distribution. DP, on the other hand, is a characteristic of the mapping function applied on the input. Consequently, the query mechanism, hence the sensitivity, should play a role in defining the fidelity criterion as translating the adversarial classification into a rate-distortion problem similarly to [42]. Ultimately, this approach inspired by rate-distortion theory could be generalized beyond misclassification attacks for various types of attacks in order to determine and manipulate limits of the impact and detection probability of an attack, and to formally characterize a tradeoff between the two. Moreover, by casting DP for adversarial classification into a source-coding problem, information-theoretic tools could be used to construct *explicit coding strategies* for privacy preservation in anomaly detection.

Furthermore, information-theoretic quantities could shed light on connections between DP and other trustworthy features of ML algorithms such as fairness and robustness. Various works show pairwise connections of DP with robustness [11, 26, 37, 38] and fairness [22]. The knowledge on these relations of DP with these properties are yet to be explored from an information-theoretic perspective.

REFERENCES

[1] Scott Aaronson and Guy N. Rothblum. 2019. Gentle measurement of quantum states and differential privacy. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (STOC'19)*. Association for Computing Machinery, New York, NY, 322–333. https://doi.org/10.1145/3313276.3316378

[2] Mário S. Alvim, Miguel E. Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. 2012. Differential privacy: On the trade-off between utility and information leakage. *Formal Aspects of Security and Trust*. Springer, Berlin, 39–54.

[3] Mário S. Alvim, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. 2010. Differential privacy versus quantitative information flow. arXiv:1012.4250. https://arxiv.org/abs/1012.4250
[4] Z. Bar Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. 2004. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences* 68, 4 (June 2004), 702–732.
[5] B. Barak, M. Braverman, X. Chen, and A. Rao. 2010. How to compress interactive communication. In *42nd ACM Symposium on Theory of Computing*. ACM, New York, NY, 67–76.
[6] G. Barthe and B. Köpf. 2011. Information-theoretic bounds for differentially private mechanisms. In *Computer Security Foundations Symposium*. IEEE, New York, NY, 191–204.
[7] M. Bellare, S. Tessaro, and A. Vardy. 2012. Semantic security for the wiretap channel. *Advances in Cryptology-CRYPTO*. Springer, Berlin, 294–311.
[8] A. Chawla. 2021. Pegasus Spyware—'A Privacy Killer'. (July 2021). SSRN.
[9] P. Cuff and L. Yu. 2016. Differential privacy as a mutual information constraint. In *CCS 2016*. Association for Computing Machinery, New York, NY, 43–54.
[10] A. De. 2012. Lower bounds in differential privacy. In *Theory of Cryptography Conference*. International Association for Cryptologic Research, 321–338.
[11] M. Du, R. Jia, and D. Song. 2020. Robust anomaly detection and backdoor attack detection via differential privacy. In *International Conference on Learning Representations (ICLR'20)*.
[12] F. du Pin Calmon and N. Fawaz. 2012. Privacy against statistical inference. In *50th Annual Allerton Conference*. IEEE, New York, NY, 1401–1408.
[13] C. Dwork. 2006. Differential privacy. *Automata, Languages and Programming*. Springer, Berlin, 1–12.
[14] C. Dwork. 2008. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation (TAMC'08)*, Lecture Notes in Computer Science, Vol. 4978. Springer, Berlin, 1–19.
[15] C. Dwork, F. McSherry, K. Nissim, and A. Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*. International Association for Cryptologic Research, 265–284.
[16] C. Dwork and A. Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9 (2014), 211–407.
[17] C. Dwork, G. N. Rothblum, and S. Vadhan. 2010. Boosting and differential privacy. In *51st Annual Symposium on Foundations of Computer Science (FOCS'10)*. IEEE Computer Society, NW Washington, DC, 51–60.
[18] C. Dwork and A. Smith. 2010. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality* 1, 2 (2010), 135–154.
[19] K. E. Emam, E. Jonker, L. Arbuckle, and B. Malin. 2011. A systematic review of re-identification attacks on health data. *Plos One PMC* 6, 12 (2011), 1–12.
[20] U. Erlingsson, V. Pihur, and A. Korolova. 2014. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *ACM SIGSAC Conference on Computer and Communications*. ACM, New York, NY, 1054–1067.
[21] A. Evfimievski, J. Gehrke, and R. Srikant. 2003. Limiting privacy breaches in privacy preserving data mining. In *22nd ACM Symposium on Principles of Database Systems*. ACM, New York, NY, 211–222.
[22] Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. 2022. Differential privacy and fairness in decisions and learning tasks: A survey. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI'22)*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 5470–5477. https://doi.org/10.24963/ijcai.2022/766 Survey Track.
[23] Q. Geng and P. Viswanath. 2014. The optimal mechanism in differential privacy. In *IEEE International Symposium on Information Theory*. IEEE, New York, NY, 2371–2375.
[24] C. Hirche, C. Rouzé, and D. S. França. 2023. Quantum differential privacy: An information theory perspective. arXiv:2202.10717. Retrieved from https://arxiv.org/abs/2202.10717.
[25] P. Kairouz, S. Oh, and P. Viswanath. 2015. The composition theorem for differential privacy. In *32nd International Conference on Machine Learning*. JMLR, Inc. and Microtome Publishing, 4037–4049.
[26] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy*. 1054–1067.
[27] P. H. Lin, C. Kuhn, T. Strufe, and E. A. Jorswieck. 2019. Physical layer privacy in broadcast channels. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS'19)*. IEEE, 1–6.
[28] Z. Lin, M. Hewett, and R. B. Altman. 2002. Using binning to maintain confidentiality of medical data. In *AMIA 2002 Annual Symposium Proceedings*. 454–458.
[29] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. 2008. Privacy: Theory meets practice on the map. In *IEEE 24th International Conference on Data Engineering*. IEEE, New York, NY, 277–286.
[30] A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. Vadhan. 2010. The limits of two-party differential privacy. In *51st Annual Symposium on Foundations of Computer Science (FOCS'10)*. IEEE Computer Society, NW Washington, DC, 81–90.

[31] D. J. Mir. 2012. Differentially-private learning and information theory. In *International Workshop on Privacy and Anonymity in the Information Society PAIS*. ACM, New York, NY, 206–210.
[32] D. J. Mir. 2012. Information theoretic foundations of differential privacy. In *International Symposium of Foundations on Practice of Security*. Springer, Berlin, 374–381.
[33] A. Narayanan and V. Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*. IEEE, New York, NY, 111–125.
[34] A. Padakandla, P. R. Kumar, and W. Szpankowski. 2020. Trade-off between privacy and fidelity vie Ehrhart theory. *IEEE Transactions on Information Theory* 66, 4 (Apr. 2020), 2549–2569.
[35] D. Pankratov. 2015. *Communication Complexity and Information Complexity*. Ph.D. Dissertation. The University of Chicago.
[36] A. Pastore and M. Gastpar. 2021. Locally differentially private randomized response for discrete distribution learning. *Journal on Machine Learning Research* 22 (July 2021), 1–56.
[37] NhatHai Phan, My T. Thai, Han Hu, Ruoming Jin, Tong Sun, and Dejing Dou. 2020. Scalable differential privacy with certified robustness in adversarial learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Article 712, 12 pages.
[38] Rafael Pinot, Florian Yger, Cedric Gouy-Pailler, and Jamal Atif. 2019. A unified view on differential privacy and robustness to adversarial examples. In *Workshop on Machine Learning for CyberSecurity at ECMLPKDD 2019*. https://hal.science/hal-02892170.
[39] Makhamisa Senekane, Mhlambululi Mafu, and Benedict Molibeli Taele. 2017. Privacy-preserving quantum machine learning using differential privacy. In *2017 IEEE AFRICON (2017 IEEE AFRICON: Science, Technology and Innovation for Africa, AFRICON 2017)*, Darryn R. Cornish (Ed.). Institute of Electrical and Electronics Engineers, Inc., 1432–1435. https://doi.org/10.1109/AFRCON.2017.8095692
[40] L. Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 557–570.
[41] A. Ünsal and M. Önen. 2021. A statistical threshold for adversarial classification in Laplace mechanisms. In *IEEE Information Theory Workshop 2021*. IEEE, New York, NY, 1–6.
[42] Ayşe Ünsal and Melek Önen. 2022. Calibrating the attack to sensitivity in differentially private mechanisms. *Journal of Cybersecurity and Privacy* 2, 4 (2022), 830–852. https://doi.org/10.3390/jcp2040042
[43] W. Wang, L. Ying, and J. Zhang. 2016. On the relation between identifiability, differential privacy and mutual information privacy. *IEEE Transactions on Information Theory* 62, 9 (Sep. 2016), 5018–5029.
[44] G. Wu, X. Xia, and Y. He. 2021. Achieving Dalenius' Goal of Data Privacy with Practical Assumptions. (May 2021). https://arxiv.org/abs/1703.07474v5.
[45] H. Zang and J. Bolot. 2011. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the International Conference on Mobile Computing and Networking 17*. ACM, New York, NY, 145–156.
[46] Li Zhou and Mingsheng Ying. 2017. Differential privacy in quantum computation. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF'17)* (2017), 249–262.
[47] S. Zhou, K. Ligett, and L. Wasserman. 2009. Differential privacy with compression. In *IEEE International Symposium on Information Theory (ISIT'09)*. IEEE, New York, NY, 2718–2722.

Received 13 April 2022; revised 27 March 2023; accepted 2 May 2023