An Information-Theoretic Characterization of Pufferfish Privacy

Theshani Nuradha
Cornell University
pt388@cornell.edu

Ziv Goldfeld
Cornell University
goldfeld@cornell.edu

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834520

***Abstract*—Pufferfish privacy (PP) is an appealing generalization of differential privacy (DP), that offers flexibility in specifying sensitive information and integrating domain knowledge into the privacy definition. Inspired by the illuminating equivalent formulation of DP in terms of mutual information proposed by Cuff and Yu [1], this work explores PP through the lens of information theory. We provide an equivalent information-theoretic formulation of PP as the conditional mutual information between the mechanism and the secret, given the public information. This formulation lends well for an information-theoretic analysis, and we use it to prove convexity, composability, and post-processing properties for PP mechanisms. We also leverage our formulation to derive noise levels for the Gaussian PP mechanisms. The obtained mechanisms are applicable under relaxed assumptions and provide improved noise levels in some regimes, compared to existing approaches,**

I. Introduction

The increased amount of personal data shared online, along with developments in data mining techniques pose serious privacy threats. Statistical privacy frameworks seek to address these threats in a principled manner with formal guarantees. Differential privacy (DP) [2] is perhaps the most popular statistical privacy framework, which enables answering aggregate queries about a database while keeping individual records private. However, DP only accounts for one type of private information (namely, individual records modeled by rows of the data matrix), and does not allow to encode domain knowledge into the framework. To address these limitations, a versatile generalization of DP, termed Pufferfish Privacy (PP), was proposed in [3]. PP allows to customize what information is regarded as private, and explicitly integrates distributional assumptions into the definition (see also [4]).

Furnishing connections between statistical privacy and information theory has gained increasing interest in the past decade [1], [5]–[9]. Such connections serve as a fertile ground for research, enabling to borrow tools and ideas from one discipline to make progress in the study of the other. In particular, [1] established an equivalent characterization of DP in terms of the conditional mutual information between the mechanism and any individual record, given the rest of the database. This reformulation lends well for an information-theoretic analysis of DP and poses it in terms of a common currency (namely, mutual information) through which privacy-utility tradeoffs may be examined [10]. It was also leveraged in [11] to study fundamental privacy-utility tradeoffs in linear regression problems. Inspired by the above, herein we target an information-theoretic characterization of PP, using which we shall explore various properties and novel mechanisms.

We consider a specialized PP framework, where private and public information is modeled as pairs of functions of the database that are coupled via a bipartite graph. This setup captures various privacy notions, from DP [2] to attribute privacy (AP) [12], as special cases. We provide an information-theoretic characterization of this PP framework via the conditional mutual information between the mechanism and the secret function given the public one. Specifically, the mutual information PP criteria is shown to be sandwiched between ϵ-PP and (ϵ, δ)-PP in terms of strength. The proof relies on representing PP constraints as bounds on appropriate divergences,[1] and comparing those to the Kullback-Leibler (KL) divergence (and thus mutual information) via monotonicity, Pinsker's inequality, and the minimax redundancy capacity theorem.

We then leverage the information-theoretic characterization to derive properties of PP mechanisms, including convexity, post-processing, and composability. Our composability results for mutual information PP offer greater flexibility than the counterparts for the classic PP framework [3]. Next, we study the Gaussian mechanism for achieving mutual information PP and derive sufficient conditions on the injected noise parameters. The derivation relies on controlling mutual information via maximum entropy arguments and the entropy power inequality. We obtain parameter bounds in terms of the conditional variance of the query, which differs from classic results that typically depend on the ℓ^1- or ℓ^2-sensitivity of the query (cf. e.g., [13], [14]). Variance-based bounds are particularly desirable under the PP framework as it encodes prior knowledge on the data distribution. Indeed, it may be the case that sensitivity explodes (e.g., for unbounded query functions) but variance is finite due to concentration properties of the distribution class.

II. Notations

Sets are denoted by calligraphic letters, e.g. $\mathcal{X}$. For $k, n \in \mathbb{N}$, we use $\mathcal{X}^{n\times k}$ for the database space of $n \times k$ matrices (columns correspond to different attributes while rows to different individuals). The (i,j)th entry of $x \in \mathcal{X}^{n\times k}$ is $x(i,j)$. The ith row and jth column of x are $x(i,\cdot)$ and $x(\cdot,j)$, respectively. The image of a function $g : \mathcal{X}^{n\times k} \to \mathbb{R}^d$ is denoted by $\mathrm{Im}(g)$. We use $\|\cdot\|_p$ for the ℓ^p, $p \geq 1$, norm in $\mathbb{R}^d$.

[1] ∞-Rényi divergence for ϵ-PP and total variation distance for $(0, \delta)$-PP.

We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space on which all random variables (RVs) are defined, with $\mathbb{E}$ designating expectation. RVs are denoted by upper case letters, e.g., X, with P_X representing the corresponding probability law. For $X \sim P_X$, we interchangeably use $\mathrm{supp}(X)$ and $\mathrm{supp}(P_X)$ for the support. Conventions for $n \times k$-dimensional random variables are the same as for deterministic elements. The space of all Borel probability measures on $\mathcal{S} \subseteq \mathbb{R}^d$ is denoted by $\mathcal{P}(\mathcal{S})$. We write $P \ll Q$ to denote that P is absolutely continuous with respect to (w.r.t.) Q. The n-fold product measure of $P \in \mathcal{P}(\mathcal{S})$ is $P^{\otimes n}$.

The mutual information between X and Y is denoted by $\mathsf{I}(X;Y)$, while $\mathsf{h}(X)$ is the differential entropy of X; conditional versions thereof given Z are denoted by $\mathsf{I}(X;Y|Z)$ and $\mathsf{h}(X|Z)$, respectively. The KL divergence between $P, Q \in \mathcal{P}(\mathcal{X})$ with $P \ll Q$ is $\mathsf{D}_{\mathsf{KL}}(P\|Q)$, while $\|P-Q\|_{\mathsf{TV}}$ designates the total variation (TV) distance. Both KL divergence and TV distance are jointly convex in (P, Q), which will be used subsequently. KL divergence and TV are related to one another via Pinsker's inequality [15], whereby $\|P-Q\|_{\mathsf{TV}} \leq \sqrt{0.5\mathsf{D}_{\mathsf{KL}}(P\|Q)}$. Also recall that $\mathsf{I}(X;Y)$ can be expressed in terms of KL divergence as $\mathsf{I}(X;Y) = \mathsf{D}_{\mathsf{KL}}(P_{XY}\|P_X \otimes P_Y)$, where P_X and P_Y are the respective marginals of X and Y.

III. Pufferfish Privacy and Mutual Information

This section establishes the mutual information based characterization of PP, where secrets and public information are modeled as arbitrary collections of functions of the database. Domain knowledge is integrated into the framework by restricting the set of data distributions.

A. Pufferfish Privacy

For a data space $\mathcal{X}^{n\times k}$, the Pufferfish framework [3] consists of three components: (i) a set of secrets $\mathcal{S}$, that contains measurable subsets of $\mathcal{X}^{n\times k}$; (ii) a set of secret pairs $\mathcal{Q} \subseteq \mathcal{S} \times \mathcal{S}$ that needs to be statistically indistinguishable in the (ϵ, δ) sense (see (1)); and (iii) a class of data distributions $\Theta \subseteq \mathcal{P}(\mathcal{X}^{n\times k})$, that captures prior beliefs or domain knowledge about potential adversaries. As formulated next, the goal of PP is to make all secret pairs in $\mathcal{Q}$ indistinguishable w.r.t. those prior beliefs $P_X \in \Theta$.

Definition 1 (Pufferfish privacy [3], [12])**.** *Fix $\epsilon, \delta > 0$. A randomized mechanism $M : \mathcal{X}^{n\times k} \to \mathcal{Y}$ is (ϵ, δ)-private in the pufferfish framework $(\mathcal{S}, \mathcal{Q}, \Theta)$ if for all $P_X \in \Theta$, secret pairs $(\mathcal{R}, \mathcal{T}) \in \mathcal{Q}$ with $P_X(\mathcal{R}), P_X(\mathcal{T}) > 0$, and measurable sets $\mathcal{A} \subseteq \mathcal{Y}$, we have*

$$\mathbb{P}\big(M(X) \in \mathcal{A}\big|\mathcal{R}\big) \leq e^{\epsilon}\, \mathbb{P}\big(M(X) \in \mathcal{A}\big|\mathcal{T}\big) + \delta. \qquad (1)$$

A randomized mechanism is described by a (regular) conditional probability distribution given the data, i.e., $P_{M|X}$.

In this work we focus on a special case of the general framework, where pairs $(\mathcal{R}, \mathcal{T}) \in \mathcal{Q}$ are decomposed into a private part (on which they should be indistinguishable) and a common one (interpreted as public information). In Remark 1 we demonstrate how the considered formulation reduces to popular privacy notions like DP [2] and AP [12]. Our formulation is constructed as follows:

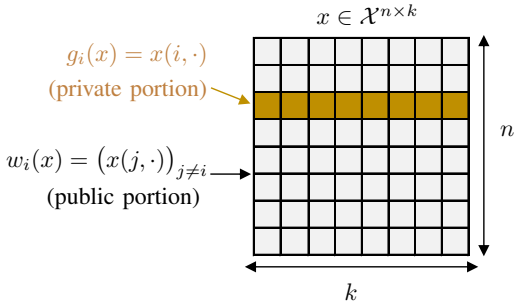

Fig. 1. Function pairs for DP: The ith row of $x \in \mathcal{X}^{n\times k}$ is the private portion, while the rest of the database is the corresponding public part.

1) Private/public functions: Let $\mathcal{G}$ and $\mathcal{W}$ be finite sets of sizes K and L, respectively, containing functions on $\mathcal{X}^{n\times k}$. For $g \in \mathcal{G}$, we interpret $g(X)$ as a private feature of the dataset $X \sim P_X \in \Theta$, while $w(X)$, $w \in \mathcal{W}$, represents publicly available information.

2) Function pairs: To encode which private-public function pairs constitute a secret (i.e., an element of $\mathcal{S}$) we use a bipartite graph. Namely, consider the graph $(\mathcal{G}, \mathcal{W}, \mathcal{E})$, where $\mathcal{E}$ is a given edge set between the two partitions $\mathcal{G}$ and $\mathcal{W}$. We write $g \sim w$ if $\{g, w\} \in \mathcal{E}$ for some $g \in \mathcal{G}$ and $w \in \mathcal{W}$. The operational meaning of an edge $g \sim w$ is that $g(X)$ must be concealed even if the adversary has access to $w(X)$ (e.g., for DP we take $g_i(x) = x(i,\cdot)$ as a specific row of the dataset and $w_i(x) = \big(x(j,\cdot)\big)_{j\neq i}$ as the rest of the data matrix, where $i = 1, \ldots, n$), and set $\mathcal{G} = \{g_i\}_{i=1}^n$, $\mathcal{W} = \{w_i\}_{i=1}^n$, and $\mathcal{E} = \big\{\{g_i, w_i\}\big\}_{i=1}^n$, as depicted in Figure 1.)

3) Secret event: Each secret event (namely, an element of $\mathcal{S}$) corresponds to a pair of specific values that a private-public function pair takes, i.e., for $\mathcal{G} \ni g \sim w \in \mathcal{W}$, $a \in \mathrm{Im}(g)$, and $c \in \mathrm{Im}(w)$, we define $\mathcal{A}_{g,w}(a,c) := \big\{g(X) = a,\, w(X) = c\big\}$.

4) Secret event pairs: Elements of $\mathcal{Q} \subseteq \mathcal{S} \times \mathcal{S}$ are pairs that share the same public information (i.e., the value for $w(X)$) but differ in their private portions (the value of $g(X)$).

We are now in position to define the considered PP framework.

Definition 2 (Specialized PP framework)**.** *Fix $\epsilon, \delta > 0$ and consider a bipartite graph $(\mathcal{G}, \mathcal{W}, \mathcal{E})$ with sets of functions $\mathcal{G}$ and $\mathcal{W}$ as above. A randomized mechanism $M : \mathcal{X}^{n\times k} \to \mathcal{Y}$ is (ϵ, δ)-private in the specialized pufferfish framework $(\mathcal{G}, \mathcal{W}, \mathcal{E}, \Theta)$ if it satisfies Definition 1 with*

$$\mathcal{S} := \big\{\mathcal{A}_{g,w}(a,c) : \mathcal{G} \ni g \sim w \in \mathcal{W},\, a \in \mathrm{Im}(g),\, c \in \mathrm{Im}(w)\big\}$$
$$\mathcal{Q} := \big\{\{\mathcal{A}_{g,w}(a,c), \mathcal{A}_{g,w}(b,c)\} : \mathcal{G} \ni g \sim w \in \mathcal{W},\, c \in \mathrm{Im}(w),\, a, b \in \mathrm{Im}(g),\, a \neq b\big\}$$

and as set of data distributions $\Theta \subseteq \mathcal{P}(\mathcal{X}^{n\times k})$.

Remark 1 (Reductions to DP and AP)**.** *The specialized PP framework presented above captures various special cases of practical importance, such as DP [2] and AP [12]. Specifically, DP corresponds to $\Theta = \mathcal{P}(\mathcal{X}^{n\times k})$, private functions $g_i(x) = x(i,\cdot)$, public functions $w_i(x) = \big(x(j,\cdot)\big)_{j\neq i}$, where*

$i = 1, \ldots, n$, and an edge set $\mathcal{E} = \big\{\{g_i, w_i\}\big\}_{i=1}^{n}$. Following the AP setup of [2], we take $g_j(x) = \tilde{g}_j\big(x(\cdot, j)\big)$, for $j = 1, \ldots, k$, where $\tilde{g}_j$ is the considered function of the jth column of the database. As AP includes no public information we further take $\mathcal{W} = \mathcal{E} = \emptyset$ and set Θ as the class of distributions of interest. Alternatively, one may consider a variant of AP with public information, in the spirit of DP, where $\mathcal{W}$ contains functions $w_j(x) = \big(x(\cdot, i)\big)_{i \neq j}$, where $j = 1, \ldots, k$, and the edge set is $\mathcal{E} = \big\{\{g_j, w_j\}\big\}_{j=1}^{k}$.

B. Information-Theoretic Characterization

We provide an information-theoretic characterization of the specialized PP framework, in the spirit of the DP characterization of [1]. To that end, we first define mutual information PP.

Definition 3 (ϵ-mutual-information PP)**.** *Let $(\mathcal{G}, \mathcal{W}, \mathcal{E})$ be a bipartite graph as in Definition 2 and $\Theta \subseteq \mathcal{P}(\mathcal{X}^{n \times k})$. A randomized mechanism $M : \mathcal{X}^{n \times k} \to \mathcal{Y}$ is ϵ-mutual-information PP, abbreviated ϵ-MI PP, in the framework $(\mathcal{G}, \mathcal{W}, \mathcal{E}, \Theta)$ if*

$$\sup_{\substack{P_X \in \Theta, \\ g \in \mathcal{G}, w \in \mathcal{W}: \\ g \sim w}} \mathsf{I}\big(g(X); M(X) | w(X)\big) \leq \epsilon.$$

Remark 2 (Reduction to ϵ-mutual-information DP, abbreviated ϵ-MI DP)**.** *ϵ-MI PP as defined above recovers information-theoretic formulation of DP proposed in [1] by taking $(\mathcal{G}, \mathcal{W}, \mathcal{E}, \Theta)$ as described in Remark 1.*

The following theorem states the equivalence between the specialized PP framework from Definition 2 and the ϵ-MI PP above. More precisely, our result shows that ϵ-MI PP sits right between $(\epsilon, 0)$-PP and (ϵ, δ)-PP in terms of its strength.

Theorem 1 (Equivalent formulations)**.** *consider the specialized (ϵ, δ)-PP framework $(\mathcal{G}, \mathcal{W}, \mathcal{E}, \Theta)$ from Definition 2. For any $\epsilon, \epsilon' > 0$, the following chain of implications holds:*

$$(\epsilon, 0)\text{-PP} \implies \epsilon\text{-MI PP} \implies (\epsilon', \sqrt{2\epsilon})\text{-PP}.$$

Furthermore, if $\big|\operatorname{supp}\big(M(X)\big)\big| < \infty$ or $\max_{g \in \mathcal{G}} |\mathrm{Im}(g)| < \infty$, then the inverse implication holds:

$$(\epsilon, \delta)\text{-PP} \implies \epsilon^\star\text{-MI PP}$$

with

$$\epsilon^\star = 2\mathsf{h}_b(\delta') + 2\delta' \log\left(\min\Big\{\big|\operatorname{supp}\big(M(X)\big)\big|, \max_{g \in \mathcal{G}} |\mathrm{Im}(g)| + 1\Big\}\right)$$

where h_b is the binary entropy function in nats and $\delta' \in [0, 1]$ with $\delta' = 1 - 2(1 - \delta)/(e^\epsilon + 1)$.

Theorem 1 is proven in Section V-A by reformulating $(\epsilon, 0)$-PP in terms of the ∞-Rényi divergence and using information-theoretic inequalities.

IV. Properties and Mechanisms

While the original PP definition is somewhat hard to manipulate, we next show that the information-theoretic formulation lands itself well for analysis, enabling the derivation of various properties and explicit noise-injection mechanisms.

A. Properties of Pufferfish Mechanisms

Modern guidelines for privacy frameworks [16] pose properties such as convexity and post-processing (also known as transformation invariance) as fundamental axioms. Composability is another important property that requires the joint distribution of the outputs of (possibly adaptively chosen) privacy mechanisms is in itself private. These properties are shown to hold for the general $(\epsilon, 0)$-PP framework in [3]. The next theorem states the ϵ-MI PP possess all these properties.

Theorem 2 (Properties of ϵ-MI PP)**.** *The following hold:*

1) Convexity: Let $\epsilon > 0$, $p \in [0, 1]$, and M_1, M_2 be ϵ-MI PP mechanisms. Then the mechanism that equals M_1 with probability (w.p.) p and M_2 w.p. $1 - p$ is also ϵ-MI PP.

2) Composability: Let $M_1, \ldots, M_k$ be sequentially and adaptively chosen $\epsilon_1 \ldots, \epsilon_k$-MI PP mechanisms, i.e., $\sup_{P_X, w \sim g} \mathsf{I}\big(g(X); M_i(X) \big| w(X), M_1, ..., M_{i-1}\big) \leq \epsilon_i$, for all $i = 1, \ldots, k$. Then M^k satisfies $\big(\sum_{i=1}^{k} \epsilon_i\big)$-MI PP.

3) Post-processing: If mechanism $M : \mathcal{X}^{n \times k} \to \mathcal{Y}$ satisfies ϵ-MI PP, then for any randomized function $A : \mathcal{Y} \to \mathcal{Z}$, the processed mechanism $A \circ M$ also satisfies ϵ-MI PP.

Theorem 2 is proven in Section V-B based on elementary properties of mutual information, such as its nullification under independence, the chain rule, and the data processing inequality. The simplicity of the argument highlights the virtue of the information-theoretic formulation of the PP framework.

Remark 3 (Non-adaptive composition)**.** *Non-adaptive composition refers to mechanisms that are conditionally independent given the database, i.e., $P_{M_1, \ldots, M_K | X} = \prod_{i=1}^{k} P_{M_i | X}$. For PP frameworks where pairs of secrets $(\mathcal{R}, \mathcal{T}) \in \mathcal{Q}$ corresponds to pairs of databases (in other words, $\mathcal{R}$ and $\mathcal{T}$ are singletons; see Definition 1), both ϵ-MI PP and regular ϵ-PP mechanism compose under this setting. The general non-adaptive setting, without assuming that secrets specify databases, was studied in [3], where it was shown that composability does not hold in general. [3] then identified a (rather restrictive) sufficient condition on the class of distributions Θ, termed universally composable (UC) distributions, under which non-adaptive composability holds for ϵ-PP. Similarly, non-adaptive composition of ϵ-MI PP also holds when all feasible distributions are UC. The UC condition is, however, unstable for ϵ-PP: if Θ contains all UC distribution, adding even a single non-UC distribution to this set will compromise the composability of the classic PP framework. ϵ-MI PP is more robust in that sense, as composability of ϵ-PP mechanisms is still ϵ-MI PP even after adding non-UC distributions to Θ, so long that all UC distributions are there (e.g., when $\Theta = \mathcal{P}(\mathcal{X}^{n \times k})$).*

B. Noise-Injection Pufferfish Mechanisms

Despite the practical importance of tractable privacy mechanisms, designing them for the general PP framework seems challenging. In [4], the Wasserstein PP mechanism was proposed, but it is computationally burdensome as it requires computing the ∞-Wasserstein distance between all pairs of conditional distributions of the published query result given

any pair of secrets from $\mathcal{Q}$. This sections shows that the information-theoretic formulation of PP allows bridging this gap, giving rise to (Gaussian and Laplace) noise-injection mechanisms whose noise level is specified in terms of elementary quantities. By virtue of Theorem 1, any of the following ϵ-MI PP mechanisms is also $(\epsilon', \sqrt{2\epsilon})$-PP in the classic sense (see Definition 2).

We first describe how a noise-injection mechanism operates. Let $f : \mathcal{X}^{n\times k} \to \mathbb{R}^d$ be the query so that our goal is to publish the value $f(X)$, for $X \sim P_X \in \Theta$, under the ϵ-MI PP framework from Definition 3. The noise-injection mechanism perturbs the published value as $M(X) = f(X) + Z$, where (in this paper) Z follows a Gaussian or Laplace distribution whose parameters are chosen so that ϵ-MI PP holds.

1) Gaussian mechanism: We characterize parameter values for the Gaussian mechanism that are sufficient for ϵ-MI PP.

Theorem 3 (Gaussian mechanism)**.** *Fix $\epsilon > 0$ and a specialized PP framework $(\mathcal{G}, \mathcal{W}, \mathcal{E}, \Theta)$. Let $f : \mathcal{X}^{n\times k} \to \mathbb{R}^d$ and consider the Gaussian mechanism $M_{\mathsf{G}}(X) := f(X) + Z_{\mathsf{G}}$, where $Z_{\mathsf{G}} \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ is a d-dimensional isotropic Gaussian of parameter $\sigma > 0$. If*

$$\sigma^2 \geq \sup_{P_X\in\Theta,\, w\in\mathcal{W}^\star} \frac{\sum_{j=1}^d \mathbb{E}\left[\mathrm{Var}\big(f_j(X)\big|w(X)\big)\right]}{d(e^{\frac{2\epsilon}{d}}-1)},$$

where $f_j(X)$ is the jth entry of $f(X) = \big(f_1(X), \ldots, f_d(X)\big)$ and $\mathcal{W}^\star = \{w \in \mathcal{W} : \exists\, g \in \mathcal{G},\, g \sim w\}$, then M_{G} is ϵ- MI PP.

The derivation of Theorem 3 is presented in Section V-C and relies, among other things, on the fact that the Gaussian distribution maximizes differential entropy subject to a second moment constraint. Theorem 3 can be specialized to ϵ-MI DP under the setup from Remark 1.

Corollary 1 (Gaussian mechanism for DP)**.** *Under the setup of Theorem 3, the Gaussian mechanism with*

$$\sigma^2 \geq \frac{\Delta_2^2(f)}{2d(e^{\frac{2\epsilon}{d}}-1)},$$

is ϵ-MI DP, where $\Delta_2(f) := \max_{x\sim x'} \|f(x) - f(x')\|_2$ is the ℓ^2-sensitivity and $x \sim x'$ denotes neighboring databases that differ by (at least one entry of) a single row, that is $x(i,\cdot) \neq x'(i,\cdot)$ for some $i = 1, \ldots, n$, and agree on all other rows.

Remark 4 (Classic Gaussian mechanisms)**.** *ϵ-MI DP implies regular $(\epsilon', \sqrt{2\epsilon})$-DP, for any $\epsilon' \geq 0$ (Theorem 1). For comparison, the classical Gaussian mechanism achieves $(\epsilon', \sqrt{2\epsilon})$-DP with $\sigma^2 \geq 2\log(1.25/\sqrt{2\epsilon})\Delta_2^2(f)/{\epsilon'}^2$ for $\epsilon' \leq 1$ [13]. It can be shown that our mechanism requires a lower noise level than the classic one whenever $\epsilon' < 2\big((e^{2\epsilon}-1)\log(1.25/\sqrt{2\epsilon})\big)^{1/2}$.*

Remark 5 (Domain knowledge for DP)**.** *A main advantage of our approach compared to sensitivity-based classic Gaussian mechanisms is that the bound in Theorem 3 depends on the variance of f and allows to incorporate domain knowledge. Consider the product Gaussian family*

$$\Theta_{\mathsf{G}}(m,s) = \Big\{ \prod_{i=1}^n \theta_i : \theta_i = \mathcal{N}(\mu_i, \sigma_i^2),\ |\mu_i| \leq m,\ \sigma_i^2 \leq s \Big\},$$

and let $f(X) = n^{-1}\sum_{i=1}^n X_i$ be the average of the database entries (the argument holds for any other linear query). Then, the noise derived from our mechanism is $(e^\epsilon - 1)^{-1}(s/n)^{-1/2}$, which is finite as the variance is bounded. On the other hand, $\Delta_2(f) = \infty$ here since X has unbounded support. Thus, the sensitivity-based mechanisms are vacuous for this case, while our bound provides feasible noise levels.

Remark 6 (Laplace mechanism)**.** *Under the setup of Theorem 3, it can be shown that the Laplace mechanism $M_{\mathsf{L}}(X) := f(X) + Z_{\mathsf{L}}$, where $Z_{\mathsf{L}} \sim \mathsf{Lap}(0,b)^{\otimes d}$ is a d-dimensional isotropic Laplace distribution with the scale parameter $b > 0$, is ϵ-MI PP if*

$$b \geq \sup_{P_X\in\Theta,\, w\in\mathcal{W}^\star} \frac{\sum_{j=1}^d \mathbb{E}\left[\sqrt{\mathrm{Var}\big(f_j(X)|w(X)\big)}\right]}{d(e^{\frac{\epsilon}{d}}-1)}.$$

This follows from the fact that the Laplace distribution maximizes differential entropy subject to an expected absolute deviation constraint in a similar fashion to the proof of Theorem 3.

2) Gaussian mechanism with dependence on $\mathcal{G}$: The noise levels derived in Theorem 3 depends on the secret function class $\mathcal{G}$ only through $\mathcal{W}^\star$. While this is sufficiently fine for DP, where there is a one-to-one correspondence between private and public functions, generally it may be desirable to capture the dependence on $\mathcal{G}$ explicitly. This is particularly relevant when there is no public information (e.g., the AP framework from [12], described in Remark 1) or if there is a single public function w corresponding to all private $g \in \mathcal{G}$. Noise levels derived from the following theorem takes this into account.

Theorem 4 (Gaussian mechanism with dependence on $\mathcal{G}$)**.** *Under the setup from Theorem 3 and assuming $\inf_{g\in\mathcal{G}, w\in\mathcal{W}:\, g\sim w} \mathsf{h}\big(f(X)\big|g(X), w(X)\big) > -\infty$, the Gaussian mechanism M_{G} achieves ϵ- MI PP, if*

$$\sigma^2 \geq \sup_{\substack{P_X\in\Theta,\\ g\in\mathcal{G}, w\in\mathcal{W}:\\ g\sim w}} \max\left\{\frac{A - d\,e^{2\epsilon/d}B}{d(e^{2\epsilon/d}-1)}, 0\right\},$$

where $A = \sum_{j=1}^d \mathbb{E}\big[\mathrm{Var}\big(f_j(X)\big|w(X)\big)\big]$ and $B = \frac{1}{2\pi}\exp\big(\frac{2}{d}\mathsf{h}\big(f(X)\big|g(X), w(X)\big) - 1\big)$.

In addition to maximum entropy arguments, the proof of Theorem 4 uses the entropy power inequality. We may replace the conditional entropy in B by any lower bound that may be easier to compute (cf. e.g., [17]), and ϵ-MI PP will still hold.

Remark 7 (Comparisons with Gaussian AP mechanism)**.** *The AP Gaussian mechanism was considered [12]. The setup assumes $\mathcal{W} = \emptyset$, and that for all $g \in \mathcal{G}$ and $P_X \in \Theta$, the query output $f(X)$ conditioned on $g(X)$ is Gaussian with a constant variance, i.e., $\mathrm{Var}\big(f(X)\big|g(X) = a\big) = \mathrm{Var}\big(f(X)\big|g(X) = b\big)$, for all $a, b \in \mathrm{Im}(g)$. Assuming $d = 1$ for simplicity, the noise level from Theorem 4 under this setup reduces to:*

$$\sup_{\substack{P_X\in\Theta,\\ g\in\mathcal{G}}} \max\left\{\frac{\mathrm{Var}(f(X)) - \ e^{2\epsilon}\mathrm{Var}\big(f(X)\big|g(X) = a\big)}{e^{2\epsilon}-1}, 0\right\}.$$

The noise level derived in [12], to achieve (ϵ,δ)-AP is

$$\sup_{\substack{P_X\in\Theta,\\ g\in\mathcal{G}}} \max\left\{\left(C\epsilon^{-1}\Delta_{\mathsf{AP}}(f)\right)^2 - \mathrm{Var}\big(f(X)\big|g(X)=a\big), 0\right\},$$

where $C=\sqrt{2\log(1.25/\delta)}$ and

$$\Delta_{\mathsf{AP}}(f)=\max_{a,b\in\mathrm{Im}(g)}\big|\mathbb{E}\big[f(X)\big|g(X)=a\big]-\mathbb{E}\big[f(X)\big|g(X)=b\big]\big|.$$

Under a multivariate extension of the product Gaussian family from Remark 5 (i.e., where each θ_i is a multivariate Gaussian) and for f and g linear functions of columns of the database (e.g., average of the column entries), $\Delta_{\mathrm{AP}}(f)$ diverges to infinity while the variance-based bound is finite and feasible.

Remark 8 (Free ϵ-MI PP regime)**.** *The bound in Theorem 4 suggests that if $A \le d\,e^{2\epsilon/d}B$ over the entire optimization domain, ϵ-MI PP holds without noise injection (i.e., $\sigma=0$). It can be shown that $A \ge dB$ for any $P_X\in\mathcal{P}(\mathcal{X})$ and functions f, g, and w. The free privacy regime therefore corresponds to cases where ϵ is large compared to $d/2$. Since large ϵ values are rarely of interest in practice, we conclude that a positive noise level is generally needed for ϵ-MI PP. For fixed ϵ and d, the above condition is related to how correlated the query and the secret are, given the public information. For instance, if $d=1$ and $f(X)$, $g(X)$, and $w(X)$ are jointly Gaussian, we have $A \le d\,e^{2\epsilon/d}B$ if the conditional correlation coefficient between $f(X)$ and $g(X)$ given $\{w(X)=c\}$ satisfies $\rho\big(f(X),g(X)\big|w(X)=c\big) \le \sqrt{(e^{2\epsilon}-1)e^{-2\epsilon}}$. Accordingly, weak correlation may lead to free privacy since the query leaks little information about the secret to begin with.*

V. Proofs

A. Proof of Theorem 1

For the first implication, note that ϵ-PP implies that

$$\sup_{\mathcal{A}}\ \log\left(\frac{\mathbb{P}\big(M(X)\in\mathcal{A}\big|\mathcal{R}\big)}{\mathbb{P}\big(M(X)\in\mathcal{A}\big|\mathcal{T}\big)}\right)\le\epsilon,\quad \forall(\mathcal{R},\mathcal{T})\in\mathcal{Q}.$$

The left-hand side above is the infinite order Rényi divergence. By monotonicity of Rényi divergences w.r.t. their order [18], we have $\mathsf{D}_{\mathsf{KL}}\left(P_{M(X)|\mathcal{R}}\|P_{M(X)|\mathcal{T}}\right)\le\epsilon$. Then,

$$\begin{aligned}&\mathsf{I}\big(g(X);M(X)|w(X)\big)\\ &\qquad\le\mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}\left(P_{M(X)|g(X),w(X)}\|P_{M(X)|g(X)',w(X)}\right)\right]\end{aligned}$$

where the inequality uses convexity of KL divergence, with $g(X)'$ as an i.i.d. copy of $g(X)$. Recalling that under the specialized PP framework secret pairs are $\big(A_{g,w}(a,c),A_{g,w}(b,c)\big)$, with $A_{g,w}(a,c)=\big\{g(X)=a,\,w(X)=c\big\}$, ϵ-MI PP follows by the KL divergence bound.

For the second implication, by the minimax redundancy capacity theorem [19], $\mathsf{I}\big(g(X);M(X)|w(X)=c\big)$ is rewritten as $\inf_Q\max_a\mathsf{D}_{\mathsf{KL}}\big(P_{M(X)|g(X)=a,w(X)=c}\big\|Q\big)$. Letting $Q^\star$ achieve the infimum above, since M is ϵ-MI PP by assumption, we have $\mathsf{D}_{\mathsf{KL}}\big(P_{M(X)|g(X)=a,w(X)=c}\big\|Q^\star\big)\le\epsilon$, for all $a\in\mathrm{Im}(g)$. Applying Pinsker's inequality together with the triangle inequality, we obtain

$$\big\|P_{M(X)|g(X)=a,w(X)=c}-P_{M(X)|g(X)=b,w(X)=c}\big\|_{\mathsf{TV}}\le\sqrt{2\epsilon},$$

which implies that M is $(0,\sqrt{2\epsilon})$-PP and hence $(\epsilon',\sqrt{2\epsilon})$-PP.

For the last implication, adapting Property 3 in [1] from DP to PP, we have that (ϵ,δ)-PP implies $(0,\delta')$-PP, with $\delta'=1-2(1-\delta)/(e^\epsilon+1)$. With this reduction, we follow the argument from the proof of Lemma 3 in [1] to show that if $\big|\operatorname{supp}\big(M(X)\big)\big|<\infty$ or $\max_{g\in\mathcal{G}}|\mathrm{Im}(g)|<\infty$, then $(0,\delta)$-PP implies $\epsilon^\star$-MI PP with $\epsilon^\star$ as stated in Theorem 1.

B. Proof of Theorem 2

For (1), let $B\sim\ \mathrm{Ber}(p_1)$, set $p_2=1-p_1$, and define $Q=2-B$. By independence of Q from $\big(X,M_1(X),M_2(X)\big)$ and since M_1 and M_2 satisfy ϵ-MI PP, we have

$$\begin{aligned}\mathsf{I}\big(g(X);M_Q(X)\big|w(X)\big)&\le\mathsf{I}\big(g(X);M_Q(X)\big|w(X),Q\big)\\ &\le\sum_{i=1}^{2}p_i\mathsf{I}\big(g(X);M_i(X)\big|w(X)\big)\le\epsilon.\end{aligned}$$

Claims (2) and (3) are direct consequences of the mutual information chain rule and the data processing inequality.

C. Proof of Theorem 3

$$\begin{aligned}&\mathsf{h}(f(X)+Z_{\mathsf{G}}|w(X))\\ &\overset{(a)}{\le}\int\frac{1}{2}\log\left((2\pi e)^d\ \left|\Sigma_{f(X)|w(X)=c}+\sigma^2\mathrm{I}_d\right|\right)\mathrm{d}P_{w(X)}(c)\\ &\overset{(b)}{\le}\int\frac{1}{2}\log\left((2\pi e)^d\prod_{j=1}^{d}\left(a_j(c)+\sigma^2\right)\right)\mathrm{d}P_{w(X)}(c)\\ &\overset{(c)}{\le}\frac{d}{2}\log\left(2\pi e\left(\frac{1}{d}\sum_{j=1}^{d}\mathbb{E}\left[\mathrm{Var}\big(f_j(X)\big|w(X)\big)\right]+\sigma^2\right)\right),\end{aligned}\tag{2}$$

where (a) follows from the Gaussian distribution maximizing differential entropy subject to a variance constant, with $|K|$ denoting the determinant of K; (b) denotes $a_j(c)=\mathrm{Var}\big(f_j(X)|w(X)=c\big)$ and uses $|K|\le\prod_{j=1}^{d}K(j,j)$, which applies to any positive semidefinite matrix; and (c) from concavity of $x\mapsto\log x$.

Combining (2) with $\mathsf{h}\big(f(X)+Z_{\mathsf{G}}\big|g(X),w(X)\big)\ge\mathsf{h}(Z_{\mathsf{G}})$, upper bounds $\mathsf{I}\big(g(X);M_{\mathsf{G}}(X)\big|w(X)\big)$ which is further bounded by ϵ. Solving for σ^2 concludes the proof.

D. Proof of Theorem 4

Similar to the proof of Theorem 3, but using the bounds

$$\begin{aligned}\mathsf{h}\big(f(X)+Z_{\mathsf{G}}\big|w(X)\big)&\le 0.5d\log\left(2\pi e\left(\frac{A}{d}+\sigma^2\right)\right)\\ \mathsf{h}\big(f(X)+Z_{\mathsf{G}}\big|g(X),w(X)\big)&\ge 0.5d\log\big(2\pi e(B+\sigma^2)\big),\end{aligned}$$

where the latter follows from entropy power inequality, with A and B specified in the theorem statement.

REFERENCES

[1] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," in *Proc. of ACM SIGSAC Conference on Computer and Communications Security*, Oct. 2016.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. of Conference on Theory of Cryptography, TCC*, pp. 265–284, 2006.

[3] D. Kifer and A. Machanavajjhala, "Pufferfish: A framework for mathematical privacy definitions," *ACM Transactions on Database Systems*, vol. 39, no. 1, 2014.

[4] S. Song, Y. Wang, and K. Chaudhuri, "Pufferfish privacy mechanisms for correlated data," in *Proc. of ACM SIGMOD*, pp. 1291–1306, 2017.

[5] G. Barthe and B. Kopf, "Information-theoretic bounds for differentially private mechanisms," in *IEEE Computer Security Foundations Symposium*, pp. 191–204, 2011.

[6] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *IEEE 54th Annual Symposium on Foundations of Computer Science(FOCS)*, pp. 429–438, 2013.

[7] S. Asoodeh, F. Alajaji, and T. Linder, "On maximal correlation, mutual information and data privacy," in *IEEE 14th Canadian Workshop on Information Theory (CWIT)*, pp. 27–31, 2015.

[8] 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.

[9] I. Issa and A. B. Wagner, "Operational definitions for some common information leakage metrics," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 769–773, 2017.

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

[11] M. Showkatbakhsh, C. Karakus, and S. Diggavi, "Privacy-utility tradeoff of linear regression under random projections and additive noise," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 186–190, 2018.

[12] W. Zhang, O. Ohrimenko, and R. Cummings, "Attribute privacy: Framework and mechanisms," *arXiv preprint arXiv:2009.04013*, 2020.

[13] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science (FnT-TCS)*, vol. 9, no. 3-4, pp. 211–407, 2014.

[14] N. Holohan, D. Leith, and O. Mason, "Rényi divergence and Kullback-Leibler divergence," *Information Sciences*, vol. 305, p. 256–268, 2015.

[15] M. D. Reid and R. C. Williamson, "Generalised Pinsker inequalities," *arXiv preprint arXiv:0906.1244*, 2009.

[16] D. Kifer and B.-R. Lin, "An axiomatic view of statistical privacy and utility," *Journal of Privacy and Confidentiality*, vol. 4, no. 1, 2012.

[17] A. Marsiglietti and V. Kostina, "A lower bound on the differential entropy of log-concave random vectors with applications," *Entropy*, vol. 20, no. 3, p. 185, 2018.

[18] T. van Erven and P. Harremos, "Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[19] T. Cover and J. A. Thomas, *Elements of information theory*. Hoboken, NJ: Wiley-Interscience, 2 edition, 2006.