Noiseless Privacy: Definition, Guarantees, and Applications

Farhad Farokhi

Abstract—In this paper, we define noiseless privacy, as a non-stochastic rival to differential privacy, requiring that the outputs of a mechanism (i.e., function composition of a privacy-preserving mapping and a query) attain only a few values while varying the data of an individual (the logarithm of the number of the distinct values is bounded by the privacy budget). Therefore, the output of the mechanism is not fully informative of the data of the individuals in the dataset. We prove several guarantees for noiselessly-private mechanisms. The information content of the output about the data of an individual, even if an adversary knows all the other entries of the private dataset, is bounded by the privacy budget. The zero-error capacity of memory-less channels using noiselessly private mechanisms for transmission is upper bounded by the privacy budget. The performance of a non-stochastic hypothesis-testing adversary is bounded again by the privacy budget. Assuming that an adversary has access to a stochastic prior on the dataset, we prove that the estimation error of the adversary for individual entries of the dataset is lower bounded by a decreasing function of the privacy budget. In this case, we also show that the maximal leakage is bounded by the privacy budget. In addition to privacy guarantees, we prove that noiselessly-private mechanisms admit composition theorem and post-processing does not weaken their privacy guarantees. We prove that quantization or binning can ensure noiseless privacy if the number of quantization levels is appropriately selected based on the sensitivity of the query and the privacy budget. Finally, we illustrate the privacy merits of noiseless privacy using multiple datasets in energy, transport, and finance.

I. INTRODUCTION

Big data revolution, equipped with novel tools for data collection, analysis, and reporting, has significant promises for answering societal challenges. These promises however come at the cost of the erosion of privacy. Therefore, there is a need for rigorous protection of the privacy of individuals.

Natural candidates for privacy protection, such as differential privacy [1], [2] and information-theoretic privacy [3], [4], either require randomized policies for privacy protection or assume statistical properties about the private dataset. The definition of differential privacy assumes the use of randomized functions as well as the probability of outputs, and conventional information-theoretic tools, such as mutual information and entropy, rely on random variables.

Heuristic-based privacy-preserving methods, such as k -anonymity [5], [6] and ℓ -diversity [7], are however deterministic in nature. They employ deterministic mechanisms, such as suppression and generalization, and do not assume stochastic

properties about the datasets. Popularity of these methods is evident from the availability of toolboxes¹.

Although providing powerful guarantees, randomized or stochastic privacy-preserving policies sometimes cause problems, such as un-truthfulness [8], that are undesirable in practice [9]. This is touted as a reason behind slow adoption of differential privacy within financial and health sectors [8]. For instance, randomized policies, stemming from differential privacy in financial auditing, complicate fraud detection [10], [11]. Randomized policies can also generate unreasonable and unrealistic outputs that might mislead investors or market operators, e.g., by reporting noisy outputs that point to lack of liquidity in a financial sector while that was not the case. For instance, the slow-decaying nature of the Laplace noise means impossible reports (e.g., negative median income) can occur with a non-negative probability [12]. Randomized privacy-preserving policies have also encountered difficulties in medical, health, or social sciences [13], [14]. Furthermore, the Laplace mechanism, a common approach to ensuring differential privacy, is shown to cause undesirable properties, e.g., the optimal estimation in the presence of Laplace noise is computationally expensive [15]. These motivate the development of non-stochastic privacy metrics and privacy-preserving policies in a rigorous manner.

Although it has been proved that noiseless policies cannot provide the strong guarantees of randomized policies, e.g., it has been proved that differential privacy cannot be delivered without noise [2], the popularity of noiseless privacy-preserving policies justifies investigating metrics for analysis and comparison. This must be done irrespective of inherent philosophical weaknesses in comparison to stochastic policies because they belong to a different category. Also, we must be open to adopting other notions of privacy when addition of noise is unacceptable, perhaps for legal reasons. This scenario might not ideal, but real, and demands an approach to privacy that is formal and clear about the guarantees provided.

In this paper, we define the new notion of noiseless privacy. Noiseless privacy implies that the outputs of a mechanism must only attain a few distinct values while varying the data of an individual. Therefore, the output of the mechanism is not very informative about the data of the individuals in a dataset. We prove the following guarantees for the noiselessly-private mechanisms:

- The information content of the output about the data of each individual, even if an adversary knows all the other entries of the private dataset, is bounded from above by the privacy budget (a constant similar to the privacy

F. Farokhi is with the Department of Electrical and Electronic Engineering at the University of Melbourne.
e-mail: farhad.farokhi@unimelb.edu.au

The work of F. Farokhi is, in part, supported by the Melbourne School of Engineering at the University of Melbourne.

¹<https://arxiv.org/overview/related-software/>

budget in differential privacy capturing the amount of the leaked information). As non-stochastic notions of information, we use the non-stochastic information leakage in [16] and the maximin information [17]. These are established measures of information in the non-stochastic information theory literature [16]–[19].

- The zero-error capacity of memory-less channels using noiselessly-private mechanisms for data transmission is upper bounded by the privacy budget. The zero-error capacity is the non-stochastic equivalent of normal capacity, also coined by Shannon, while investigating non-stochastic communication channels and worst-case behaviours [20].
- The performance of an adversary performing non-stochastic hypothesis tests [21] on the data of an individual, while knowing all the other entries of the private dataset, is bounded again by the privacy budget.
- Assuming that an adversary has access to a stochastic prior about the dataset, we prove that the error of an adversary for estimating the data each individual is lower bounded by a decreasing function of the privacy budget. Therefore, by reducing the privacy budget, the estimation error of the adversary worsens. In this case, we also show that the maximal information leakage (in the sense of [22]) is upper bounded by the privacy budget. Hence, by reducing the privacy budget, we can also reduce the maximal information leakage.

In addition to these privacy guarantees, we prove the following important properties:

- Noiselessly-private mechanisms admit composition theorem, i.e., the privacy budgets of the mechanisms add up when reporting on multiple queries on the same private dataset.
- Post-processing of noiselessly-private mechanisms does not weaken their privacy guarantees, i.e., the privacy budget can only be decreased by post-processing.

We also prove that quantization operators can ensure noiseless privacy. We provide a recipe for determining the number of quantization levels based on the sensitivity of the query and the privacy budget. Finally, we illustrate the privacy merits of noiseless privacy using multiple datasets in energy, transport, and finance.

A. Related Studies

Anonymization: Anonymization is widely used within public and private sectors for releasing sensitive datasets² for public competitions and analysis. Although popularly adopted, anonymization is often insufficient for privacy preservation [23]–[25] and hence, systematic methods with provable guarantees are required.

Differential Privacy: Differential privacy offers provable privacy guarantees [2], [26]–[31]. This method uses randomization to provide plausible deniability for the data of an individual by ensuring that the statistics of privacy-preserving outputs do not change significantly by varying the data of

individual. Additive Laplace and Gaussian noise with scales proportional to the sensitivity of the submitted query with respect to the individual entries of the dataset are proved to guarantee differential privacy [2]. By definition, differential privacy requires randomization.

Information-Theoretic Privacy: Information-theoretic privacy, a rival to differential privacy, dates back to studying secrecy [32] and its generalizations [3], [4], [33], [34]. Information-theoretic guarantees have been also used to measure the quantity of leaked private information when using differential privacy [35], [36]. Other studies have used information theory to modify, relax, or generalize the definition of privacy [37], [38]. In information-theoretic privacy, entropy, mutual information, Kullback-Leibler divergence, and Fisher information have been repeatedly used as measures of privacy [39]–[45]. Maximal leakage based on guessing theory was recently developed based on an axiomatic approach to measuring private information leakage [22]. This measure is similar to max-information in [46]. In [47], sparse data analysis is used to ensure information-theoretic privacy. Information theory, starting with Shannon [48], assumes that data source and communication channels are random, and is powerful in modeling and analyzing communication systems. Traditional notions in information theory, such as mutual information, are not useful for analyzing non-stochastic/noiseless settings and deterministic privacy-preserving policies.

Deterministic Privacy-Preserving Policies: Noiseless privacy-preserving policies are often heuristic-based making them vulnerable to attacks, e.g., k -anonymity is vulnerable to homogeneity attack [7]. This is because we do not possess sensible measures/definitions for privacy that extend to noiseless privacy-preserving policies on deterministic datasets. Therefore, we cannot prove, in any sense, privacy guarantees of noiseless privacy-preserving. The popularity of noiseless privacy-preserving policies justifies investigating metrics for analysis and comparison. Here, we propose a non-stochastic or noiseless rival to differential privacy.

In this paper, we are proposing a noiseless notion of privacy without any statistical assumptions. This is entirely different from the so-called noiseless database privacy in [10]. In [10], [49], although additive noise and randomization are not used to preserve privacy, the underlying private dataset is assumed to be randomized. This stochasticity assumption is then utilized to ensure a differential-privacy-alike condition on the outputs. This makes the formulation and the guarantees of noiseless database privacy in [10], [49] dependent on the probability distribution of the dataset and the auxiliary information available to the adversary, which can make the framework susceptible to wrong assumptions. This is not the case in this paper.

We particularly use non-stochastic information theory [17], non-stochastic hypothesis testing [21], and the zero-error capacity [20] to investigate the merits of the noiseless privacy definition. These analyses do not require statistical assumptions or knowledge about the private dataset. Non-stochastic information theory dates back to early studies of information transmission [17], [50]–[53]. It has been recently revived [54]–[56]. Most recently, non-stochastic information theory was

²See <https://data.gov.au> and <https://www.kaggle.com> for examples.

used in [16] for investigating deterministic privacy-preserving policies. Interestingly, in [16], it was proved that k -anonymity is not privacy-preserving using non-stochastic information theory, a fact that was only observed using adversarial attacks in [7].

B. Paper Organization

The rest of the paper is organized as follows. Background material on non-stochastic information theory and hypothesis testing are presented in Section II. Noiseless privacy is defined in Section III. In this section, guarantees and properties of noiseless privacy are also presented. A method for ensuring noiseless privacy is presented in Section IV. Experimental results are presented in Section V. Finally, the paper is concluded in Section VI.

II. UNCERTAIN VARIABLES, HYPOTHESIS TESTING, AND NON-STOCHASTIC INFORMATION THEORY

We start by reviewing a few necessary concepts from non-stochastic information theory.

A. Uncertain Variables

Let Ω be an uncertainty set whose elements, i.e., $\omega \in \Omega$, model/capture the source of uncertainty. An uncertain variable X is defined as a mapping on Ω . For any uncertain variable $X : \Omega \rightarrow \mathbb{X}$, $X(\omega)$ is the realization of uncertain variable X (corresponding to the realization of uncertainty $\omega \in \Omega$). *Marginal range* of uncertain variable X is $\llbracket X \rrbracket := \{X(\omega) : \omega \in \Omega\} \subseteq \mathbb{X}$. *Joint range* of uncertain variables $X : \Omega \rightarrow \mathbb{X}$ and $Y : \Omega \rightarrow \mathbb{Y}$ is defined as $\llbracket X, Y \rrbracket := \{(X(\omega), Y(\omega)) : \omega \in \Omega\} \subseteq \mathbb{X} \times \mathbb{Y}$. *Conditional range* of uncertain variable X , conditioned on the realizations of uncertain variable Y belonging to the set $\mathcal{Y}$, i.e., $Y(\omega) \in \mathcal{Y} \subseteq \llbracket Y \rrbracket$, is given by $\llbracket X|\mathcal{Y} \rrbracket := \{X(\omega) : \exists \omega \in \Omega \text{ such that } Y(\omega) \in \mathcal{Y}\} \subseteq \llbracket X \rrbracket$. If $\mathcal{Y}$ is a singleton, i.e., $\mathcal{Y} = \{y\}$, we use $\llbracket X|y \rrbracket$ instead of $\llbracket X|\{y\} \rrbracket = \llbracket X|\mathcal{Y} \rrbracket$. The definition of uncertain variables and their properties are similar to those of random variables with the exception of not requiring a measure on Ω . Finally, if the marginal range $\llbracket X \rrbracket$ is uncountably infinite for uncertain variable X , we refer to X as a *continuous* uncertain variable, similar to a continuous random variable. If the marginal range $\llbracket X \rrbracket$ is countable for uncertain variable X , we call X a *discrete* uncertain variable. For a more comprehensive review of uncertain variables, see [16], [17].

B. Non-Stochastic Information Theory

Non-stochastic entropy of discrete uncertain variable X is $H_0(X) := \log_2(|\llbracket X \rrbracket|) \in \mathbb{R} \cup \{\pm\infty\}$. For continuous uncertain variable X , the non-stochastic (differential) entropy is given by $h_0(X) := \log_e(\mu(|\llbracket X \rrbracket|)) \in \mathbb{R} \cup \{\pm\infty\}$, where $\mu(\cdot)$ is the Lebesgue measure. In the context of information-theoretic privacy [16], the non-stochastic information leakage is defined as $L_0(X; Y) := \max_{y \in \llbracket Y \rrbracket} \log_2(|\llbracket X \rrbracket|/|\llbracket X|y \rrbracket|)$ for discrete uncertain variables. Similarly, for continuous uncertain variables, the non-stochastic information leakage is $L_0(X; Y) := \text{ess sup}_{y \in \llbracket Y \rrbracket} \log_e(\mu(\llbracket X \rrbracket)/\mu(\llbracket X|y \rrbracket))$. With

slight adaptation, Kolmogorov had previously defined ‘combinatorial’ conditional entropy using $\log(|\llbracket X|y \rrbracket|)$ and the information gain as $|\llbracket X \rrbracket|/|\llbracket X|y \rrbracket|$ in [51]. The combinatorial conditional entropy and the information gain are only defined for a fixed realization $Y(\omega) = y$ while $L_0(X; Y)$ is based on the worst-case, i.e., the most informative, output. Finally, note that $L_0(X; Y)$ is not symmetric. We define the symmetrized non-stochastic information leakage as

$$L_0^s(X; Y) := \min\{L_0(X; Y), L_0(Y; X)\}. \quad (1)$$

Note that, by construction, $L_0^s(X; Y) = L_0^s(Y; X)$. For a more comprehensive review of non-stochastic information theory in privacy, see [16].

a) *Maximin Information:* In [17], the maximin information was introduced as a symmetric measure of information and its relationship with zero-error capacity was explored. To present the definition of the maximin information, we need to introduce the notion of overlap partitions:

- $x, x' \in \llbracket X \rrbracket$ are $\llbracket X|Y \rrbracket$ -overlap connected if there exists a finite sequence of conditional ranges $\{\llbracket X|y_i \rrbracket\}_{i=1}^n$ such that $x \in \llbracket X|y_1 \rrbracket$, $x' \in \llbracket X|y_n \rrbracket$, and $\llbracket X|y_i \rrbracket \cap \llbracket X|y_{i+1} \rrbracket \neq \emptyset$ for all $i = 1, \dots, n-1$;
- $\mathcal{A} \subseteq \llbracket X \rrbracket$ is $\llbracket X|Y \rrbracket$ -overlap connected if all $x, x' \in \mathcal{A}$ are $\llbracket X|Y \rrbracket$ -overlap connected;
- $\mathcal{A}, \mathcal{B} \subseteq \llbracket X \rrbracket$ are $\llbracket X|Y \rrbracket$ -overlap isolated if there does not exist $x \in \mathcal{A}, x' \in \mathcal{B}$ such that x, x' are $\llbracket X|Y \rrbracket$ -overlap connected;
- An $\llbracket X|Y \rrbracket$ -overlap partition is a partition of $\llbracket X \rrbracket$ such that each member set is $\llbracket X|Y \rrbracket$ -overlap connected and any two member sets are $\llbracket X|Y \rrbracket$ -overlap isolated.

A unique $\llbracket X|Y \rrbracket$ -overlap partition always exists [17], and is denoted by $\llbracket X|Y \rrbracket_\star$. The maximin information is

$$I_\star(X; Y) := \log_2(|\llbracket X|Y \rrbracket_\star|). \quad (2)$$

In [17], it was proved that $|\llbracket X|Y \rrbracket_\star| = |\llbracket Y|X \rrbracket_\star|$ and thus $I_\star(X; Y) = I_\star(Y; X)$. In [57], it is shown that $I_\star(X; Y) \leq L_0(X; Y)$ for discrete uncertain variables. In Appendix A in Supplementary Material, we generalize this result to the case where only Y needs to be a discrete uncertain variable (i.e., X can be a continuous uncertain variable). We also improve the bound by relating $I_\star(X; Y)$ and $L_0^s(X; Y)$. For a more comprehensive review of maximin information, see [17].

An uncertain time series X is a sequence of uncertain variables $X[k] : \Omega \rightarrow \mathbb{X}$ for all $k \in \mathbb{N}$. Alternatively, we can think of uncertain time series X as a mapping from the sample space Ω to the set of functions from $\mathbb{N}$ to $\mathbb{X}$. Now, we can define a memory-less uncertain communication channel. A memory-less uncertain channel maps any uncertain time series X to uncertain time series Y such that $\llbracket Y[k], \dots, Y[1]|X[k](\omega) = x[k], \dots, X[1](\omega) = x[1] \rrbracket = \llbracket Y[k]|X[k](\omega) = x[k] \rrbracket \times \dots \times \llbracket Y[1]|X[1](\omega) = x[1] \rrbracket$ for all $(x[k], \dots, x[1]) \in \llbracket X[k], \dots, X[1] \rrbracket$ and all $k \in \mathbb{N}$. A code of length k is a finite set $\mathcal{F} \subseteq \mathbb{X}^k$ with each codeword $f \in \mathcal{F}$ denoting a distinct message. Define $\mathcal{X}(y[k], \dots, y[1]) :=$

$\llbracket X[k], \dots, X[1] | Y[k](\omega) = y[k], \dots, Y[1](\omega) = y[1] \rrbracket$. The zero-error capacity is

$$C_0 := \lim_{k \rightarrow \infty} \sup_{\substack{\mathcal{F} \subseteq \mathbb{X}^k : \\ |\mathcal{F} \cap \mathcal{X}(y[k], \dots, y[1])| \leq 1, \\ \forall (y[k], \dots, y[1]) \in \mathbb{Y}^k}} \frac{\log_2(|\mathcal{F}|)}{k}. \quad (3)$$

In Appendix B in Supplementary Material, we provide an upper bound for the zero-error capacity based on the symmetrized non-stochastic information leakage.

C. Non-Stochastic Hypothesis Testing

Consider uncertain variable X denoting the original uncertain variable. An adversary is interested in testing the validity of a hypothesis for the realizations of X using *uncertain measurements* denoted by Y . This is captured by that $Y(\omega) = g_Y(X(\omega))$ for a mapping $g_Y : \llbracket X \rrbracket \rightarrow \llbracket Y \rrbracket$. The *hypothesis* is an uncertain variable H with binary range $\llbracket H \rrbracket = \{p_0, p_1\}$, where p_0 denotes the *null hypothesis* and p_1 denotes the *alternative hypothesis*. We assume that there exists a mapping $g_H : \llbracket X \rrbracket \rightarrow \llbracket H \rrbracket$ such that $H = g_H \circ X$; the hypothesis is constructed based on the uncertain variable X as $H(\omega) = g_H(X(\omega))$. A *test* is a function $T : \llbracket Y \rrbracket \rightarrow \llbracket H \rrbracket = \{p_0, p_1\}$. If $T(Y) = p_1$, the test rejects the null hypothesis in favour of the alternative hypothesis; however, if $T(Y) = p_0$, the test accepts the null hypothesis (and rejects the alternative hypothesis). Following [21], a test T is correct at a particular realization $Y(\omega) = y \in \llbracket Y \rrbracket$, if $\llbracket H | \llbracket X | y \rrbracket \rrbracket = \{T(y)\}$. The set of all outputs at which test T is correct is equal to $\aleph(T) := \{y \in \llbracket Y \rrbracket : \llbracket H | \llbracket X | y \rrbracket \rrbracket = \{T(y)\}\}$. If Y is a continuous uncertain variable, the performance of test T is $\mathcal{P}(T) := \log_e(\mu(\aleph(T)))$. Similarly, if Y is a discrete uncertain variable, the performance of test T is equal to $\mathcal{P}(T) := \log_2(|\aleph(T)|)$. In [21], it is proved that the performance of any test T is bounded by $\mathcal{P}(T) \leq \log_e(\mu(\llbracket Y | p_0 \rrbracket \Delta \llbracket Y | p_1 \rrbracket))$ if Y is a continuous uncertain variable, and by $\mathcal{P}(T) \leq \log_2(|\llbracket Y | p_0 \rrbracket \Delta \llbracket Y | p_1 \rrbracket|)$ if Y is a discrete uncertain variable, where Δ denotes the symmetric difference operator on the sets, i.e., $\mathcal{A} \Delta \mathcal{B} = (\mathcal{A} \setminus \mathcal{B}) \cup (\mathcal{B} \setminus \mathcal{A})$. For a more comprehensive review of non-stochastic hypothesis testing, see [21].

III. NOISELESS PRIVACY:

DEFINITION, GUARANTEES, AND PROPERTIES

We model a *private dataset* by a realization of a *vector-valued uncertain variable* $X : \Omega \rightarrow \mathbb{R}^n$ with n denoting the number of individuals whose data is in the dataset. The dataset is therefore in the form of

$$X(\omega) = \begin{bmatrix} X_1(\omega) \\ X_2(\omega) \\ \vdots \\ X_n(\omega) \end{bmatrix},$$

where $X_i(\omega) \in \mathbb{R}$ is the data of the i -th individual. Evidently, each $X_i : \Omega \rightarrow \mathbb{R}$, $1 \leq i \leq n$, is itself an uncertain variable.

A data curator, in possession of the realization of uncertain variable X , i.e., the private dataset $X(\omega)$, must return a response to a query $f : \llbracket X \rrbracket \rightarrow \mathbb{R}^q$ for some $q \in \mathbb{N}$. The

curator employs a mechanism $\mathfrak{M} : \mathbb{R}^q \rightarrow \mathbb{R}^q$ to generate a privacy-preserving response. Throughout this paper, $\mathfrak{M} \circ f$ is referred to as the mechanism of the curator. This is the same language used in the privacy literature albeit without the presence of the randomness [1], [58]. Therefore, the curator provides the response $Y(\omega) = \mathfrak{M} \circ f(X(\omega))$. By definition, $Y = \mathfrak{M} \circ f \circ X$ is an uncertain variable. In the remainder of this paper, we use the notation x_{-i} to denote $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ for vectors and X_{-i} to denote $(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ for uncertain variables alike. In this notation, $-i$ refers essentially refers to the set of all individuals except the i -th one.

Definition 1 (Noiseless Privacy). *A mechanism $\mathfrak{M} \circ f$ is η -noiselessly private, for $\eta > 0$, if*

$$\|\llbracket Y | X_{-i}(\omega) = x_{-i} \rrbracket\| \leq 2^\eta, \quad \forall x_{-i} \in \llbracket X_{-i} \rrbracket, \forall i, \quad (4)$$

where $Y = \mathfrak{M} \circ f \circ X$.

Note that this definition is akin to a noiseless differential privacy. This is because, instead of bounding the information leakage as in information-theoretic privacy [16], the output realizations are restricted if one individual entry of the dataset changes. Note that, when the data of i -th individual changes, the output can take all the values within the set $\llbracket Y | X_{-i}(\omega) = x_{-i} \rrbracket$. If this set is not informative, i.e., it does not contain many elements, reverse engineering the data of i -th individual changes with knowledge of $X(\omega)$ even in the presence of side-channel information is a difficult task. The privacy budget η is similar to the privacy budget in differential privacy, which is also often denoted by ϵ . In fact, by reducing η , we can improve the privacy guarantees. We prove this by showing that the privacy budget is the upper bound for multiple notions of information leakage. This allows us to set η by investigating the physical interpretations of these notations of information leakage in practical scenarios. For instance, we show that maximal leakage in [22] is bounded by η when ensuring η -noiseless privacy. Therefore, we can use the operational definition of leakage in [22] based on guessing theory to set η . The same goes for other discussed notions of information leakage. Similar to differential privacy, we can also define a local version of noiseless privacy.

Note that, in Definition 1, $\|\llbracket Y | X_{-i}(\omega) = x_{-i} \rrbracket\|$ informs the privacy level. However, to ensure η -noiseless privacy, we require that $\|\llbracket Y | X_{-i}(\omega) = x_{-i} \rrbracket\|$ is upper bounded by 2^η instead of setting $\|\llbracket Y | X_{-i}(\omega) = x_{-i} \rrbracket\|$ equal to 2^η . This is similar to the definition of differential privacy in which the ratio of the probabilities is bounded by the exponential of the privacy budget, instead of being set equal to it. This implies that any mechanism that is η -noiselessly or -differentially private is also η' -noiselessly or -differentially private for $\eta' \geq \eta$.

Example 1. Table I shows two datasets containing the heights of five individuals. These datasets only differ in the height of one individual, i.e., Bob. Knowing that the heights are always between one and two meters, we can model each row of the dataset, i.e., the height of each individual, with an uncertain variable X_i whose marginal range is $\llbracket X_i \rrbracket = [1, 2]$. We are

Name	X_i	X'_i
Alice	1.5	1.5
Bob	1.8	1.5
Mathew	1.6	1.6
Jane	1.7	1.7
Libby	1.6	1.6

TABLE I: Two datasets containing heights of five individuals for an illustrative example.

interested in reporting the average height of the group. The average height in the first dataset is 1.64 while it is 1.58 in the second dataset. Assume that, instead of the average height, we report the binned or quantized average height with bins $\{[1.00, 1.25), [1.25, 1.50), [1.50, 1.75), [1.75, 2]\}$. This is 2-noiselessly private; see Theorem 8. We can see that, in this illustrative example, the output is $[1.50, 1.75)$ for both datasets. Therefore, the variation in Bob's height would not change what we report. Therefore, Bob's privacy is maintained. In differential privacy, we need to add noise to ensure privacy. For instance, if we report the average height perturbed by zero-mean Laplace noise with scale $1/10$, we can ensure 2-differential privacy. In this case, there exists a non-zero probability that the output is smaller than 1 or larger than 2. This is undesirable as the output can only be within $[1, 2]$ by construction.

Definition 2 (Local Noiseless Privacy). Assume that $f_i : (x_i)_{i=1}^n \mapsto x_i$ for each i . A mechanism $\mathcal{M}$ is η -locally noiselessly private if $\mathcal{M} \circ f_i$ is η -noiselessly private for all i .

Local noiseless privacy is the preferred notion of privacy when the third-party data aggregators is untrusted. This is because the data owners can generate reports that satisfy local noiseless privacy so that the untrusted aggregator cannot access their private data. This is the same as the relationship between local differential privacy and differential privacy.

We make the following standing assumptions regarding the query and the underlying uncertain variables to simplify the analysis.

Assumption 1. Query f is (i) continuous and (ii) has connected preimage, i.e., $f^{-1}(y) := \{x \in \llbracket X \rrbracket \mid f(x) = y\}$ is a connected set for all $y \in \llbracket f(X) \rrbracket$.

Assumption 1 (a) calls for continuous query functions. Many query functions, e.g., all linear queries as well as empirical mean and variance, are continuous functions. However, some queries, such as percentile statistics, are discontinuous. In those cases, we must approximate the query function with continuous functions, potentially at the risk of degrading the utility (because of the combined utility degradation caused by continuous function approximation and noiseless privacy). Assumption 1 (b) states that the query function must have a connected preimage. Comparing to Assumption 1 (a), this can be more conservative. It should be noted that linear queries satisfy this assumption. This condition is met for any query function that is the composition of a linear function and a non-decreasing continuous function.

Assumption 2. $\llbracket X_i \rrbracket$ is a connected set for all $i \in \{1, \dots, n\}$.

Assumption 2 requires that the range of each private uncertain variable X_i is a connected set. We can always use an outer approximation of the range that is connected, e.g., ellipsoid or box sets, to guarantee this assumption. However, such an approximation can potentially degrade the utility if the volume/count of the range changes significantly.

In what follows, we use non-stochastic information theory to establish the extend of the privacy guarantees by noiseless privacy.

A. Guarantee: Non-Stochastic Information Leakage

We define a function $\psi_{i,v_{-i}} : \llbracket X \rrbracket \rightarrow \llbracket X_i \rrbracket \times \{v_{-i}\}$ replacing the value of $X_j(\omega)$ or the realization of X_j , for all $1 \leq j \leq$ except for $j = i$, with given constants v_j , i.e., $\psi_{i,v_{-i}}(X(\omega)) = (v_1, \dots, v_{i-1}, X_i(\omega), v_{i+1}, \dots, v_n)$. Let $\Psi_i = \{\psi_{i,v_{-i}} \mid v_{-i} \in \llbracket X_{-i} \rrbracket\}$ be the set of all such functions for $i \in \{1, \dots, n\}$. The uncertain variable $\psi_{i,v_{-i}} \circ X$ becomes unrelated (in the sense of [17]) to X_{-i} for all $\psi_{i,v_{-i}} \in \Psi_i$.

This definition allows us to measure the amount of information that the curator's mechanism leaks about the data of the i -th individual $X_i(\omega)$. For a given $\psi_{i,v_{-i}} \in \Psi_i$, let us define $Y = \mathcal{M} \circ f \circ \psi_{i,v_{-i}} \circ X$. Now, the information between Y and X_i captures how much more information can an adversary extract from Y knowing the data of all the individuals except the i -th individual. This is because, here, we let the adversary to select any possible $\psi_{i,v_{-i}}$.

Theorem 1 (Non-Stochastic Information vs Noiseless Privacy). Assume X is a discrete uncertain variable, $Y = \mathcal{M} \circ f \circ \psi_{i,v_{-i}} \circ X$, and $\mathcal{M} \circ f$ is η -noiselessly private. For any $\psi_{i,v_{-i}} \in \Psi_i$,

$$0 \leq I_\star(X_i; Y) \leq L_0^s(X_i; Y) \leq L_0(Y; X_i) \leq \eta. \quad (5)$$

Proof. See Appendix C in Supplementary Material. $\square$

Theorem 1 shows that, by reducing η , we can reduce the amount of leaked information about each individual measured using the non-stochastic information leakage and the maximin information. This makes sense. Consider the case where $\eta = +\infty$. In this case, the curator can report the output of the query $f(\psi_{i,v_{-i}} \circ X(\omega))$ completely (i.e., $\mathcal{M}$ can be chosen to be equal identity) and the adversary, knowing v_{-i} , can compute the data of the data of the i -th individual $X_i(\omega)$ (at least if the adversary select the query to be linear with non-zero weight for the i -th individual). On the other hand, if $\eta = 0$, the output becomes a constant that is independent of $X_i(\omega)$ and thus the adversary learns nothing new about the data of the i -th individual $X_i(\omega)$.

Remark 1. Note that the upper bound in (5) is tight in the sense that there exist problem formulations for which the equality can be achieved. To observe this, assume that $n = 1$, X_1 is an uncertain variable with $\llbracket X_1 \rrbracket = [0, 1]$, and that we are interested in receiving a response to the query $f(X_1) = X_1$. This can be seen as the setup for local noiseless privacy. Let $Y = \mathcal{M} \circ f(X_1)$ using any quantizer $\mathcal{M}$ with 2^n levels over $[0, 1]$. Evidently, $\mathcal{M} \circ f$ is η -noiselessly private (because the quantizer at most can generate 2^n outputs). We can also observe that $L_0(Y; X_1) =$

$\sup_{x_1 \in [0,1]} \log_2(\|\llbracket Y \rrbracket\| / \|\llbracket Y|X_1(\omega) = x_1 \rrbracket\|) = \log_2(2^\eta) = \eta$ because $\llbracket Y|X_1(\omega) = x_1 \rrbracket$ is a singleton for each x_1 .

B. Guarantee: Zero-Error Capacity

Let us consider a memory-less noiselessly-private communication channel. This can be seen as a non-stochastic equivalent of differentially-private communication channels in [59].

Let $\mathfrak{M} \circ f$ be a η -noiselessly private for some $\eta > 0$. For any given sequence of mappings $\{\psi_{i,v-i}^t\}_{t \in \mathbb{N}}$ with $\psi_{i,v-i}^t \in \Psi_i$, a memory-less η -noiselessly-private channel maps any uncertain time series $X = (X[k], \dots, X[1])$ to uncertain time series $Y = (Y[k], \dots, Y[1])$ such that $Y[\ell](\omega) = \mathfrak{M} \circ f \circ \psi_{i,v-i}^\ell(X(\omega))$ for all $1 \leq \ell \leq k$ and $k \in \mathbb{N}$.

This setup can be seen as a scenario in which the curator is reporting on a stream of data from the individuals. We can assume that an extremely strong adversary can set the realizations of the data of all individuals except the i -th individual. The capacity of the channel captures the amount of information that passes through a η -noiselessly private mechanism over time.

Theorem 2 (Zero-Error Capacity vs Noiseless Privacy). *Assume, for all k , $X[k]$ is a discrete uncertain variable, $Y[k] = \mathfrak{M} \circ f \circ \psi_{i,v-i}^k \circ X[k]$, and $\mathfrak{M} \circ f$ is η -noiselessly private. For any $\psi_{i,v-i} \in \Psi_i$, the zero-error capacity of memory-less η -noiselessly-private channel is bounded by*

$$C_0 \leq \eta. \quad (6)$$

Proof. The rest of the proof follows from Proposition 2 in Appendix B in Supplementary Material and Theorem 1. $\square$

Theorem 2 shows that the zero-error capacity of memoryless noiselessly-private channels is upper bounded by the privacy budget η . This constraints dynamical systems that can be estimated or stabilized through privacy-preserving communication channels [17], [60].

C. Guarantee: Non-Stochastic Hypothesis Testing

In this part, the analysis is motivated by the definition of semantic security or indistinguishability under chosen plaintext attack [61]. Assume that an adversary selects $i \in \{1, \dots, n\}$, $x_i, x'_i \in \llbracket X_i \rrbracket$, and provides this information to the curator. The curator uses uncertain variable $\overline{X}_i : \Omega \rightarrow \llbracket \overline{X}_i \rrbracket := \{x_i, x'_i\}$ to construct uncertain variable $\overline{X} = (X_{-i}, \overline{X}_i)$. Fix $\psi_{i,v-i} \in \Psi_i$. The curator then generates a realization $\overline{X}(\omega)$, computes $Y(\omega) = \mathfrak{M} \circ f \circ \psi_{i,v-i}(\overline{X}(\omega))$, and provides $Y(\omega)$ to the adversary. The adversary tests whether the realization of the data of individual i is equal to x_i or x'_i knowing that it is bound to be one of those values and knowing that the value of the data of all the other individuals is fixed to v_{-i} . We define the hypothesis uncertain variable H using $g_H : \overline{X}(\omega) \mapsto H(\omega)$ as

$$H(\omega) = g_H(\overline{X}(\omega)) = \begin{cases} p_0, & \overline{X}_i(\omega) = x_i, \\ p_1, & \overline{X}_i(\omega) = x'_i. \end{cases}$$

The following theorem bounds the performance of the adversary for performing its hypothesis test.

Theorem 3 (Hypothesis Testing vs Noiseless Privacy). *Assume $Y = \mathfrak{M} \circ f \circ \psi_{i,v-i} \circ \overline{X}$ and $\mathfrak{M} \circ f$ is η -noiselessly private. Then, for any test T and any $\psi_{i,v-i} \in \Psi_i$, the performance of the adversary is bounded by*

$$\mathcal{P}(T) \leq \eta + 1. \quad (7)$$

Proof. See Appendix D in Supplementary Material. $\square$

Bounding the performance of a hypothesis-testing adversary by the privacy budget η in Theorem 3 is in essence close to identifiability [62], [63] for which privacy preservation relates to the potential of an adversary identifying the private data of individuals based on the received outputs.

D. Guarantee: Performance of Adversaries with Stochastic Priors

In this subsection, we briefly assume that the dataset is randomly distributed according to the probability density function p , i.e., for any Lebesgue-measurable³ set $\mathcal{X} \subseteq \llbracket X \rrbracket$, $\mathbb{P}\{X \in \mathcal{X}\} = \int_{x \in \llbracket X \rrbracket} \xi(x) \mu(x)$. We also consider an adversary that knows the realizations of all the entries of the dataset except the entry of the i -th individual. It constructs an estimate of the missing entry X_i using an estimator $\hat{X}_i(x_{-i}, \mathfrak{M} \circ f(X))$ using its prior information $X_{-i}(\omega) = x_{-i}$ and the response $\mathfrak{M} \circ f(X)$.

Theorem 4 (Stochastic Prior vs Noiseless Privacy). *Assume that $\rho = \inf_{x \in \llbracket X \rrbracket} \xi(x) > 0$, $\mathfrak{M} \circ f$ is η -noiselessly private, and $f^{-1} \circ \mathfrak{M}^{-1}(y)$ is a connected set for any $y \in \llbracket Y \rrbracket$. For any even p ,*

$$\begin{aligned} \mathbb{E}\{|X_i - \hat{X}_i(X_{-i}, \mathfrak{M} \circ f(X))|^p | X_{-i}(\omega) = x_{-i}\} \\ \geq \left(\frac{\rho \mu(\llbracket X_i \rrbracket)^{p+1}}{2^p(p+1)} \right) 2^{-\eta(p+1)}. \end{aligned} \quad (8)$$

Proof. See Appendix E in Supplementary Material. $\square$

The lower bound on the adversary's estimation performance in Theorem 4 is an decreasing function of the privacy budget η . Therefore, as expected and in-line with the earlier results, by decreasing η and thus increasing the lower bound in (8), we can reduce the adversary's ability to infringe on the privacy of any individual in the dataset even if the adversary knows the data of all the other individuals.

E. Guarantee: Stochastic Maximal Leakage

In this section, we can recreate the stochastic framework for information leakage in [22] by again endowing all the uncertain variables with a measure. This way, the maximal leakage from X to Y is given by

$$\begin{aligned} \mathcal{L}_c(X \rightarrow Y) &= \log \left(\sum_{y \in \llbracket Y \rrbracket} \max_{x \in \llbracket X \rrbracket : P_X(x) > 0} P_{Y|X}(y|x) \right) \\ &= I_\infty(X; Y). \end{aligned}$$

³We implicitly assume restriction to Euclidean spaces and use the σ -algebra of all Lebesgue-measurable sets (containing the Borel σ -algebra and inducing a complete measure space with the Lebesgue measure) with the uncertainty set and the mentioned probability distribution to define the probability spaces.

This notion of information leakage is the same as the maximum information in [46].

Theorem 5 (Maximal Leakage vs Noiseless Privacy). *Assume $Y = \mathfrak{M} \circ f \circ \psi_{i,v_{-i}} \circ X$ and $\mathfrak{M} \circ f$ is η -noiselessly private. Then,*

$$\mathcal{L}_c(X_i \rightarrow Y) \leq \eta. \quad (9)$$

Proof. Note that $\sup_{P_{Y|X}} \mathcal{L}_c(X \rightarrow Y) \leq H_0(Y)$ because of [22, Lemma 1 & Example 6]. Furthermore, $|\llbracket \mathfrak{M} \circ f \circ \psi_{i,v_{-i}} \circ X \rrbracket| = |\llbracket Y | X_{-i}(\omega) = v_{-i} \rrbracket| \leq 2^\eta$. $\square$

Evidently, the amount of the leaked information is upper bounded by the privacy budget η . Hence, by reducing the privacy budget, we can minimize the leaked information.

Remark 2. *The bound in (9) is tight in the sense that there exist problem formulations for which the equality can be achieved. First, note that for deterministic relationship between X and Y (e.g., when using a binning or quantization algorithm), $\mathcal{L}_c(X \rightarrow Y) = H_0(Y)$ [22, Example 6]. Furthermore, if we employ the setup of Remark 1 with a quantizer with 2^η levels, we get $H_0(Y) = \eta$.*

F. Property: Composition of Noiselessly-Private Mechanisms

Composition of differentially-private mechanisms [2], [64], [65] is an important result showing that the privacy budgets add up when reporting on multiple queries on the same private dataset. In what follows, we show that the same also applies to noiseless privacy.

Theorem 6 (Composition of Noiselessly-Private Mechanisms). *Let $\mathfrak{M}_1$ and $\mathfrak{M}_2$ be such that $\mathfrak{M}_1 \circ f$ and $\mathfrak{M}_2 \circ f$ are η_1 -noiseless private and η_2 -noiseless private, respectively. Then, $(\mathfrak{M}_1, \mathfrak{M}_2) \circ f$ is $(\eta_1 + \eta_2)$ -noiseless private.*

Proof. See Appendix F in Supplementary Material. $\square$

Theorem 6 shows that noiselessly-private mechanisms admit composition theorem, i.e., the privacy budgets of the mechanisms add up when reporting on multiple queries on the same private dataset.

G. Property: Post-Processing of Noiselessly-Private Mechanisms

Finally, an important property of differentially-private mechanisms and information-theoretic privacy is that the privacy guarantees do not weaken by post-processing privacy-preserving outputs [2]. In what follows, this also holds for noiselessly-private mechanisms as well.

Theorem 7 (Post-Processing of Noiselessly-Private Mechanisms). *Let $\mathfrak{M}$ be such that $\mathfrak{M} \circ f$ is η -noiseless private. Then, $g \circ \mathfrak{M} \circ f$ is also η -noiseless private for any mapping g .*

Proof. See Appendix G in Supplementary Material. $\square$

Theorem 7 shows that post-processing noiselessly-private mechanisms does not weaken their privacy guarantees, i.e., the privacy budget can only be decreased by post-processing.

We have summarized the results of this section in Table II.

IV. NOISELESS PRIVACY: SATISFACTION

We can ensure noiseless privacy using non-stochastic approaches, such as binning or quantization. To do so, first, we define linear quantizers.

Definition 3 (Linear Quantizer). *A q -level quantizer $\mathcal{Q} : [x_{\min}, x_{\max}] \rightarrow \{b_1, \dots, b_q\}$ is a piecewise constant function defined as*

$$\mathcal{Q}(x) = \begin{cases} b_1, & x \in [x_1, x_2), \\ b_2, & x \in [x_2, x_3), \\ \vdots & \vdots \\ b_{q-1}, & x \in [x_{q-1}, x_q), \\ b_q, & x \in [x_q, x_{q+1}], \end{cases}$$

where $(b_i)_{i=1}^q$ are distinct symbols and $x_1 \leq x_2 \leq \dots \leq x_q$ are real numbers such that $x_1 = x_{\min}$, $x_{q+1} = x_{\max}$, $x_{i+1} - x_i = (x_{\max} - x_{\min})/q$ for all $1 \leq i \leq q$.

We can show that linear quantizers can achieve noiseless privacy for any query on private datasets. This is proved in the next theorem.

Theorem 8. *Let $\llbracket f(X) | X \rrbracket \subseteq [y_{\min}, y_{\max}]$. Define sensitivity of query f as*

$$\begin{aligned} \mathcal{S}(f) &:= \sup_{x_{-i} \in \llbracket X_{-i} \rrbracket} \mu(f(\llbracket X_i \rrbracket \times \{x_{-i}\})) \\ &= \sup_{x_{-i} \in \llbracket X_{-i} \rrbracket} \sup_{x_i, x'_i \in \llbracket X_i \rrbracket} |f(x'_i, x_{-i}) - f(x_i, x_{-i})|. \end{aligned}$$

The mechanism $\mathcal{M} \circ f$ is η -noiseless private if $\mathcal{M}$ is a q -level quantizer with

$$q \leq \frac{2^\eta (y_{\max} - y_{\min})}{\mathcal{S}(f)}.$$

Proof. See Appendix H in Supplementary Material. $\square$

Theorem 8 shows that linear quantizers are a family functions that can provide noiseless privacy. However, in general, arbitrary piecewise constant functions can provide better utility guarantees. Note the similarity between the problem of finding optimal utility for a given budget in noiseless privacy and function approximation for a given accuracy with piecewise constant functions [66]. This relationship was explored for a different notion of non-stochastic privacy in [16]. Note that when the output of the query takes continuous values, e.g., coefficients of linear regression, we can still quantize the output. This quantization does not need to be coarse necessarily and the quantized values can be very close to the nominal values; however, the number of quantization levels determines the privacy guarantee. This demonstrates the usual narrative on privacy-utility trade-off. In discrete settings, the problem of finding optimal utility for a given budget in noiseless privacy can be rewritten as an integer program and sub-optimal computationally efficient algorithms can be used for solving it [57].

Definitions	Noiseless Privacy (Definition 1)	$\ \mathfrak{M} \circ f \circ X X_{-i}(\omega) = x_{-i}\ \leq 2^\eta, \forall x_{-i} \in [X_{-i}], \forall i$
	Local Noiseless Privacy (Definition 2)	$\ \mathfrak{M} \circ f_i \circ X X_{-i}(\omega) = x_{-i}\ \leq 2^\eta$ with $f_i : (x_i)_{i=1}^n \mapsto x_i$
Guarantees	Non-Stochastic Information (Theorem 1)	$0 \leq I_*(X_i; Y) \leq L_0^s(X_i; Y) \leq L_0(Y; X_i) \leq \eta$
	Zero-Error Capacity (Theorem 2)	$C_0 \leq \eta$
	Non-Stochastic Hypothesis Testing (Theorem 3)	$\mathcal{P}(T) \leq \eta + 1$
	Adversaries with Stochastic Priors (Theorem 4)	$\mathbb{E}\{ X_i - \hat{X}_i(X_{-i}, \mathfrak{M} \circ f(X)) ^p X_{-i}(\omega) = x_{-i}\} = \Omega(2^{-\eta(p+1)})$
	Stochastic Maximal Leakage (Theorem 5)	$\mathcal{L}_c(X_i \rightarrow Y) \leq \eta$
Properties	Composition (Theorem 6)	$\mathfrak{M}_1 \circ f \text{ is } \eta_1\text{-private} \implies (\mathfrak{M}_1, \mathfrak{M}_2) \circ f \text{ is } (\eta_1 + \eta_2)\text{-private}$
	Post-Processing (Theorem 7)	$\mathfrak{M} \circ f \text{ is } \eta\text{-private} \implies g \circ \mathfrak{M} \circ f \text{ is } \eta\text{-private}$

TABLE II: A summary of the definitions and the results presented in Sections III.

V. EXPERIMENTS

A. Synthetic Data: Reporting Average

In this section, we use randomly generated datasets to compare the utility of noiseless privacy and differential privacy. The dataset is assumed to contain the data of $n = 32$ individuals. The private data X_i of each data owner is generated or sampled from a uniform distribution over $[0, 1]$. We are interested in responding to the mean query:

$$f(X_1, \dots, X_n) = \frac{1}{n}(X_1 + \dots + X_n).$$

To ensure η -noiseless privacy, we use a linear quantizer $\mathcal{M}$ with $n2^\eta$ levels; see Theorem 8. In this case, the privacy-preserving report is $Y = \mathcal{M}(f(X_1, \dots, X_n))$. The difference between Y and $f(X_1, \dots, X_n)$ is the error caused by privacy-preserving mechanism $\mathcal{M}$. We want to compare the outcome of noiseless privacy with differential privacy. To ensure ϵ -differential privacy, we need to add a zero-mean Laplace noise with scale of n/ϵ . To be able to fairly compare noiseless privacy and differential privacy, for a given noiseless privacy budget η , we make sure to select differential privacy budget ϵ so that the mean or expected value of the error is the same in both cases. Empirically, this seems to be the case when $\epsilon = 2^\eta$. This is an interesting observation demonstrating the difference between the scale of privacy budget for noiseless privacy and differential privacy.

Figure 1 shows the empirical cumulative distribution function of the absolute value of the reporting error, i.e., privacy-preserving noise or quantization error, for η -noiseless privacy and 2^η -differential privacy with $\eta = 2$ [left], $\eta = 4$ [middle], and $\eta = 8$ [right]. The empirical cumulative distribution function of error for differential privacy shows a heavier tail in comparison with the noiseless privacy case. In fact the error for the noiseless privacy has a bounded support. This implies that in applications where we need to be able to provide an *a priori* bound for the error, noiseless privacy is more palatable.

B. Energy Data: Reporting Aggregate

For this part, we use a publicly available dataset from the Ausgrid⁴ containing half-hour smart meter measurements for 300 randomly-selected homes with rooftop solar systems over the period from 1 July 2010 to 30 June 2013. In this paper, we use the data over July 2012 to June 2013.

Let $x_{i,t}$ denote the consumption of house i at day t . We consider reporting aggregate statistics

$$y_t = f((x_{i,t})_{i=1}^n) = \frac{1}{n}(x_{1,t} + \dots + x_{n,t}), \forall t.$$

Here, f denotes the query. We particularly use the mechanism in Theorem 8 to report noiselessly-private outputs. In this experiment, we test the ability of an adversary for inferring if a particular household has contributed to the aggregate or not as in [67]. We use a game, as in [67], [68], to evaluate the ability of the adversary. The setup of the game is summarized in Figure 2. At first, the adversary can select two households i_0, i_1 . The curator select one of those households uniformly at random i_j . It also selects an additional $n-1$ households. Then it reports the privacy-preserving aggregate outputs. Based on the reported output, the adversary guesses the participating household $i_{\hat{j}}$. The adversary's success or advantage is then defined as $\text{Adv} := 2|\mathbb{P}\{j = \hat{j}\} - 1/2|$. Small Adv means that the adversary is as successful as randomly guessing and large Adv implies that the adversary is successful in recognizing the household participating in the aggregate.

Similar to [67], we use three adversary policies. The first one is based on correlation. In this case, the adversary selects $j \in \{0, 1\}$ based on the correlation between $(x_{i_j,t})_t$ and $(y_t)_t$. The second policy is based on mean square error. In this case, the adversary selects $j \in \{0, 1\}$ by minimizing the square error $\|(x_{i_j,t})_t - (y_t)_t\|_2$. Finally, the last policy uses the relative peaks of each load profile $(x_{i_0,t})_t$, $(x_{i_1,t})_t$, and $(y_t)_t$. In this case, the adversary selects $j \in \{0, 1\}$ based on

⁴<https://www.ausgrid.com.au/Industry/Innovation-and-research/Data-to-share/Solar-home-electricity-data>

Fig. 1: Empirical cumulative distribution function of the absolute value of the reporting error for η -noiseless privacy and $\epsilon(= 2^\eta)$ -differential privacy with $\eta = 2$ [left], $\eta = 4$ [middle], and $\eta = 8$ [right]. For a fair comparison of noiseless privacy and differential privacy, we select the differential privacy budget ϵ as a function of the noiseless privacy budget η so that the expected value of the error remains the same (in average they have the same utility degradation). Empirically, this happens for $\epsilon = 2^\eta$. This is an interesting observation demonstrating the difference between the scale of privacy budget for noiseless privacy and differential privacy.

Fig. 2: The timing of a game used for evaluating the ability of an adversary in guessing if the data of a particular individual belongs to a publicly-released noiselessly-private aggregate statistics. Note that $x_{i,t}$ denotes the consumption of house i at day t .

the most common peaks between $(x_{i_0, t})_t$ and $(y_t)_t$, or $(x_{i_1, t})_t$ and $(y_t)_t$.

Figure 3 illustrates the advantage of the adversary Adv when using the correlation-based policy (left), the mean square error policy (center), and the peak-based policy (right). As η grows larger, the adversary’s advantage tends toward the non-private case in [67]. Clearly, even for moderate η when considering small groups, the adversary’s advantage is very small. This is not the case for non-private outputs as observed in [67]. For instance, for small groups and moderate privacy budgets, such as $n = 4$ and $\eta = 2$ or $n = 8$ and $\eta = 3$, the adversary’s advantage is negligible (almost zero). This shows that combining noiseless privacy with aggregation is an excellent tool for providing privacy to individuals.

C. Energy Data: Reporting Single Consumption

Non-intrusive load monitoring provides tools for extracting appliance-specific energy consumption statistics from the smart meter readings of a household and is one of privacy concerns behind releasing energy data [69], [70]. In this section, we use Theorem 8 to report high-frequency energy consumption of a household in a privacy-preserving manner using local noiseless privacy. We then proceed to see the effect

of privacy budget on an adversary performing non-intrusive load monitoring.

We use the low frequency data from the first house in the REDD database⁵ [71], which contains the consumption of various appliances in the house every 3-4 seconds. This data in conjunction with the consumption of the entire house is used for training and verification of a non-intrusive load monitoring algorithm. The consumption of the entire house is measured every second. The data is for the period of April 23–May 21, 2011. The part of the data prior to April 30th is used for training and the rest for validation purposes. We select the top 5 appliances in energy consumption for disaggregation purposes, namely, fridge, microwave, socket (in the kitchen), light, and dish washer. For non-intrusive load monitoring, we have used the NILMTK⁶ toolbox in Python [72]. We have used a frequently utilized combinatorial optimization method for non-intrusive load monitoring. We report the success of the non-intrusive load monitoring using the f -score.

Figure 4 shows the f -score of the non-intrusive load monitoring algorithm based on combinatorial optimization versus the privacy budget. As we can see the f -score gets rapidly bad as η decrease. This means an adversary would not be able to identify the appliances that are used within the household reliably. This illustrates the power of local noiseless privacy in reporting energy consumption of households for analysis while protecting the privacy of the households.

D. Transport Data: Reporting Individual Source-Destinations

We use New York City Taxi Cab trips⁷ for 2014. We use the first million trips and focus on trips that begin and end within the New York City in Figure 5. Here, we consider reporting the start-end point of the taxi rides in a locally noiselessly private manner. In this subsection, we again use Theorem 8 to report start-end points of taxi rides in the New York City in a privacy-preserving manner using local noiseless privacy. In

⁵<http://redd.csail.mit.edu/>

⁶<https://github.com/nilmtk/nilmtk>

⁷<https://www.kaggle.com/kentonnlp/2014-new-york-city-taxi-trips>

Fig. 3: The advantage of the adversary Adv when using the correlation-based policy (left), the mean square error policy (center), and the peak-based policy (right).

Fig. 4: f -score of the non-intrusive load monitoring algorithm based on combinatorial optimization versus the privacy budget.

Fig. 7: Performance of the machine learning model versus the privacy budget for various training dataset sizes.

Fig. 5: Map of New York City: latitude in $[40.92^\circ, 40.49^\circ]$ and longitude in $[-74.27^\circ, -73.68^\circ]$

Fig. 6: Portion of taxi rides with unique start-end points (i.e., routes) versus the privacy budget.

Fig. 8: Advantage of the adversary in membership inference attacks versus the privacy budget for various training dataset sizes.

this case, for each η , we split the latitude and the longitude into $2^{\eta/2}$ boxes. Therefore, the privacy of the total privacy budget for the reported outputs is η , following Theorem 6.

Figure 6 illustrates the portion of unique start-end points of taxi rides versus the privacy budget. As we can see, the portion of unique start-end points is negligible for small η . This means an adversary would not be able to attribute a specific taxi ride to an individual, thus protecting the privacy of individuals.

E. Financial Data: Machine Learning for Loan Interest Rates

Finally, we demonstrate the effect of noiseless privacy on membership inference attacks. To this aim, we use a dataset

of anonymized loan application⁸ containing loan information, such as loan size, and applicant information, such as credit rating, state of residence, and age. We only focus on 10,000 individuals in this dataset. We remove unique identifiers, such as identity numbers, and irrelevant information, such as web addresses. We further use integer numbers to encode categorical attributes, such as state of residence. We use Principal Component Analysis (PCA) for selecting top ten features. We also scale all the features to be between zero and one. The features, i.e., the inputs to the machine learning model, are denoted by $x_i \in \mathbb{R}^n$. The interest rates of the loans are the outputs of the machine learning model and are denoted by y_i . Before training the machine learning model, we again use Theorem 8 to make the data of each individual, captured in x_i , noiselessly private. If we make each feature η -noiselessly private, the total feature vector or input is 10η -noiselessly private. Let us denote the private version of x_i by $\bar{x}_i$. We train a linear regression model $\theta^\top x_i$ by solving the optimization problem $\min_{\theta} (1/N) \sum_{i=1}^N \|y_i - \theta^\top \bar{x}_i\|_2^2$, where N is the size of the training dataset. For assessing the performance of the trained model, we use the original non-private data to compute the training cost $(1/N) \sum_{i=1}^N \|y_i - \theta^\top x_i\|_2^2$. Figure 7 shows the performance of the machine learning model captured using the training cost versus the privacy budget η for various training dataset sizes. The performance of the machine learning model clearly improves as η increases. However, we expect that the privacy guarantees to worsen. We demonstrate this by a membership inference attack.

For membership inference, we use the membership experiment from [73]. We train a model θ based on the training dataset. We select b with equal probability from $\{0, 1\}$. Then, select a single record from the training dataset if $b = 1$ or from outside the training dataset if $b = 0$. We transmit the selected record to the adversary. The adversary estimates the realization of random variable b , denoted by $\hat{b} \in \{0, 1\}$, based on the selected record and trained model following the threshold-based adversary in [73]. The adversary's advantage (in comparison to randomly selecting an estimate) is given by $\text{Adv} := 2|\mathbb{P}\{\hat{b} = b\} - 1/2|$. Figure 8 shows the advantage of the adversary in membership inference attacks versus the privacy budget η for various training dataset sizes. As expected, the attacks become more successful as η increases.

VI. DISCUSSIONS AND FUTURE WORK

In this paper, we defined noiseless privacy, as a non-stochastic rival to differential privacy, requiring that the outputs of the mechanism to attain very few values while varying the data of an individual remains. We proved that noiseless-private mechanisms admit composition theorem and post-processing does not weaken their privacy guarantees. We proved that quantization operators can ensure noiseless privacy. We finally illustrated the privacy merits of noiseless privacy and local noiseless privacy using multiple datasets in energy, transport, and finance.

An important avenue for future research involves investigating the robustness of noiselessly-private outcomes to side

channel information. Experimental and theoretical investigations must be done to understand the success of adversaries that can access additional information about the private data from other sources. Finally, we must investigate the merits of existing non-stochastic privacy policies, e.g., k -anonymity or ℓ -diversity, in terms of noiseless privacy.

REFERENCES

- [1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*, pp. 265–284, Springer, 2006.
- [2] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.
- [3] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, 2013.
- [4] H. Yamamoto, "A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers," *IEEE Transactions on Information Theory*, vol. 29, no. 6, pp. 918–923, 1983.
- [5] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.
- [6] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.
- [7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian, " ℓ -diversity: privacy beyond k -anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, pp. 24–24, 2006.
- [8] R. Bild, K. A. Kuhn, and F. Prasser, "SafePub: A truthful data anonymization algorithm with strong privacy guarantees," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 1, pp. 67–87, 2018.
- [9] G. Poulis, A. Gkoulalas-Divanis, G. Loukides, S. Skiadopoulos, and C. Tryfonopoulos, "SECRETA: A tool for anonymizing relational, transaction and RT-datasets," in *Medical Data Privacy Handbook* (A. Gkoulalas-Divanis and G. Loukides, eds.), pp. 83–109, Springer International Publishing, 2015.
- [10] R. Bhaskar, A. Bhowmick, V. Goyal, S. Laxman, and A. Thakurta, "Noiseless database privacy," in *International Conference on the Theory and Application of Cryptology and Information Security*, pp. 215–232, Springer, 2011.
- [11] S. U. Nabar, B. Marthi, K. Kenthapadi, N. Mishra, and R. Motwani, "Towards robustness in query auditing," in *Proceedings of the 32nd International Conference on Very Large Data Bases*, pp. 151–162, VLDB Endowment, 2006.
- [12] J. Bambauer, K. Muralidhar, and R. Sarathy, "Fool's gold: an illustrated critique of differential privacy," *Vanderbilt Journal of Entertainment & Technology Law*, vol. 16, p. 701, 2013.
- [13] F. K. Dankar and K. El Emam, "Practicing differential privacy in health care: A review," *Transactions on Data Privacy*, vol. 6, no. 1, pp. 35–67, 2013.
- [14] J. Mervis, "Researchers object to census privacy measure," *Science*, vol. 363, no. 6423, pp. 114–114, 2019.
- [15] F. Farokhi, J. Milosevic, and H. Sandberg, "Optimal state estimation with measurements corrupted by Laplace noise," in *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pp. 302–307, 2016.
- [16] F. Farokhi, "Development and analysis of deterministic privacy-preserving policies using non-stochastic information theory," *IEEE Transactions on Information Forensics and Security*, vol. 14, pp. 2567–2576, Oct 2019.
- [17] G. N. Nair, "A nonstochastic information theory for communication and state estimation," *IEEE Transactions on Automatic Control*, vol. 58, no. 6, pp. 1497–1510, 2013.
- [18] P. Duan, F. Yang, S. L. Shah, and T. Chen, "Transfer zero-entropy and its application for capturing cause and effect relationship between variables," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 3, pp. 855–867, 2014.
- [19] T. J. Lim and M. Franceschetti, "A deterministic view on the capacity of bandlimited functions," in *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 691–698, IEEE, 2014.
- [20] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 8–19, 1956.

⁸<https://www.kaggle.com/wendykan/lending-club-loan-data>

- [21] F. Farokhi, "Non-stochastic hypothesis testing with application to privacy against hypothesis-testing adversary," in *Decision and Control (CDC), 2019 IEEE 58th Conference on*, 2019.
- [22] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1625–1657, 2020.
- [23] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 111–125, 2008.
- [24] J. Su, A. Shukla, S. Goel, and A. Narayanan, "De-anonymizing web browsing data with social networks," in *Proceedings of the 26th International Conference on World Wide Web*, pp. 1261–1269, 2017.
- [25] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, p. 1376, 2013.
- [26] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings* (M. Agrawal, D. Du, Z. Duan, and A. Li, eds.), pp. 1–19, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.
- [27] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pp. 429–438, 2013.
- [28] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Advances in Neural Information Processing Systems*, pp. 2879–2887, 2014.
- [29] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," in *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pp. 277–286, 2008.
- [30] R. Hall, A. Rinaldo, and L. Wasserman, "Random differential privacy," *Journal of Privacy and Confidentiality*, vol. 4, no. 2, pp. 43–59, 2012.
- [31] A. Padakandla, P. Kumar, and W. Szpankowski, "Preserving privacy and fidelity via Ehrhart theory," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 696–700, 2018.
- [32] A. D. Wyner, "The wire-tap channel," *Bell System Technical Journal*, vol. 54, no. 8, pp. 1355–1387, 1975.
- [33] T. Courtade, "Information masking and amplification: The source coding setting," in *Proceedings of the IEEE International Symposium on Information Theory Proceedings (ISIT)*, pp. 189–193, 2012.
- [34] H. Yamamoto, "A rate-distortion problem for a communication system with a secondary decoder to be hindered," *IEEE Transactions on Information Theory*, vol. 34, no. 4, pp. 835–842, 1988.
- [35] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, and C. Palamidessi, "On the relation between differential privacy and quantitative information flow," in *Automata, Languages and Programming* (L. Aceto, M. Henzinger, and J. Sgall, eds.), vol. 6756 of *Lecture Notes in Computer Science*, pp. 60–76, Springer Berlin Heidelberg, 2011.
- [36] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1401–1408, 2012.
- [37] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275, IEEE, 2017.
- [38] M. S. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi, and G. Smith, "Axioms for information leakage," in *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, pp. 77–92, IEEE, 2016.
- [39] F. Farokhi, H. Sandberg, I. Shames, and M. Cantoni, "Quadratic Gaussian privacy games," in *Proceedings of the 54th IEEE Conference on Decision and Control*, pp. 4505–4510, 2015.
- [40] M. J. Wainwright, M. I. Jordan, and J. C. Duchi, "Privacy aware learning," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 1430–1438, 2012.
- [41] Y. Liang, H. V. Poor, and S. Shamai, "Information theoretic security," *Foundations and Trends in Communications and Information Theory*, vol. 5, no. 4–5, pp. 355–580, 2009.
- [42] L. Lai, S.-W. Ho, and H. V. Poor, "Privacy–security trade-offs in biometric security systems—Part I: Single use case," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 1, pp. 122–139, 2011.
- [43] Z. Li and T. Oechtering, "Privacy on hypothesis testing in smart grids," in *IEEE Information Theory Workshop (ITW) 2015, Jeju, Korea, Oct. 11-15, 2015*, pp. 337–341, 2015.
- [44] G. Bassi, M. Skoglund, and P. Piantanida, "Lossy communication subject to statistical parameter privacy," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1031–1035, 2018.
- [45] F. Farokhi and H. Sandberg, "Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4726–4734, 2018.
- [46] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *Advances in Neural Information Processing Systems*, pp. 2350–2358, 2015.
- [47] S. Zhou, J. Lafferty, and L. Wasserman, "Compressed and privacy-sensitive sparse regression," *IEEE Transactions on Information Theory*, vol. 55, no. 2, pp. 846–866, 2009.
- [48] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [49] R. Bassily, A. Groce, J. Katz, and A. Smith, "Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 439–448, IEEE, 2013.
- [50] R. V. L. Hartley, "Transmission of information," *Bell System Technical Journal*, vol. 7, no. 3, pp. 535–563, 1928.
- [51] A. N. Kolmogorov and V. M. Tikhomirov, " ϵ -entropy and ϵ -capacity of sets in function spaces," *Uspekhi Matematicheskikh Nauk*, vol. 14, no. 2, pp. 3–86, 1959. English translation American Mathematical Society Translations, series 2, vol. 17, pp. 277–364.
- [52] A. Renyi, "On measures of entropy and information," in *Proc. of the Fourth Berkeley Symp. on Math. Statist. and Prob.*, vol. 1, pp. 547–561, 1961.
- [53] D. Jagerman, " ϵ -entropy and approximation of bandlimited functions," *SIAM Journal on Applied Mathematics*, vol. 17, no. 2, pp. 362–377, 1969.
- [54] G. N. Nair, "A nonstochastic information theory for feedback," in *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 1343–1348, 2012.
- [55] P. Duan, F. Yang, S. L. Shah, and T. Chen, "Transfer zero-entropy and its application for capturing cause and effect relationship between variables," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 3, pp. 855–867, 2015.
- [56] M. Wiese, K. H. Johansson, T. J. Oechtering, P. Papadimitratos, H. Sandberg, and M. Skoglund, "Uncertain wiretap channels and secure estimation," in *Information Theory (ISIT), 2016 IEEE International Symposium on*, pp. 2004–2008, 2016.
- [57] N. Ding and F. Farokhi, "Developing non-stochastic privacy-preserving policies using agglomerative clustering," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3911–3923, 2020.
- [58] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, "Broadening the scope of differential privacy using metrics," in *International Symposium on Privacy Enhancing Technologies Symposium*, pp. 82–102, Springer, 2013.
- [59] G. Barthe and B. Kopf, "Information-theoretic bounds for differentially private mechanisms," in *2011 IEEE 24th Computer Security Foundations Symposium*, pp. 191–204, 2011.
- [60] A. S. Matveev and A. V. Savkin, "Shannon zero error capacity in the problems of state estimation and stabilization via noisy communication channels," *International Journal of Control*, vol. 80, no. 2, pp. 241–255, 2007.
- [61] J. Katz and Y. Lindell, *Introduction to Modern Cryptography, Second Edition*. Chapman & Hall/CRC Cryptography and Network Security Series, Taylor & Francis, 2 ed., 2014.
- [62] W. Wang, L. Ying, and J. Zhang, "On the relation between identifiability, differential privacy, and mutual-information privacy," *IEEE Transactions on Information Theory*, vol. 62, no. 9, pp. 5018–5029, 2016.
- [63] A. Bkakra, N. Cuppens-Boulahia, and F. Cuppens, "Linking differential identifiability with differential privacy," in *International Conference on Information and Communications Security*, pp. 232–247, 2018.
- [64] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60, 2010.
- [65] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 4037–4049, 2017.
- [66] H. Konno and T. Kuno, "Best piecewise constant approximation of a function of single variable," *Operations research letters*, vol. 7, no. 4, pp. 205–210, 1988.
- [67] N. Buescher, S. Boukoro, S. Bauregger, and S. Katzenbeisser, "Two is not enough: Privacy assessment of aggregation schemes in smart metering," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 4, pp. 198–214, 2017.
- [68] J.-M. Bohli, C. Sorge, and O. Ugus, "A privacy model for smart metering," in *2010 IEEE International Conference on Communications Workshops*, pp. 1–5, 2010.

- [69] A. Zoha, A. Gluhak, M. A. Imran, and S. Rajasegarar, "Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey," *Sensors*, vol. 12, no. 12, pp. 16838–16866, 2012.
- [70] O. Parson, S. Ghosh, M. Weal, and A. Rogers, "Non-intrusive load monitoring using prior models of general appliance types," in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012.
- [71] J. Z. Kolter and M. J. Johnson, "REDD: A public data set for energy disaggregation research," in *Workshop on Data Mining Applications in Sustainability (SIGKDD)*, vol. 25, pp. 59–62, 2011.
- [72] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, "NILMTK: An open source toolkit for non-intrusive load monitoring," in *5th International Conference on Future Energy Systems (ACM e-Energy)*, 2014.
- [73] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018.

Farhad Farokhi (S'11–M'15–SM'20) received B.Sc. and M.Sc. degrees from Sharif University of Technology in 2008 and 2010, respectively, the Ph.D. degree from the KTH Royal Institute of Technology in 2014.

In 2014, he joined The University of Melbourne, where he is currently a Lecturer (equivalent to Assistant Professor in North America). From 2018–2020, he was also a Research Scientist at the CSIRO's Data61. In 2013, he held a Visiting Researcher position at the University of California, Berkeley.

Dr. Farokhi has been the recipient of the VESKI Victoria Fellowship from the Victoria State Government, Australia, and the McKenzie Fellowship, the 2015 Early Career Researcher Award, and the Melbourne School of Engineering (MSE) Excellence Award for Early Career Research from The University of Melbourne. He has been involved in multiple projects on data privacy and cyber-security funded by the Australian Research Council, the Defence Science and Technology Group, the Department of the Prime Minister and Cabinet, the Department of Environment and Energy, and the CSIRO. He is the associate editor for IET Smart Grid and Results in Control and Optimization.

APPENDIX A

RELATIONSHIP BETWEEN MAXIMIN INFORMATION AND
NON-STOCHASTIC INFORMATION LEAKAGE

Proposition 1. For discrete uncertain variable Y , $I_*(X; Y) \leq L_0(X; Y)$.

Proof. First, we prove that $I_*(X; Y) \leq L_0(X; Y)$. Let $I_*(X; Y) = m$. Then, $\llbracket X|Y \rrbracket_* = \{P_1, \dots, P_{2^m}\}$. Each P_i is non-empty. Therefore, there exists at least one x such that $x \in P_i$. Note that x must also belong to $\llbracket X|y \rrbracket$ for any $y \in \llbracket Y|x \rrbracket$. We prove that $\llbracket X|y \rrbracket \subseteq P_i$. Assume that this not the case. Therefore, there exist an element of $x' \in \llbracket X|y \rrbracket$, distinct from x , that belongs to another P_j , $j \neq i$, because $\{P_1, \dots, P_{2^m}\}$ covers $\llbracket X \rrbracket$. We know that P_i and P_j are $\llbracket X|Y \rrbracket$ -overlap isolated by the definition of partition $\llbracket X|Y \rrbracket_*$. On the other hand, we evidently x, x' are $\llbracket X|Y \rrbracket$ -overlap connected (by the definition of $\llbracket X|Y \rrbracket$ -overlap connectedness). This is a contradiction and thus $\llbracket X|y \rrbracket$ must be a subset of P_i . This results in $|\llbracket X|y \rrbracket| \leq |P_i|$ and hence

$$\min_{y \in \llbracket Y \rrbracket} |\llbracket X|y \rrbracket| \leq |P_i|, \quad \forall i \in \{1, \dots, 2^m\}. \quad (10)$$

On the other hand, $\bigcup_{i=1}^{2^m} P_i = \llbracket X \rrbracket$ because $\{P_1, \dots, P_{2^m}\}$ is a partition for $\llbracket X \rrbracket$. Because of the non-overlapping nature of the sets $\{P_1, \dots, P_{2^m}\}$, we get

$$\sum_{i=1}^{2^m} |P_i| = |\llbracket X \rrbracket|. \quad (11)$$

Combining (10) and (11) results in $2^m \min_{y \in \llbracket Y \rrbracket} |\llbracket X|y \rrbracket| \leq |\llbracket X \rrbracket|$. This implies that $I_*(X; Y) \leq L_0(X; Y)$. Similarly, we can show that $I_*(Y; X) \leq L_0(Y; X)$. By symmetry of the maximin information [17], i.e., $I_*(X; Y) = I_*(Y; X)$, we get that $I_*(X; Y) \leq L_0(X; Y)$ and $I_*(X; Y) \leq L_0(Y; X)$. This concludes the proof. $\square$

APPENDIX B

RELATIONSHIP BETWEEN ZERO-ERROR CAPACITY AND
NON-STOCHASTIC INFORMATION LEAKAGE

In what follows, we only consider sequence of discrete uncertain variables $Y[k]$.

Proposition 2. Any memory-less uncertain channel satisfies $C_0 \leq \sup_{\llbracket X[k] \rrbracket \subseteq \mathbb{X}} L_0^s(X[k]; Y[k])$.

Proof. Let $\aleph(y[k], \dots, y[1])$ denote the statement $Y[k](\omega) = y[k], \dots, Y[1](\omega) = y[1]$. Note that

$$\begin{aligned} & 2^{L_0(X[k], \dots, X[1]; Y[k], \dots, Y[1])} \\ &= \max_{(y[i])_{i=1}^k \in \llbracket Y \rrbracket^k} \frac{|\llbracket X[k], \dots, X[1] \rrbracket|}{|\llbracket X[k], \dots, X[1] | \aleph(y[k], \dots, y[1]) \rrbracket|} \\ &= \frac{|\llbracket X[k], \dots, X[1] \rrbracket|}{\min_{(y[i])_{i=1}^k \in \llbracket Y \rrbracket^k} |\llbracket X[k], \dots, X[1] | \aleph(y[k], \dots, y[1]) \rrbracket|} \\ &= \frac{\prod_{\ell=1}^k |\llbracket X[\ell] \rrbracket|}{\prod_{\ell=1}^k \min_{y[\ell] \in \llbracket Y \rrbracket} |\llbracket X[\ell] | Y[\ell](\omega) = y[\ell] \rrbracket|} \\ &= \prod_{\ell=1}^k \frac{|\llbracket X[\ell] \rrbracket|}{\min_{y[\ell] \in \llbracket Y \rrbracket} |\llbracket X[\ell] | Y[\ell](\omega) = y[\ell] \rrbracket|} \end{aligned}$$

$$\begin{aligned} &= \prod_{\ell=1}^k L_0(X[\ell]; Y[\ell]) \\ &= (L_0(X[\ell]; Y[\ell]))^k. \end{aligned}$$

Thus, $L_0(X[k], \dots, X[1]; Y[k], \dots, Y[1]) = k L_0(X[\ell]; Y[\ell])$, and, as a result, $L_0(X[k], \dots, X[1]; Y[k], \dots, Y[1])/k = L_0(X[\ell]; Y[\ell])$. Similarly, we can show that $L_0(Y[k], \dots, Y[1]; X[k], \dots, X[1])/k = L_0(Y[\ell]; X[\ell])$. Combining these inequalities with Proposition 1 in this paper and Theorem 4.1 in [17] proves the result. $\square$

APPENDIX C

PROOF OF PROPOSITION 1

We have $Y(\omega) = \mathfrak{M} \circ f(v_1, \dots, v_{i-1}, X_i(\omega), v_{i+1}, \dots, v_n)$ for all $\omega \in \Omega$ because, by definition, $\psi_{i, v_{-i}} \circ X(\omega) = (v_1, \dots, v_{i-1}, X_i(\omega), v_{i+1}, \dots, v_n)$ for all $\omega \in \Omega$. This implies that $\llbracket Y \rrbracket = \{Y(\omega) = \mathfrak{M} \circ f(v_1, \dots, v_{i-1}, X_i(\omega), v_{i+1}, \dots, v_n), \forall \omega \in \Omega\} = \{Y(\omega) = \mathfrak{M} \circ f(X(\omega)), X_j(\omega) = v_j, \forall j \in \{1, \dots, n\} \setminus \{i\}, \forall \omega \in \Omega\} = \llbracket Y | X_{-i}(\omega) = v_{-i} \rrbracket$. Note that

$$\begin{aligned} 2^{L_0^s(X_i; Y)} &\leq 2^{L_0(Y; X_i)} = \sup_{x_i \in \llbracket X_i \rrbracket} \frac{|\llbracket Y \rrbracket|}{|\llbracket Y | X_i(\omega) = x_i \rrbracket|} \\ &= |\llbracket Y \rrbracket| \\ &= |\llbracket Y | X_{-i}(\omega) = v_{-i} \rrbracket| \\ &\leq 2^\eta, \end{aligned} \quad (12)$$

where the second equality follows from that the realization of Y can be uniquely determined based on the realization of X_i , i.e., $\llbracket Y | X_i(\omega) = x_i \rrbracket$ is a singleton. The last inequality in (12) follows from the definition of noiseless privacy. Therefore, $L_0^s(X_i; Y) \leq L_0(Y; X_i) \leq \eta$. The rest follows from Proposition 1.

APPENDIX D

PROOF OF THEOREM 3

Since Y is a discrete uncertain variable, for any test T , the upper bound in [21] states that the performance of the adversary is bounded from the above by $\mathcal{P}(T) \leq \log_2(|\llbracket Y | X_i(\omega) = x_i \rrbracket \Delta \llbracket Y | X_i(\omega) = x'_i \rrbracket|)$. Since $Y = \mathfrak{M} \circ f \circ \psi_{i, v_{-i}} \circ \bar{X}$, we know that $\llbracket Y | X_i(\omega) = x_i \rrbracket = \llbracket Y | X_i(\omega) = x_i, X_{-i}(\omega) = v_{-i} \rrbracket$ and $\llbracket Y | X_i(\omega) = x'_i \rrbracket = \llbracket Y | X_i(\omega) = x'_i, X_{-i}(\omega) = v_{-i} \rrbracket$. Now, note that $\llbracket Y | X_i(\omega) = x_i, X_{-i}(\omega) = v_{-i} \rrbracket \Delta \llbracket Y | X_i(\omega) = x'_i, X_{-i}(\omega) = v_{-i} \rrbracket \subseteq \llbracket Y | X_{-i}(\omega) = v_{-i} \rrbracket$. Therefore, $|\llbracket Y | X_i(\omega) = x_i \rrbracket \Delta \llbracket Y | X_i(\omega) = x'_i \rrbracket| = |\llbracket Y | X_i(\omega) = x_i, X_{-i}(\omega) = v_{-i} \rrbracket \Delta \llbracket Y | X_i(\omega) = x'_i, X_{-i}(\omega) = v_{-i} \rrbracket| \leq |\llbracket Y | X_{-i}(\omega) = v_{-i} \rrbracket| \leq 2 \times 2^\eta$. This concludes the proof.

APPENDIX E

PROOF OF THEOREM 4

For fixed $X_{-i}(\omega) = x_{-i}$, define function $g_{x_{-i}}(\cdot)$ such that $g_{x_{-i}}(X_i) = \mathfrak{M} \circ f(X)$. There must exists $y' \in \llbracket Y | X_{-i}(\omega) = x_{-i} \rrbracket$ such that $\mu(g_{x_{-i}}^{-1}(y')) \geq \mu(\llbracket X_i \rrbracket) 2^{-\eta}$. We prove this contradiction. If this statement does not hold,

$\mu(g_{x_{-i}}^{-1}(y)) < \mu(\llbracket X_i \rrbracket)2^{-\eta}$ for all $y \in \llbracket Y|X_{-i}(\omega) = x_{-i} \rrbracket$ and thus

$$\begin{aligned} \mu(\llbracket X_i \rrbracket) &= \mu\left(\bigcup_{y \in \llbracket Y|X_{-i}(\omega) = x_{-i} \rrbracket} g^{-1}(y)\right) \\ &\leq \sum_{y \in \llbracket Y|X_{-i}(\omega) = x_{-i} \rrbracket} \mu(g^{-1}(y)) \\ &< \sum_{y \in \llbracket Y|X_{-i}(\omega) = x_{-i} \rrbracket} \mu(\llbracket X_i \rrbracket)2^{-\eta} \\ &= \mu(\llbracket X_i \rrbracket)2^{-\eta} |\llbracket Y|X_{-i}(\omega) = x_{-i} \rrbracket| \\ &\leq \mu(\llbracket X_i \rrbracket). \end{aligned}$$

This is a contradiction. Hence, we get

$$\begin{aligned} &\mathbb{E}\{|X_i - \hat{X}_i(X_{-i}, \mathfrak{M} \circ f(X))|^p | X_{-i}\} \\ &\geq \rho \sum_{y \in \llbracket Y|X_{-i}(\omega) = x_{-i} \rrbracket} \int_{g^{-1}(y)} |X_i - \hat{X}_i(X_{-i}, \mathfrak{M} \circ f(X))|^p d\mu(X_i) \\ &\geq \rho \int_{g^{-1}(y')} |X_i - \hat{X}_i(X_{-i}, \mathfrak{M} \circ f(X))|^p d\mu(X_i), \end{aligned}$$

where $\rho = \inf_{X \in \llbracket X \rrbracket} \xi(X)$. Since $g^{-1}(y')$ is a connected set⁹, there must exist $\underline{x}_i, \bar{x}_i$ such that closure of the $g^{-1}(y')$ is equal to $[\underline{x}_i, \bar{x}_i]$. Hence, we get

$$\begin{aligned} &\int_{g^{-1}(y')} |X_i - \hat{X}_i(X_{-i}, \mathfrak{M} \circ f(X))|^p d\mu(X_i) \\ &= \int_{\underline{x}_i}^{\bar{x}_i} |X_i - \hat{X}_i(X_{-i}, \mathfrak{M} \circ f(X))|^p d\mu(X_i) \\ &\geq \int_{\underline{x}_i}^{\bar{x}_i} (z - (\underline{x}_i + \bar{x}_i)/2)^p dz \\ &\geq \frac{(\bar{x}_i - \underline{x}_i)^{p+1}}{2^p(p+1)} \\ &= \frac{\mu(g^{-1}(y'))^{p+1}}{2^p(p+1)} \\ &\geq \frac{\mu(\llbracket X_i \rrbracket)^{p+1} 2^{-\eta(p+1)}}{2^p(p+1)}. \end{aligned}$$

APPENDIX F PROOF OF THEOREM 6

Note that $\llbracket (\mathfrak{M}_1, \mathfrak{M}_2) \circ f(X) | X_{-i}(\omega) = x_{-i} \rrbracket = \llbracket \mathfrak{M}_1 \circ f(X), \mathfrak{M}_2 \circ f(X) | X_{-i}(\omega) = x_{-i} \rrbracket \subseteq \llbracket \mathfrak{M}_1 \circ f(X) | X_{-i}(\omega) = x_{-i} \rrbracket \times \llbracket \mathfrak{M}_2 \circ f(X) | X_{-i}(\omega) = x_{-i} \rrbracket$, and as a result $\mu(\llbracket (\mathfrak{M}_1, \mathfrak{M}_2) \circ f(X) | X_{-i}(\omega) = x_{-i} \rrbracket) \leq \mu(\llbracket \mathfrak{M}_1 \circ f(X) | X_{-i}(\omega) = x_{-i} \rrbracket) \times \mu(\llbracket \mathfrak{M}_2 \circ f(X) | X_{-i}(\omega) = x_{-i} \rrbracket)$. Hence, $\log_e(\mu(\llbracket (\mathfrak{M}_1, \mathfrak{M}_2) \circ f(X) | X_{-i}(\omega) = x_{-i} \rrbracket)) \leq \log_e(\mu(\llbracket \mathfrak{M}_1 \circ f(X) | X_{-i}(\omega) = x_{-i} \rrbracket)) + \log_e(\mu(\llbracket \mathfrak{M}_2 \circ f(X) | X_{-i}(\omega) = x_{-i} \rrbracket))$. The proof for discrete uncertain variables follows the same approach.

APPENDIX G PROOF OF THEOREM 7

The proof follows from that $\llbracket g \circ \mathfrak{M} \circ f(X) | X_{-i}(\omega) = x_{-i} \rrbracket \leq \llbracket \mathfrak{M} \circ f(X) | X_{-i}(\omega) = x_{-i} \rrbracket$ for any x_{-i} .

APPENDIX H PROOF OF THEOREM 8

For any given $x_{-i} \in \llbracket X_{-i} \rrbracket$, due to continuity of f , we know that $f(\llbracket X_i \rrbracket \times \{x_{-i}\}) = f \circ \psi_{i, x_{-i}}(\llbracket X_i \rrbracket) \subseteq [y_{\min}, y_{\max}]$ is a connected set (because $\llbracket X_i \rrbracket$ is connected). Therefore, if $\mathfrak{M}$ is a q -level quantizer, $\llbracket Y | X_{-i}(\omega) = x_{-i} \rrbracket = \mathfrak{M} \circ f(\llbracket X_i \rrbracket \times \{x_{-i}\})$ can at most contain $q\mu(f(\llbracket X_i \rrbracket \times \{x_{-i}\}))/ (y_{\max} - y_{\min})$ points. Therefore, $|\llbracket Y | X_{-i}(\omega) = x_{-i} \rrbracket| \leq q\mathcal{S}_f / (y_{\max} - y_{\min})$. This concludes the proof.

⁹This follows from Assumption 1 (ii) and the fact that quantizers have connected preimage.

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Farokhi, F

Title:

Noiseless Privacy: Definition, Guarantees, and Applications

Date:

2021

Citation:

Farokhi, F. (2021). Noiseless Privacy: Definition, Guarantees, and Applications. IEEE Transactions on Big Data, PP (1), pp.1-1. <https://doi.org/10.1109/tbdata.2021.3104021>.

Persistent Link:

<http://hdl.handle.net/11343/281337>