Bayes Security: A Not So Average Metric

Konstantinos Chatzikokolakis, Giovanni Cherubin, Catuscia Palamidessi, Carmela Troncoso

Bayes Security: A Not So Average Metric

Konstantinos Chatzikokolakis†
University of Athens
kostasc@di.uoa.gr

Giovanni Cherubin†
Microsoft Research
gcherubin@microsoft.com

Catuscia Palamidessi†
INRIA, École Polytechnique
catuscia@lix.polytechnique.fr

Carmela Troncoso†
EPFL
carmela.troncoso@epfl.ch

***Abstract*—Security system designers favor worst-case security metrics, such as those derived from differential privacy (DP), due to the strong guarantees they provide. On the downside, these guarantees result in a high penalty on the system's performance. In this paper, we study Bayes security, a security metric inspired by the cryptographic advantage. Similarly to DP, Bayes security i) is independent of an adversary's prior knowledge, ii) it captures the worst-case scenario for the two most vulnerable secrets (e.g., data records); and iii) it is easy to compose, facilitating security analyses. Additionally, Bayes security iv) can be consistently estimated in a black-box manner, contrary to DP, which is useful when a formal analysis is not feasible; and v) provides a better utility-security trade-off in high-security regimes because it quantifies the risk for a specific threat model as opposed to threat-agnostic metrics such as DP.**

We formulate a theory around Bayes security, and we provide a thorough comparison with respect to well-known metrics, identifying the scenarios where Bayes Security is advantageous for designers.

***Index Terms*—Leakage, Quantitative Information Flow, Bayes risk, Bayes security metric, Local differential privacy**

I. Introduction

Quantifying the level of protection given by security and privacy-preserving mechanisms is a fundamental process in secure system engineering. To perform a quantitative analysis, one needs to define appropriate metrics that capture the adversary's gain, and, ultimately, what are the risks for the system's users.

A common way of evaluating threats in security and privacy applications is to quantify the probability that an adversary guesses some secret information; this metric is referred to as the *success rate* or *accuracy* of an attacker. For example, membership inference attacks against machine learning (ML) models [31], where the attacker aims at guessing if a data record was used for training the model, have been for long evaluated w.r.t. the attacker's accuracy.

Average-case metrics. The success rate (or accuracy) has a very clear interpretation: it measures the probability that an adversary succeeds in the attack. An important special case, the *Bayes vulnerability* (e.g., [32]), is the accuracy of the (Bayes) optimal adversary, who has maximal information about the underlying uncertainty. Both Bayes vulnerability and accuracy rely on the the *prior* probability of the secret information that the attacker is trying to guess; unfortunately, this can result in misleading conclusions about an attack's strength [32]. In the membership inference example, if the prior probability that a data record is low (say, 0.1), a strawman attack that always guesses "non-member" will achieve 90% accuracy; yet, this is a rather weak attack. This shows that the accuracy metric does not characterize well the risk of this attack.

An alternative criterion, used to evaluate cryptographic primitives, is the *advantage* (e.g., [5]). Advantage defines the prior probability over the secrets to be uniform, by construction, and it relates this prior probability to the probability that the adversary succeeds after having access to the model (accuracy). *Intuitively*, this metric disregards the contribution of the prior, and it quantifies the information leakage of the algorithm itself; however, to the best of our knowledge, no known result shows that this metric is prior-independent.

Both cryptographic advantage and Bayes vulnerability are *threat-specific*, i.e., they are connected to the threat model under which security is quantified. This gives a precise interpretation of what attacks they protect against. However, they are rarely used to study complex real-world systems, such as ML training algorithms. The main reason is that, due to complexity, one often needs to evaluate the security of individual parts of the algorithm and then *compose them*; this is not known to be possible with these metrics.

Worst-case metrics. At the other end of the spectrum, Differential Privacy (DP) has become the golden standard in privacy analysis [14]. In DP, a parameter ε bounds the probability that an algorithm's output *leaks any information*. There are several reasons why DP is generally preferred over other metrics: 1) DP is easy to compose analytically; e.g., if two algorithms are resp. ε_1- and ε_2-DP, their cascade composition is $(\varepsilon_1 + \varepsilon_2)$-DP. 2) DP is *prior-independent*: it measures the risk of releasing a secret via the algorithm, independently of the secret's prior probability; 3) DP protects against virtually any threat model: its guarantees hold whether the adversary wishes to learn an entire data record or just one bit of information; we refer to this property as being *threat-agnostic*. 4) DP considers the *worst-case* scenario over the outputs, ensuring robustness against any threat: it bounds the best gain an adversary can have, even if their maximum gain is achieved with negligible probability.

DP, however, also comes with disadvantages. First, DP is often too strict of a requirement: in many security settings, such as traffic analysis, side channel protection, and privacy-preserving ML (PPML), DP mechanisms that provide high protection levels incur severe utility loss. This mostly comes from the fact that DP is threat-agnostic. Second, it is theoretically impossible to estimate empirically ε-DP in a consistent

† Equal contribution.

manner (e.g., [10]) for black-box mechanisms; for example, an empirical DP estimate would fail to properly assess a mechanism that violates ε-DP with negligible probability.

Bayes security. In this paper, we study the multiplicative risk leakage (β), a metric that generalizes the cryptographic advantage defined for a Bayes-optimal adversary [9]. Specifically, we focus on the minimizer of β across all prior probability distributions, which we call the *Bayes security metric* (β^*). We identify several properties about β^* that make it suitable for studying security and privacy threats of complex algorithms: 1) similarly to the Bayes vulnerability and the advantage, Bayes security is threat-specific; 2) differently from them, it quantifies the risk for the two most vulnerable secrets (*worst-case*). 3) Similarly to DP, the *composition* of Bayes secure algorithms is easy to study; 4) the formal analysis of complex algorithms via Bayes security is further aided by its direct relation with the total variation distance between the output distributions of the algorithm; 5) when a formal analysis is not possible (e.g., one needs to study a black-box system), there are consistent methods for estimating Bayes security (Section VII).

Finally, because of its construction, Bayes security captures the *average* (i.e., expected) risk for the *worst-case* pair of inputs (e.g., data records, users). In this sense, it can be regarded as a middle way between average and worst-case security metrics; we argue that it gains benefits from both.

We summarize our contributions as follows:

- ✓ We study multiplicative risk leakage [9], β, a generalization of the advantage. We show that it reaches its least secure setting, β^*, when assigning a uniform prior to the two most vulnerable secrets. We call this minimum Bayes security, which captures the expected risk for the two secrets that are the easiest to distinguish for an optimal adversary (Section III).
- ✓ We study compositionality rules for algorithms that satisfy Bayes security (Section IV).
- ✓ We study the relation between Bayes security and mainstream security and privacy notions. We provide a game-based interpretation of Bayes security, which is equivalent to IND-CPA, and one for local DP. We derive bounds w.r.t. DP and local DP, which improve on the bound by Yeom et al. [39]. Our analysis shows Bayes security is in between worst-case metrics (e.g., DP) and average ones (Section V). This enables system designers to explore new security-utility trade-offs (Section VIII).
- ✓ We derive the Bayes security of three mainstream privacy mechanisms: Randomized Response, and the Laplace and Gaussian mechanisms (Section VI), and discuss suitable applications (Section VIII).
- ✓ Finally, we provide efficient means to compute β^* in white- and black-box scenarios (Section VII).

Due to space constraints, proofs are given in the appendix.

II. Preliminaries

We consider a *system* $(\pi, \mathcal{C})$, where a *channel* or *mechanism* $\mathcal{C}$ protects secrets $s \in \mathbb{S}$. Let $\mathcal{D}(S)$ be the set of probability distributions over a set S. Secrets are selected as inputs to the channel according to a *prior* probability distribution $\pi \in \mathcal{D}(\mathbb{S})$; we write $\pi_s \stackrel{\text{def}}{=} P(s)$. The channel is a matrix defining the posterior probability of observing an output $o \in \mathbb{O}$ given an input $s \in \mathbb{S}$: $\mathcal{C}_{s,o} \stackrel{\text{def}}{=} P(o \mid s)$. We denote by $\mathcal{C}_s \in \mathcal{D}(\mathbb{O})$ the s-th row of $\mathcal{C}$ (which is a distribution over $\mathbb{O}$), and by $\mathcal{C}_{\mathbb{S}}$ the set of all rows of $\mathcal{C}$. Table I summarizes our notation.

Adversarial Goal. We consider a passive adversary $\mathcal{A}$ who, given an output o, aims at inferring which secret s was input to the mechanism. We model this adversary with the following indistinguishability game, which we call IND-BAY:

IND-BAY$^{\mathcal{A}}_{\mathcal{C}}$

1 : $\mathcal{A} \leftarrow \mathcal{C}, \pi$
2 : $s \xleftarrow{\pi} \mathbb{S}$
3 : $o \xleftarrow{P(o|s)} \mathbb{O}$
4 : $s' \leftarrow \mathcal{A}(o)$
5 : **return** $s = s'$

We consider an optimal adversary that has perfect knowledge of the channel $\mathcal{C}$ and of the prior distribution over the secret inputs π (line 1). A challenger samples a secret s according to the prior π (line 2), and inputs it to the channel $\mathcal{C}$ to obtain an observable output o (line 3). For simplicity, the game considers an individual observation, but we note that sequences of observations (e.g., representing multiple uses of the same channel to hide one secret, or simultaneous use of two channels with the same secret) can be accounted for by redefining o to be a vector. Upon observing the output o. the adversary produces a prediction s' (line 4). The adversary wins if they guess the secret correctly: $s = s'$ (line 5). We evaluate the adversary $\mathcal{A}$ with respect to their expected prediction error according to the 0-1 loss function: $R^{\mathcal{A}} \stackrel{\text{def}}{=} P(s \neq \mathcal{A}(o)) = P(s \neq s')$. Extensions to further loss functions are possible, but out of the scope of this paper.

This formulation is different from typical cryptographic games because of the following reasons. First, we assume an optimal adversary: instead of providing them with knowledge of the cryptographic algorithm except for the key, and let them query the primitive to learn its statistical behavior, we assume that the adversary has perfect knowledge of the probabilistic behavior of the channel. Second, we compute the advantage with respect to the adversary's error, while cryptographic games compute the adversary's probability of success. Third, this game captures an *eavesdropping* adversary that *cannot* influence the secret used by the challenger to produce the observable output. This is considered to be a weak adversary in cryptography, where typically the adversary is allowed to provide inputs to the algorithm under attack. However, it corresponds to many security and privacy problems where the adversary cannot influence the secret and only observes channel outputs: website fingerprinting [21], [37], privacy-preserving distribution estimation [18], [29], [30], side channel attacks [25], [26], [33], or pseudorandom number generation.

Adversarial Models. In this paper, we consider the *Bayes*

adversary, an idealized adversary who knows both prior π and channel matrix $\mathcal{C}$, and guesses according to the Bayes rule:

$$s' = \mathcal{A}^*(o, \pi, \mathcal{C}) = \arg\max_{s \in \mathbb{S}} P(s|o) = \arg\max_{s \in \mathbb{S}} \mathcal{C}_{s,o}\pi_s \,.$$

The expected error of the Bayes adversary (*Bayes risk*) is:

$$R^{\mathcal{A}^*}(\pi, \mathcal{C}) = R^*(\pi, \mathcal{C}) = 1 - \sum_{o \in \mathbb{O}} \max_{s \in \mathbb{S}} \mathcal{C}_{s,o}\pi_s \,.$$

When this adversary is confronted with a perfect channel, whose outputs leak nothing about the inputs, their best strategy is to guess according to priors: $s' = \arg\max_{s \in \mathbb{S}} \pi_s$. The expected error of this strategy is the *random guessing error*: $G(\pi) = 1 - \max_{s \in \mathbb{S}} \pi_s$. Naturally, $G(\pi) \geq R^*(\pi, \mathcal{C})$.

Multiplicative Bayes risk leakage. We study the properties of a metric, β, defined as [9]:

$$\beta(\pi, \mathcal{C}) \stackrel{\text{def}}{=} \frac{R^*(\pi, \mathcal{C})}{G(\pi)} \,,$$

where it is assumed that $G(\pi) > 0$; we refer to β as the *multiplicative Bayes risk leakage*. Inspired from the cryptographic advantage (Section V-A), β captures how much better than random guessing an adversary can do. It takes values in $[0, 1]$; $\beta = 1$ when the system is perfectly secure (i.e., it exhibits no leakage), and $\beta = 0$ when the adversary always guesses the secret correctly. In the next section, we define the Bayes security metric $\beta^*(\mathcal{C})$ to be the minimizer (i.e., least secure configuration) of $\beta(\pi, \mathcal{C})$ for any prior π. We then study its properties, which we argue make it suitable for analyzing the security of complex real-world algorithms.

We note that β is closely related to the *multiplicative Bayes vulnerability leakage* [6], which is defined as: $\mathcal{L}^\times(\pi, \mathcal{C}) \stackrel{\text{def}}{=} \frac{1-R^*(\pi,\mathcal{C})}{1-G(\pi)}$. Differently from β, $\mathcal{L}^\times$ is defined for the adversary's probability of *success* (i.e., Bayes *vulnerability*) rather than failure, and it takes values in $[0, n]$. Despite their similar definition, these two metrics behave very differently (Section V). An important result for $\mathcal{L}^\times$ is that it takes its worse value (i.e., least secure configuration) on a uniform prior over the secrets [6]. Therefore, even when the real priors are unknown, a security analyst can easily compute a bound on the security of the system. In the next section, we derive the counterpart result for the β: it reaches its least secure configuration when setting a uniform prior on the two most vulnerable secrets, and a 0 prior probability elsewhere; the proof is substantially more involved than the one for $\mathcal{L}^\times$. We further discuss the relation between the Bayes Security metric and multiplicative leakage in Section V-E.

III. The Bayes security metric

Leakage notions based on the Bayes risk generally depend on the prior distribution over the secrets. This makes them unsuitable for measuring security in real-world applications where the true priors are unknown, e.g., traffic analysis [17], [38] or membership inference attacks [31], and result in an overestimation of a mechanism's security if the real prior implies more leakage than the prior considered in the analysis.

TABLE I: Notation

Symbol	Description
$\mathbb{S} = \{1, ..., n\}$	The secret space.
$\mathbb{O} = \{1, ..., m\}$	The output space.
$\mathcal{C}_{s,o}$ (abbr. $\mathcal{C}$)	A channel matrix, where $\mathcal{C}_{s,o} = P(o \mid s)$ for $s \in \mathbb{S}$, $o \in \mathbb{O}$.
$\mathcal{C}_s$	The s-th row of a channel matrix. It corresponds to the probability distribution $P(o \mid s), \forall o \in \mathbb{O}$.
$\pi \in [0, 1]^n$	A vector of prior probabilities over the secret space. The i-th entry of the vector is π_i.
π_{ij}	A prior vector with exactly 2 non-zero entries, in position i and j, with $i \neq j$.
$\upsilon = (1/n, ...1/n)$	Uniform priors for a secret space of size $\lvert\mathbb{S}\rvert = n$.
$R^*(\pi, \mathcal{C})$ (abbr. R^*)	The Bayes risk of a channel.
$G(\pi)$ (abbr. G)	The random guessing error (error when only priors' knowledge is available).
$\beta(\pi, \mathcal{C})$	Bayes security of a channel.
$\beta^*(\mathcal{C})$ (abbr. β^*)	Min Bayes security of a channel.

Given the similarity between multiplicative *risk* leakage and multiplicative *vulnerability* leakage, one could expect that the uniform prior also represents the worst case for the latter [6]. Unfortunately, this is not the case: Theorem 7 (Appendix A) shows that, for secret spaces $|\mathbb{S}| > 2$, there exists a prior π for which β is smaller than the one achieved for a uniform prior.

Prior-independence for β. In this section, we show that the multiplicative Bayes risk leakage, $\beta(\pi, \mathcal{C})$, for a channel $\mathcal{C}$, is minimized when the prior π assigns equal weight to the two secrets that are maximally distant (according to posterior distribution), and 0 to all other secrets. We refer to this minimizer, representing the highest risk for the channel w.r.t. adversary's prior knowledge, as the *Bayes security metric*; we denote it with $\beta^*(\mathcal{C})$ (omitting the argument if no confusion arises). This result makes the Bayes security metric prior-independent: for any prior knowledge the attacker may have in practice, β^* bounds their success.

For simplicity, we present our result in the *one-try* attack scenario, as formalized by the IND-BAY game: the adversary observes just *one* output of the system before guessing the secret input. In Section IV we extend this result to cases where the adversary can collect more observations.

Theorem 1. *Consider a channel $\mathcal{C}$ on a secret space with $|\mathbb{S}| \geq 2$. There exists a prior vector $\pi^* \in \mathcal{D}(\mathbb{S})$ of the form*

$$\pi^* = \{0, ..., 0, 1/2, 0, ..., 0, 1/2, 0, ..., 0\}$$

such that

$$\beta^*(\mathcal{C}) = \beta(\pi^*, \mathcal{C}) = \min_{\pi \in \mathcal{D}(\mathbb{S})} \beta(\pi, \mathcal{C}) \,.$$

In the following, we provide an intuition of the concepts involved with this proof.

We denote with $\mathcal{U}^{(k)} \subset \mathcal{D}(\mathbb{S})$, for $k = 1, ..., |\mathbb{S}|$, the set of distributions whose support has cardinality k, and with a uniform distribution over its non-zero components:

$$\mathcal{U}^{(k)} \stackrel{\text{def}}{=} \left\{ u \in \mathcal{D}(\mathbb{S}) \mid u_s \in \{0, \frac{1}{k}\} \text{ for all } s \in \mathbb{S} \right\} .$$

For example, if $n = 3$, then: $\mathcal{U}^{(1)} = \{(1, 0, 0), ..., (0, 0, 1)\}$, $\mathcal{U}^{(2)} = \{(1/2, 1/2, 0), (1/2, 0, 1/2), (0, 1/2, 1/2)\}$, and $\mathcal{U}^{(3)} = \{(1/3, 1/3, 1/3)\}$.

For a fixed channel $\mathcal{C}$, the proof of Theorem 1 is based on demonstrating the following two steps:

1) the function $\beta(\pi, \mathcal{C}) = {}^{R^*(\pi,\mathcal{C})}/_{G(\pi)}$ has its minimum in the set $\mathcal{U} = \mathcal{U}^{(1)} \cup ... \cup \mathcal{U}^{(|\mathbb{S}|)}$. The elements of $\mathcal{U}$ are known in the literature as the *corner points* of $G(\pi)$;
2) the minimizing prior π^* of $\beta(\pi, \mathcal{C})$ has cardinality 2; that is, $\pi^* \in \mathcal{U}^{(2)}$.

The proof for the first step comes from the observation that the function β is the ratio between a concave function, R^*, and a function G that is convexly generated by $\mathcal{U}$. Lemma 2 (Appendix B) shows that the minima of this ratio exist, and that they must come from the set of corner points of G (i.e., the set $\mathcal{U}$). This determines the form of the minimizing priors.

For the second step, under the constraints given by Lemma 2, the Bayes risk $R^*(\pi, \mathcal{C})$ decreases quicker than $G(\pi)$ as we increase the number of 0'es in $\pi \in \mathcal{U}$. By excluding the solution $\pi^* \in \mathcal{U}^{(1)}$, which would force the denominator $G = 0$, it follows that the minimizer of $\beta(\pi, \mathcal{C})$, π^*, has exactly 2 non-zero elements; that is, $\pi^* \in \mathcal{U}^{(2)}$.

Discussion. Theorem 1 has several consequences. First, the fact that $\beta^*(\mathcal{C})$ does not depend on a prior means that it captures the actual leakage of the channel, excluding any prior knowledge that the adversary may have. Conveniently, after Bayes security is computed for the channel, one can recover the success rate of the Bayes optimal adversary for desired levels of attacker's knowledge (Section V-E). Second, the fact that Bayes security represents the *risk for the two leakiest secrets* means that:

- If the two leakiest secrets can be determined a priori, this makes the security analysis straightforward (Section VI);
- If the two leakiest secrets cannot be determined a priori, one only needs $O(n^2)$ computations (instead of $O(n!)$) to recover them.

Finally, Theorem 1 suggests that Bayes security can be interpreted as a middle way between worst-case and average-case security metrics: it represents the expected (i.e., average) risk for the two most vulnerable (i.e., worst-case) secrets. We argue that this, paired with the fact that Bayes security is threat-specific, favors the interpretability of this metric. In the next sections, we prove properties about Bayes security which make it suitable for studying complex mechanisms.

Bayes security and total variation. We now introduce an important result for β^* which helps analyzing mechanisms in practice: Bayes security is the complement of the total variation of the two maximally distant rows of the channel:

Theorem 2. *For any channel $\mathcal{C}$, it holds that*

$$\beta^*(\mathcal{C}) \;=\; 1 - \frac{1}{2} \max_{a,b \in \mathbb{S}} \|\mathcal{C}_a - \mathcal{C}_b\|_1 \;=\; 1 - \max_{a,b \in \mathbb{S}} \mathrm{tv}(\mathcal{C}_a, \mathcal{C}_b) \;.$$

This result gives a clear interpretation of what β^* represents: it measures the maximal distance between the pairwise posterior distributions of the outputs w.r.t. the secret inputs (Figure 1). Further, thanks to this result: i) it is easy to analyze mechanisms both analytically (Section VI) and via estimation techniques (Section VII) by exploiting the plethora of results surrounding the total variation distance between distributions.

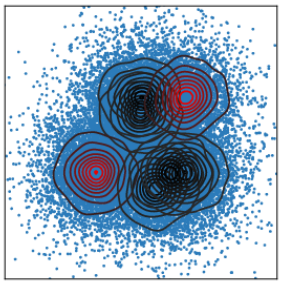


Fig. 1: Posterior probability distribution for 5 secrets obfuscated with a two-dimensional Laplace. The Bayes security metric is the complement of the total variation distance between the posterior of the most distinguishable secrets (shown in red).

IV. The Bayes Security Metric under Composition

Some of the properties that made DP so popular for studying complex algorithms are its compositionality rules: given DP-compliant mechanisms, it is very easy to determine the privacy of a mechanism that combines them (e.g., by chaining them). Further, compositionality enables studying complex threat scenarios. For example, while so far we have only considered an adversary who observes the channel's output once, it is common that a real adversary observes more than one channel at a time; e.g., observing obfuscated locations at different layers [35] or combining side channels [33]. Moreover, they can observe the output of two sequential channels, e.g., users' privacy-preserving interactions with a database through anonymous communication channels [34]; or they can observe more than one output from one channel, e.g., by gathering several side channel measurements from a hardware running cryptographic routines [25], or observing more than one visit to a website through an anonymous communication channel [24], [37].

In this section, we uncover compositionality rules for Bayes security, which enable it to tackle the above examples.

A. Parallel composition

We first consider an adversary who has access to the outputs of two channels that have as input the same secret [33], [35] or where an adversary observes multiple channel outputs belonging to the same secret [24], [25], [37].

Given two channels, $\mathcal{C}_1 : \mathbb{S} \to \mathbb{O}^1$ and $\mathcal{C}_2 : \mathbb{S} \to \mathbb{O}^2$, their parallel composition is the channel $\mathcal{C}^1 || \mathcal{C}^2 : \mathbb{S} \to \mathbb{O}^1 \times \mathbb{O}^2$, defined by $(\mathcal{C}^1 || \mathcal{C}^2)_{s,(o_1,o_2)} = \mathcal{C}^1_{s,o_1} \cdot \mathcal{C}^2_{s,o_2}$.

Theorem 3. *For all channels $\mathcal{C}^1, \mathcal{C}^2$ it holds that*

$$\beta^*(\mathcal{C}^1 || \mathcal{C}^2) \;\geq\; \beta^*(\mathcal{C}^1) \cdot \beta^*(\mathcal{C}^2) \,.$$

In layman's terms, the composition of two channels that are respectively β^*_1-secure and β^*_2-secure leads to a $\beta^*_1\beta^*_2$-secure channel. This bound is tight.

Note that the security of this new channel is *not necessarily* minimized by the secrets that minimize the composing channels $\mathcal{C}^1, \mathcal{C}^2$, not even when the channel is composed with itself:

Proposition 1. *Let $\mathcal{C}$ be a channel for which β is minimized for secrets (s_1, s_2). Then the composition channel $\mathcal{C}' := \mathcal{C}||\mathcal{C}$ is not necessarily minimized by secrets (s_1, s_2).*

Proof. A counterexample follows.

$$\mathcal{C} = \begin{bmatrix} 0.9 & 0.1 & 0.0 \\ 0.8 & 0.2 & 0.0 \\ 0.5 & 0.5 & 0.0 \\ 0.5 & 0.1 & 0.4 \end{bmatrix}$$

$\beta^*(\mathcal{C}) = 0.6$ is obtained for secrets (s_1, s_3); $\beta^*(\mathcal{C}||\mathcal{C}) = 0.36$ is achieved for secrets (s_2, s_4). □

B. Chaining mechanisms

Another typical configuration, used to strengthen the security of the system, is to put in place a cascade of security mechanisms (in-depth security). More formally, consider two channels $\mathcal{C}^1 : \mathbb{S}^1 \mapsto \mathbb{S}^2$ and $\mathcal{C}^2 : \mathbb{S}^2 \mapsto \mathbb{O}$. Their *cascade* composition is the channel $\mathcal{C}^1\mathcal{C}^2$ in which the secret is input in $\mathcal{C}^1$ and this channel's output is post-processed by $\mathcal{C}^2$.

It is well understood that post-processing cannot decrease the security of a mechanism. Therefore, $\mathcal{C}^1\mathcal{C}^2$ should be at least as secure ask $\mathcal{C}^1$. Indeed, based on the concavity of R^*, it is easy to show that $R^*(\pi, \mathcal{C}^1\mathcal{C}^2) \geq R^*(\pi, \mathcal{C}^1)$ for any prior π. Consequently, $\beta(\pi, \mathcal{C}^1\mathcal{C}^2) \geq \beta(\pi, \mathcal{C}^1)$.

Understanding the effect of $\mathcal{C}^1$ on $\mathcal{C}^2$ is less straightforward. The composition $\mathcal{C}^1\mathcal{C}^2$ can be seen as the *pre*-processing of $\mathcal{C}^2$, which is not necessarily a safe operation. Note that $\mathcal{C}^2$ receives as input the output of $\mathcal{C}^1$, which is not necessarily the same as $\mathbb{S}^1$. Hence, the prior π on the input secret in $\mathbb{S}^1$ is meaningless for $\mathcal{C}^2$. Remarkably, as β^* does not depend on the prior, it allows to compare $\mathcal{C}^1\mathcal{C}^2$ and $\mathcal{C}^2$ despite the different input spaces. From Theorem 2 we know that $\beta^*(\mathcal{C}^1\mathcal{C}^2)$ is given by the maximum ℓ_1 distance between the rows of $\mathcal{C}^1\mathcal{C}^2$. The key observation is that the rows of $\mathcal{C}^1\mathcal{C}^2$ are *convex combinations* of the rows of $\mathcal{C}^2$; but convex combinations cannot increase distances, which brings us to the following result.

Theorem 4. *For all channels $\mathcal{C}^1, \mathcal{C}^2$ it holds that*

$$\beta^*(\mathcal{C}^1\mathcal{C}^2) \;\geq\; \max\{\beta^*(\mathcal{C}^1), \beta^*(\mathcal{C}^2)\}\,.$$

This means that *neither pre-processing nor post-processing* decreases the Bayes security provided by a mechanism.

V. Relation with other notions

In the previous sections, we presented Bayes security, discussed its properties and showed how to compute it in an efficient manner. In this section, we compare it with three well-known security notions: cryptographic advantage, a mainstream threat-specific metric in the security community; DP, the paradigmatic worst-case metric; and multiplicative Bayes vulnerability leakage, which is closely related to β but comes with different properties.

IND-MINBAY$^{\mathcal{A}}_{\mathcal{C}}$	IND-LDP$^{\mathcal{A}}_{\mathcal{C}}$
1 : $\mathcal{A} \leftarrow \mathcal{C}, \pi$	1 : $\mathcal{A} \leftarrow \mathcal{C}, \pi$
2 : $\mathcal{A}$ selects $s_1, s_2 \in \mathbb{S}$	2 : $\mathcal{A}$ selects $s_1, s_2 \in \mathbb{S}$
3 : $s \overset{\pi_{1,2}}{\leftarrow} \{s_1, s_2\}$	3 : $\mathcal{A}$ selects $o \in \mathbb{O}$
4 : $o \overset{P(o\|s)}{\leftarrow} \mathbb{O}$	4 : $s \overset{P(s\|o)}{\leftarrow} \{s_1, s_2\}$
5 : $s' \leftarrow \mathcal{A}(o)$	5 : $s' \leftarrow \mathcal{A}(o)$
6 : **return** $s = s'$	6 : **return** $s = s'$

Fig. 2: Security games for β^* (left) and LDP (right).

A. Cryptographic advantage

In cryptography, the advantage Adv of an adversary $\mathcal{A}$ is defined assuming that there are two secrets ($|\mathbb{S}| = 2$) with a uniform prior as input to a channel $\mathcal{C}$. Formally (e.g. [39]):

$$\mathsf{Adv}(\mathcal{C}, \mathcal{A}) \;\overset{\text{def}}{=}\; 2|R^{\mathcal{A}}(\upsilon, \mathcal{C}) - 1/2|\,.$$

The factor 2 serves to scale Adv within the interval $[0, 1]$.

Denoting by $\mathsf{Adv}(\mathcal{C})$ the advantage of the optimal (Bayes) adversary and considering a uniform prior $\pi = \upsilon$, we derive:

$$\beta(\upsilon, \mathcal{C}) \;=\; 1 - 2|R^*(\upsilon, \mathcal{C}) - G(\upsilon)| = 1 - \mathsf{Adv}(\mathcal{C}). \qquad (1)$$

Hence the Bayes security metric can be seen as a generalization of $1 - \mathsf{Adv}$ for which the secret space $\mathbb{S}$ may contain more than two secrets the prior is not necessarily uniform.[1]

Bayes security as IND-CPA security. In Section II, we introduced the IND-BAY game to formalize the adversarial setting captured by $\beta(\pi, \mathcal{C})$. When considering this game in the light of the minimizer, $\beta^*(\mathcal{C})$, and our main result Section III (β *is minimized on the two leakiest secrets*), the IND-BAY game becomes a version of the traditional IND-CPA cryptographic game that we call IND-MINBAY (Figure 2, left).

First, recall that the adversary has perfect knowledge of the prior π and the channel $\mathcal{C}$ (line 1). Then, as opposed to the IND-BAY game, where the adversary cannot influence the input, we allow $\mathcal{A}$ to select the secrets and provide them to the challenger (line 2). This is analogous to classical IND-CPA, and it allows to capture the worst-case inputs. Then the challenger selects one of the two secrets according to the prior π (line 3), and returns to the adversary an obfuscated version according to the channel probability matrix (line 4). The adversary guesses one of the two secrets (line 5), and wins the game if the guess is the secret selected by the challenger. The advantage of this adversary is equivalent to that of a CPA adversary guessing what message was encrypted by the challenger.

This equivalence of games reinforces that the Bayes security metric sits in the middle between average metrics (measuring the expected risk) and worst-case metrics (measuring the worst-case risk across the secrets). In the next part of this section we explore this relation further.

[1]Note that in cryptography the advantage is usually defined for a generic (and not necessarily optimal) adversary.

B. Local Differential Privacy

We investigate the relation between the privacy guarantees induced by Bayes security (and, more in general, β), and those induced by DP metrics.

For a parameter $\varepsilon \geq 0$, we say that a mechanism is ε-LDP (local DP) [13] if for every i, h, j:

$$\mathcal{C}_{s_i,o_j} \;\leq\; \exp(\varepsilon)\,\mathcal{C}_{s_h,o_j}\,. \tag{2}$$

LDP is a worst-case metric, while recall that β has the characteristics of an average metric. Therefore, we expect that LDP implies a lower bound on β, but not vice versa. The rest of this section is dedicated to analyzing this implication.

A game for LDP. We first illustrate the difference between the threat model of Bayes security and the one considered by Local Differential Privacy using security games. Figure 2, right, represents the game for local differential privacy (IND-LDP).[2] A first remarkable difference with respect to typical security games is that, in addition to selecting the secrets as the IND-MINBAY game, the adversary also chooses the observation (line 3). This captures a worst-case in which the adversary not only picks the most vulnerable inputs, but also the output that makes them easier to distinguish. Upon receiving the secrets and the observation, the challenger selects one of the secrets according to the probability that it caused the observation (line 4). A second difference with respect to typical games, and IND-MINBAY, is that the challenger *does not show the chosen value to the adversary*. Otherwise, it would be a trivial win. The adversary guesses a secret (line 5), and wins if this is the secret the challenger chose (line 6). Note that, because the adversary has much greater freedom in their choices, their chances to win are considerably greater than than in IND-MINBAY or traditional games. Therefore, the LDP game captures a stronger attacker than most cryptographic games, but it is much harder to map it to a realistic threat scenario.

LDP induces a lower bound on Bayes security. In general, if there are no restrictions on the channel matrix, the lowest possible value for β is 0; this is achieved when the adversary can identify the value of the secret from every observable with probability 1. Assuming that $|\mathbb{S}| \geq 2$ and that π is not concentrated on one single secret,[3] and that $\mathbb{S}$ contains at least two elements, β can only be zero if and only if the channel contains at most one non-0 value for each column.

If the matrix is $\exp(\varepsilon)$-LDP, however, then the ratio between two values on the same column is at most $\exp(\varepsilon)$. Intuitively, under this restriction, β cannot be 0 anymore: the best case for the adversary is when the ratio is as large as possible, i.e., when it is *exactly* $\exp(\varepsilon)$. In particular, in a 2×2 channel, we expect that the minimum β is achieved by a matrix that has the values $\exp(\varepsilon)/1+\exp(\varepsilon)$ on the diagonal, and $1/1+\exp(\varepsilon)$ in the other positions (or vice versa). The next theorem confirms this intuition, and extends it to the general case $n \times m$.

$$\overbrace{\begin{bmatrix} a & \cdots & a & b & \cdots & b \\ b & \cdots & b & a & \cdots & a \\ c & \cdots & c & c & \cdots & c \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c & \cdots & c & c & \cdots & c \end{bmatrix}}^{k \quad m-k} \qquad \overbrace{\begin{bmatrix} d & e & 0 & \cdots & 0 \\ e & d & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}}^{m-2}$$

Fig. 3: Two examples of $n \times m$ matrices $\mathcal{C}^*$ which achieve minimum β value $\beta^*(\mathcal{C}^*) = \frac{2}{1+\exp(\varepsilon)}$. In the first matrix: $a = \frac{\exp(\varepsilon)}{k(1+\exp(\varepsilon))}$, $b = \frac{1}{(m-k)(1+\exp(\varepsilon))}$ and $c = \frac{1}{m}$. In the second matrix: $d = \frac{\exp(\varepsilon)}{1+\exp(\varepsilon)}$ and $e = \frac{1}{1+\exp(\varepsilon)}$.

Theorem 5.

1) *If $\mathcal{C}$ is ε-LDP, then for every π we have $\beta(\pi, \mathcal{C}) \geq \frac{2}{1+\exp(\varepsilon)}$.*

2) *For every $n, m \geq 2$ there exists a $n \times m$ ε-LDP channel $\mathcal{C}^*$ such that $\beta^*(\mathcal{C}^*) = \frac{2}{1+\exp(\varepsilon)}$. Examples of such $\mathcal{C}^*$'s are illustrated in Figure 3.*

Bayes security does not induce a lower bound on LDP. Theorem 5 shows that ε-LDP induces a bound on Bayes security, and that we can express a strict bound that depends only on ε. The other direction does not hold. The main reason is that if a column contains both a 0 and a positive element, then ε-LDP cannot hold, independently from the value of β.

C. Approximate Differential Privacy

One may consider (ε, δ)-LDP [15]. This is a variant of LDP in which small violations to Equation 2 are tolerated. Precisely, a mechanism is (ε, δ)-LDP if for every $s_i, s_j \in \mathbb{S}$ and $O \subseteq \mathbb{O}$:

$$\sum_{o \in O} \left(\mathcal{C}_{s_i,o} - \exp(\varepsilon)\,\mathcal{C}_{s_j,o}\right) \;\leq\; \delta\,. \tag{3}$$

With (ε, δ)-LDP, a column may contain 0 and non-0 values, as long as the latter are smaller than δ. Similarly to pure DP, approximate DP is threat-agnostic; this makes it harder to match (ε, δ) values to the risk of an attack occurring.

Surprisingly, we observe a direct relation between Bayes security and the special case $(0, \delta)$-DP:

Proposition 2. *Let $\mathcal{C}$ be a β^*-secure channel. Then it is also $(0, \delta)$-LDP, with $\delta = 1 - \beta^*$.*

This comes from the fact that, for $\varepsilon = 0$, the LHS of Equation 3 becomes $\sum_{o \in O} \left(\mathcal{C}_{s_i,o} - \mathcal{C}_{s_j,o}\right)$, which is maximized for $O^* = \{o \in \mathbb{O} \mid \mathcal{C}_{s_i,o} > \mathcal{C}_{s_j,o}\}$. Observe that $\sum_{o \in O^*} \left(\mathcal{C}_{s_i,o} - \mathcal{C}_{s_j,o}\right)$ corresponds to the total variation between $\mathcal{C}_{s_i,o}$ and $\mathcal{C}_{s_j,o}$. Applying the equivalence between β^* and total variation (Theorem 2) concludes the argument.

[2] We use this game for a qualitative comparison between the metrics. However, we observe that the parameter ε of LDP can be recovered from this game by ensuring a uniform prior when sampling from $P(s \mid o) = P(o|s)/(2P(o))$ (i.e., $P(s_1) = P(s_2) = 1/2$), and by evaluating the game with the following success metric: $\varepsilon = \ln(V^*/1-V^*)$, where $V^* = \max_s P(s \mid o)$ is the probability that a Bayes-optimal adversary guesses the secret correctly.

[3] If the probability mass of π is concentrated on one secret, then $G(\pi) = 0$ and $\beta(\pi, \mathcal{C})$ is undefined. However also $R^*(\pi, \mathcal{C}) = 0$, and $\lim_{\pi' \to \pi} \beta(\pi', \mathcal{C}) = 1$.

The special case of $(0, \delta)$-LDP mechanisms is not commonly studied. Intuitively, it corresponds to a mechanism that is completely vulnerable, but only with probability δ. We hope that the direct correspondence between β^*, δ, and the cryptographic advantage can give further insights in the decision of the parameter choices for approximate DP.

D. Differential privacy [16]

Differential privacy is similar to LDP, except that it involves the notion of adjacent databases. Two databases x, x' are adjacent, denoted as $x \sim x'$, if x is obtained from x' by removing or adding one record.

The definition of ε-differential-privacy (ε-DP), in the discrete case, is as follows. A mechanism $\mathcal{K}$ is ε-DP if for every x, x' such that $x \sim x'$, and every y, we have

$$P(\mathcal{K}(x) = y) \;\leq\; \exp(\varepsilon)\, P(\mathcal{K}(x') = y).$$

A relation between the Bayes security and DP follows from an analogous result in [1] for the multiplicative Bayes leakage $\mathcal{L}^{\times}(\pi, \mathcal{K})$, and the correspondence between the latter and the Bayes security (cfr. Section V-E), which is given by

$$\beta(\pi, \mathcal{C}) = \frac{1 - (\max_s \pi_s)\, \mathcal{L}^{\times}(\pi, \mathcal{C})}{1 - \max_s \pi_s}. \tag{4}$$

The following result, proven by Alvim et al. [1], states that ε-DP induces a bound on the multiplicative vulnerability leakage, where the set of secrets are all the possible databases. The theorem is given for the bounded DP case, where we assume that the number of records present in the database is at most a certain number n, and that the set of values for the records includes a special value $\perp$ representing the absence of the record. The adjacency relation is modified accordingly: $x \sim x'$ means that x and x' differ for the value of exactly one record. We also assume that the cardinality v of the set of values is finite. Hence also the number of secrets (i.e., the possible databases) is finite.

Theorem 6 (From [1], Theorem 15)**.** *If $\mathcal{K}$ is ε-DP, then, for every π, $\mathcal{L}^{\times}(\pi, \mathcal{C})$ is bounded from above as*

$$\mathcal{L}^{\times}(\pi, \mathcal{C}) \;\leq\; \left(\frac{\mathsf{v}\, \exp(\varepsilon)}{\mathsf{v} - 1 + \exp(\varepsilon)}\right)^n.$$

and this bound is tight when π is uniform.

From Theorem 6 and Equation 4 we immediately obtain a bound also for the Bayes security:

Corollary 1. *If $\mathcal{K}$ is ε-DP, then, for every π, $\beta(\pi, \mathcal{C})$ is bounded from below as*

$$\beta(\pi, \mathcal{C}) \;\geq\; \frac{1 - (\max_s \pi_s) \left(\frac{\mathsf{v}\, \exp(\varepsilon)}{\mathsf{v}-1+\exp(\varepsilon)}\right)^n}{1 - \max_s \pi_s}.$$

and this bound is tight when π is the uniform distribution υ which assigns $1/\mathsf{v}^n$ to every database, in which case it case it can be rewritten as

$$\beta(\pi, \mathcal{C}) \;\geq\; \frac{\mathsf{v}^n - \left(\frac{\mathsf{v}\, \exp(\varepsilon)}{\mathsf{v}-1+\exp(\varepsilon)}\right)^n}{\mathsf{v}^n - 1}.$$

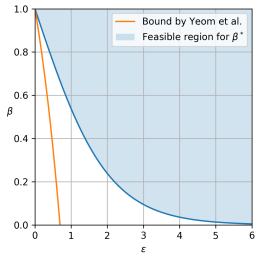


Fig. 4: The blue line illustrates the lower bound of ε-DP on β expressed by Corollary 2). The orange line represents the lower bound on β derived from the one proved in Yeom et al. [40] for the advantage of a membership inference adversary.

Alvim et al. [1] show that the reverse of Theorem 6 does not hold, and as a consequence the reverse of Corollary 1 does not hold either. The reason is analogous to the case of LDP: a 0 in a position of a non-0-column implies that the mechanism cannot be DP, independently from the value of β.

Membership inference. In Corollary 1 the secrets are the whole databases. Often, however, in DP we assume that the attacker is not interested at discovering the whole database, but only whether a certain record belongs to the database or not. We can model this case by isolating a generic pair of adjacent databases x and x', and then restricting the space of secrets to be just $\{x, x'\}$. On this space, the mechanism can be represented by a stochastic channel $\mathcal{C}^{\{x,x'\}}$ that has only the two inputs x and x', and as outputs the (obfuscated) answers to the query. It is immediate to see that $\mathcal{K}$ is ε-DP iff $\mathcal{C}^{\{x,x'\}}$ is ε-LDP for any pair of adjacent databases x and x'. Hence, the relations we proved between Bayes security and LDP hold also for DP. In particular, the following is an immediate consequence of Theorem 5.

Corollary 2. *If $\mathcal{K}$ is ε-DP, then for every pair of adjacent databases x and x' and every π we have*

$$\beta(\pi, \mathcal{C}^{\{x,x'\}}) \geq \frac{2}{1 + \exp(\varepsilon)}.$$

and this bound is strict.

A similar investigation was done by Yeom et al. [39]. They studied the privacy of $\mathcal{C}^{\{x,x'\}}$ in terms of the advantage, defined in the context of membership inference attacks (MIA). The authors established that, if a mechanism is ε-DP, then the following lower bound for holds for $\mathsf{Adv}(\mathcal{C}^{\{x,x'\}})$, for any adjacent databases x and x':

$$\mathsf{Adv}(\mathcal{C}^{\{x,x'\}}) \leq \exp(\varepsilon) - 1 \tag{5}$$

By using the relation between the advantage and the Bayes security metric (Equation 1), we derive the following bound:

$$\beta(\upsilon, \mathcal{C}^{\{x,x'\}}) \geq 2 - \exp(\varepsilon)\,. \tag{6}$$

where υ is the uniform distribution.

By exploiting the equivalence between Bayes security and the advantage, we conclude that the bound by Yeom et al. [39] is

loose; from Corollary 2 and Equation 1, we derive the following (strict) bound for the advantage of an ε-DP mechanism:

$$\mathsf{Adv}(\mathcal{C}^{\{x,x'\}}) \leq \frac{\exp(\varepsilon)-1}{\exp(\varepsilon)+1}.$$

which is much tighter than their bound (see Figure 4).

Concurrent work by Humphries et al. [22] proved a similar bound for (ε,δ)-DP in the context of membership inference attacks. Their bound is more general than ours, since it captures approximate DP; however, we prove tightness for our bound. Whether tightness can be proven for the bound by Humphries et al. is to our knowledge an open problem.

E. Leakage notions from Quantitative Information Flow

We discuss multiplicative risk leakage (β) and its minimizer (β^*) from the point of view of Quantitative Information Flow (QIF), and compare it with similar metrics stemming from the field. QIF measures the information leakage of a system by comparing its vulnerability *before* and *after* observing its output. It starts with a *vulnerability* metric $V(\pi)$, expressing how vulnerable the system is when the adversary has knowledge π about the secret. The *posterior* vulnerability is defined as $V(\pi,\mathcal{C}) = \sum_o p(o)V(\delta^o)$, where δ^o is the posterior distribution on $\mathbb{S}$ produced by the observation o; intuitively, it expresses how vulnerable the system is, on average, *after observing* the system's output. *Leakage* is defined by comparing the two, either *multiplicatively* or *additively*:

$$\mathcal{L}^{\times}(\pi,\mathcal{C}) = \frac{V(\pi,\mathcal{C})}{V(\pi)}\ , \quad \mathcal{L}^{+}(\pi,\mathcal{C}) = V(\pi,\mathcal{C}) - V(\pi)\ .$$

One of the most widely used vulnerability metrics is *Bayes vulnerability* [32], defined as $V(\pi) = \max_s \pi_s = 1 - G(\pi)$; it expresses the adversary's probability of guessing the secret correctly in one try.[4] For the posterior version, it holds that $V(\pi,\mathcal{C}) = 1 - R^*(\pi,\mathcal{C})$. The multiplicative risk leakage follows the same core idea: $G(\pi)$ can be thought of as a *prior* version of R^*: indeed, it holds that $R^*(\pi,\mathcal{C}) = \sum_o p(o)G(\delta^o)$ where δ^o are the posteriors of the channel. Hence, β can be considered to be a variant of multiplicative vulnerability leakage, using Bayes risk instead of Bayes vulnerability.

Since the two are closely related, one would expect to be able to directly translate results about $\mathcal{L}^{\times}(\pi,\mathcal{C})$ to similar results on $\beta(\pi,\mathcal{C})$. This would be the case for *additive leakage*, since $V(\pi,\mathcal{C}) - V(\pi) = G(\pi) - R^*(\pi,\mathcal{C})$, but in the multiplicative case, the "one minus" in both sides of the fraction completely changes the behavior of the function.

Capacity vs β^*. One should first note that, while β takes lower values to indicate a worse level of security, $\mathcal{L}^{\times}$ takes higher values. In both cases, a natural question is to find the prior π that provides the worst level of security; in the case of leakage, its maximum value is known as *channel capacity*, denoted by $\mathcal{ML}^{\times}(\mathcal{C}) = \max_\pi \mathcal{L}^{\times}(\pi,\mathcal{C})$.

$\mathcal{ML}^{\times}(\mathcal{C})$ is given by the *uniform prior* [6], and $\mathcal{ML}^{\times}(\mathcal{C}) = \sum_o \max_s \mathcal{C}_{s,o}$. Our main result (Theorem 1) shows that $\beta(\pi,\mathcal{C})$ is minimized on a uniform prior over 2 secrets. Hence, despite the similarity between Bayes vulnerability and Bayes risk, the corresponding leakage and security metrics behave very differently. Note that this difference makes $\mathcal{ML}^{\times}(\mathcal{C})$ easier to compute for arbitrary channels; it is linear on both $|\mathbb{S}|$ and $|\mathbb{O}|$, while β^* is quadratic on $|\mathbb{S}|$. We discuss fast ways for computing (or estimating) Bayes security in Section VII.

Channel composition. Despite their difference w.r.t. the prior that realizes each notion, $\mathcal{ML}^{\times}$ and β^* behave similarly w.r.t. parallel and cascade composition. It was shown that [19]:

$$\mathcal{ML}^{\times}(\mathcal{C}^1||\mathcal{C}^2) \leq \mathcal{ML}^{\times}(\mathcal{C}^1)\cdot\mathcal{ML}^{\times}(\mathcal{C}^2)\ , \quad \text{and}$$
$$\mathcal{ML}^{\times}(\mathcal{C}^1\mathcal{C}^2) \leq \min\{\mathcal{ML}^{\times}(\mathcal{C}^1), \mathcal{ML}^{\times}(\mathcal{C}^2)\}\ .$$

The same bounds are given in Section IV for β^*. Note, however, that the proofs for β^* are completely different and cannot be directly obtained from those of $\mathcal{ML}^{\times}$.

Bounds on Bayes risk. The goal of a security analyst is to quantify how much information is leaked by a mechanism in the *worst case*. This is captured by both $\mathcal{ML}^{\times}$ and β^* which focus on the prior that produces the highest leakage instead of the true prior. The user, however, is mostly interested in the actual threat that he is facing: how likely it is for the adversary to guess his secret, given a particular prior π that captures the user's behavior. In this sense, the Bayes risk, $R^*(\pi,\mathcal{C})$, has a clear operational interpretation for the user.

Fortunately, having computed either $\mathcal{ML}^{\times}$ or β^*, we can obtain direct bounds for the prior π of interest:

$$R^*(\pi,\mathcal{C}) \geq \beta^*(\mathcal{C})\cdot G(\pi)\ , \quad \text{and}$$
$$R^*(\pi,\mathcal{C}) \geq 1 - \mathcal{ML}^{\times}(\mathcal{C})\cdot V(\pi)\ .$$

The goodness of either bound depends on the application. Intuitively, how good the bound is depends on how close π is to the one achieving $\mathcal{ML}^{\times}$ or β^*. Concretely, since the former implies uniform priors, and the latter a vector with only 2 non-empty (uniform) entries, the tightness of these bounds depends on the sparsity of the real prior vector π.

We study empirically how tight these bounds are. We consider two channels with $|\mathbb{S}| = 10$ inputs and $|\mathbb{O}| = 1K$ outputs. The first channel (hereby referred to as the *random* channel) is obtained by sampling at random its conditional probability distribution $P(o \mid s)$; the second one (*geometric* channel) has a geometric distribution as used by Cherubin et al. [10], with a noise parameter $\nu = 0.1$. To evaluate the effect of sparsity, we set a sparsity level to $\sigma \in 0, 1, ..., n-2$, and we sample a prior that is σ-sparse uniformly at random. We compute the values of $\mathcal{L}^{\times}$ and β, and measure their absolute distance respectively from $\mathcal{ML}^{\times}$ and β^*. The experiment is repeated $1K$ times for each sparsity level.

Figure 5 shows the results. As expected, the multiplicative vulnerability leakage bound is tighter for vectors that are less sparse, and the Bayes security one for higher sparsity levels. However, we observe that the Bayes security bound is loose for high values of sparsity in the case of the geometric channel, but not for the random one. The reason is that if the real prior has maximum sparsity (i.e., only 2 non-zero entries), then it is

[4] The tightly connected notion of *min-entropy*, defined as $\log V(\pi)$, is used by many authors instead of Bayes vulnerability.

TABLE II: Security metrics comparison: Local Differential Privacy (LDP), Multiplicative leakage capacity, and Bayes Security (β^*). Note that the cryptographic advantage is a special case of β^*, and therefore not included in this table. "Consistent Black-box Estimation" refers to the existence of a statistically consistent estimator for the security metric (e.g., [10]).

Property	**LDP** (ε)	$\mathcal{ML}^\times$	β^*
Range	$[0, \infty)$ Smaller is more secure	$[1, n]$ Smaller is more secure	$[0, 1]$ Larger is more secure
Attacks	Any attack	Concrete attack	Concrete attack
Security Guarantee	Worst case for 2 leakiest secrets	Expected risk among all secrets	Expected risk for 2 leakiest secrets
Qualitative Intuition	Bounds probability of ever distinguishing two secrets	Bounds the probability of guessing among all secrets	(Complement of) the advantage of an attacker in guessing the secret w.r.t. the random baseline
Quantitative Intuition	None for general case. Can be defined w.r.t. mechanism	$\mathcal{ML}^\times = k$ means the adversary is k times more likely to guess the secret	$1 - 1/2\beta^*$ is the probability that the adversary guesses the secret correctly
Composable	✓	✓	✓
Consistent Black-box Estimation	✗	✓	✓
Prior-agnostic	✓	✓	✓

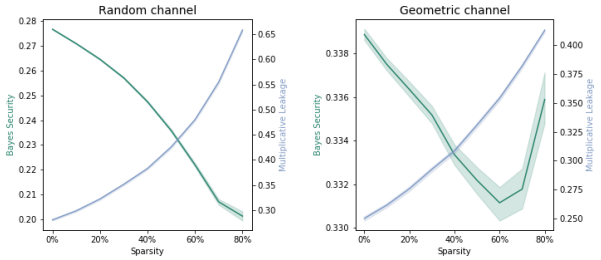


Fig. 5: Tightness of the bounds on Bayes security and multiplicative leakage with respect to sparsity. Note that, because the two metrics have different scales, these plots are useful to compare their behavior, and not their actual values.

more likely that the secrets on which β is minimized are not the same 2 secrets on which the prior is not empty.

As a consequence of this analysis, we suggest $\mathcal{ML}^\times$ is better suited to analyze deterministic mechanisms with a large number of secrets distributed close to uniformly (see [32]). For deterministic programs, β^* is always 0, unless the program is non-interfering; the bound obtained by β^* is then trivial, while $\mathcal{ML}^\times$ can provide meaningful bounds if the number of outputs is limited. On the other hand, β^* is advantageous if π is very sparse (e.g., website fingerprinting, where the user may be visiting only a small number of websites): since π is very different than the uniform one, and more similar to the one achieving β^*, the latter will provide much better bounds. We discuss application examples for Bayes security in Section VIII.

Miracle theorem. Both Bayes vulnerability and Bayes risk can be thought of as instantiations of a general family of metrics parameterized by a *gain function* g (for vulnerability) or a *loss function* ℓ (for risk). For generic choices of g and ℓ, we can define g-leakage $\mathcal{L}_g^\times(\pi, \mathcal{C})$ and the corresponding security notion $\beta_\ell(\pi, \mathcal{C})$, in a natural way (detailed in Appendix F).

A result by Alvim et al. [2], known as "miracle" due to its arguably surprising nature, states that

$$\mathcal{L}_g^\times(\pi, \mathcal{C}) \;\leq\; \mathcal{ML}^\times(\mathcal{C})\;,$$

for all priors π and all *non-negative* gain functions g. This gives a direct bound for a very general family of leakage metrics.

For β_ℓ, however, we know that a corresponding result does not hold in general, even if we restrict to the family of $[0, 1]$ loss functions. Identifying families of loss functions that provide similar bounds is left as future work.

VI. Case Studies: Bayes Security of Well-Known Mechanisms

We now exploit the various properties we proved about Bayes security to study well-known mechanisms: Randomized Response, and the Gaussian and Laplace mechanisms. These are often used as building blocks for more complex ones.

A. Randomized Response

Randomized Response (RR) is a simple obfuscation protocol that guarantees ε-LDP. It randomly assigns a data record to a new data record from the same range. Assuming $\mathbb{S} = \mathbb{O}$, RR is represented as the following channel matrix, $\mathcal{R} : \mathbb{S} \mapsto \mathbb{O}$:

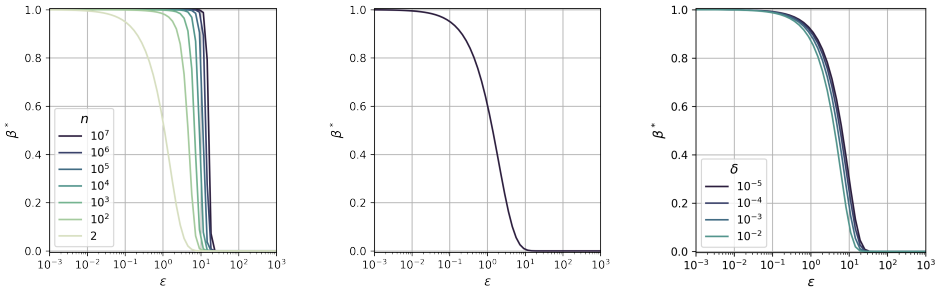
$$\mathcal{R}_{s,o} = P(o|s) \;\stackrel{\text{def}}{=}\; \begin{cases} \frac{exp(\varepsilon)}{n+\exp(\varepsilon)-1} & \text{if } o = s \\ \frac{1}{n+\exp(\varepsilon)-1} & \text{otherwise}\,. \end{cases}$$

We can derive β^* easily for RR because it obfuscates each secret according to the same distribution. For any two secrets s_i and s_j, the rows of the channel matrix are identical except for positions i and j, where their values are inverted; therefore, the Bayes security is the same between any two secrets, and thus all are equally vulnerable. Using the results in Section VII-A, we just need to look at any two rows, e.g., the first two. Let $\mathcal{R}_{ab}$ indicate the sub-channel matrix containing only the first two rows, and let $\upsilon = 1/2$. The corresponding Bayes risk is $R^*(\upsilon, \mathcal{R}_{ab}) = \frac{n}{2(\exp(\varepsilon)+n-1)}$; hence the Bayes security is:

$$\beta^*(\mathcal{R}) = \frac{n}{\exp(\varepsilon) + n - 1}\,, \tag{7}$$

where n is the number of secrets and observables.

Discussion. Equation 7 captures the risk that an optimal adversary can distinguish between any two data records (secrets) from the RR output; by Theorem 1, these are the easiest two



(a) Randomized Response (b) Laplace mechanism (c) Gaussian mechanism

Fig. 6: Relation between Bayes security and DP for various mechanisms; (ε, δ)-DP is used for the Gaussian mechanism.

records to distinguish, and it implies Bayes security for any other subset of the secret space.

We use this equation to relate Bayes security and ε w.r.t. the number of data records (secrets) in Figure 6 (a). We observe that the number of data records is essential for security. For a rather loose DP parameter of $\varepsilon = 10$, having a dataset of $1M$ gives $\beta^* \approx 0.978$; assuming the two data records have the same prior, the probability that the adversary guesses correctly is 0.511. With the same ε, having a dataset of $10M$ ensures a practically perfect Bayes security of 0.998 (Bayes vulnerability for a uniform prior: 0.501). Overall, this shows that, if we are interested in a specific threat model(s), then a threat-specific metric such as Bayes security can reassure us on the security of the mechanism even when DP suggests it is not.

In the appendix (Section H), we include an empirical study of RR for the `Census1990` dataset. We observe that, in this specific case, a good utility (95%) is only achieved for a rather large $\varepsilon = 3.3$; in principle, one would disregard the mechanism to be unsafe in this particular instance. Yet, Bayes security ($\beta^* = 0.99999$ for $\varepsilon = 3.3$, and $\beta^* = 0.99995$ for $\varepsilon = 4.8$) reassures us on its security within this threat model.

B. Laplace mechanism

For parameter a λ, we define the Laplace mechanism as $\mathcal{L} : s \mapsto s + \Lambda(0, \lambda)$, where $\Lambda(\mu, \lambda)$ is a μ-centered Laplace distribution with scale λ.

Proposition 3. *$\mathcal{L}$ is β^*-secure with*

$$\beta^* = \exp\left(-\max_{s_i, s_j \in \mathbb{S}} \frac{|s_i - s_j|}{2\lambda}\right),$$

Discussion. We can use this analysis to compare β^* with DP. Let $f : \mathcal{D} \mapsto \mathbb{R}$ be a real-valued function with sensitivity $\Delta f \stackrel{\text{def}}{=} \max_{x,y \in \mathcal{D}} f(x) - f(y)$. The mechanism $f(x) + \mathcal{L}(0, \lambda)$ with scale parameter $\lambda = \frac{\Delta f}{\varepsilon}$ is ε-DP. By using the last result, the Bayes security of this mechanism is $\beta^* = \exp(-\frac{\varepsilon}{2})$.

Now, suppose we care about the probability of an adversary at distinguishing the two maximally distant points $s_1, s_2 = \arg\max_{x,y \in \mathcal{D}} f(x) - f(y)$. For a relatively strong DP level of $\varepsilon = 0.1$, we get $\beta^* \approx 0.95$; This implies a non-negligible advantage for the adversary; e.g., assuming the two points have identical prior $1/2$, the probability that the optimal adversary distinguishes between them is roughly 0.525. Figure 6 (b) shows the overall behavior.

C. Gaussian mechanism

For parameter σ, the Gaussian mechanism adds noise to a secret s from a Gaussian distribution: $\mathcal{G} : s \mapsto s + \mathcal{N}(0, \sigma^2)$.

Proposition 4. *$\mathcal{G}$ is β^*-secure with $\beta^* = 1 - (\Phi(\alpha) - \Phi(-\alpha))$, where Φ is the CDF of $\mathcal{N}(0, 1)$, and $\alpha = \max_{s_i, s_j} \frac{|s_i - s_j|}{2\sigma}$.*

Discussion. Because the Gaussian mechanism does not satisfy pure DP, we compare Bayes security with approximate DP. For a function f with sensitivity Δf, and for $\varepsilon < 1$, the following mechanism satisfies (ε, δ)-DP: $f(x) + \mathcal{N}(0, \frac{2\ln(1.25/\delta)(\Delta f)^2}{\varepsilon^2})$.

By applying Proposition 4 we obtain $\beta^* = 1 - (\Phi(\alpha) - \Phi(-\alpha))$ with $\alpha = \varepsilon/2\sqrt{2\ln(1.25/\delta}$. As desired, security does not depend on the sensitivity of the function.

We observe a similar behavior to what we observed for the Laplace mechanism (Figure 6 (c)). Consider a dataset containing $N = 1K$ records, for which an appropriate choice of δ according to the literature is $\delta = 1/N^2$. For a relatively secure setting ($\varepsilon = 1$), we have $\beta^* = 0.925$. As before, an interpretation of this value is that an optimal attacker will distinguish the two most vulnerable secrets with probability 0.538; this is clearly non-negligible. We note that only a stricter value such as $\varepsilon = 0.1$ ensures a strong guarantee against the attack ($\beta^* = 0.992$).

Overall, Bayes security enabled us to interpret the privacy guarantees of various mechanisms, by matching them back to the probability of success of an optimal attacker under a specific threat model.

VII. Computational estimation of β^*

Suppose that, differently from the cases we just analyzed (Section VI), a simple closed-form expression of the mechanism does not exist: how can we determine its Bayes security? Theorem 1 shows that to quantify Bayes security, the minimizer of multiplicative risk leakage β, we just need to estimate β for all pairs of secrets; this requires $\mathcal{O}(n^2)$ measurements. A measurement for a pair of secrets is obtained by estimating the Bayes risk of the mechanism for those two secrets; we can do this analytically, if we have white-box knowledge of

the mechanism, or in a black-box manner[5]. In either case, if the mechanism is complex enough (e.g., large input or output space), each measurement may need a non-negligible computational time, from seconds to tens of minutes.

In this section, we investigate techniques for improving the search time. Since the bottleneck is the time it takes to measure $\beta(\pi, \mathcal{C})$ for one prior π, we seek to reduce the number of such measurements. We first assume white-box knowledge of the system (subsections VII-A-VII-C), and then study the black-box case (subsection VII-D).

Initial observations. Denote by π_{ab} be the sparse prior vector $(0, ..., 0, 1/2, 0, ..., 0, 1/2, 0, ..., 0)$ such that the two non-zero elements of value $1/2$ are in positions a and b, and $a \neq b$. Given a channel $\mathcal{C}$, from the definition of β we get that

$$\beta(\pi_{ab}, \mathcal{C}) = 2 - \sum_o \max_{s \in \{a,b\}} \mathcal{C}_{s,o} \ .$$

The crucial observation (shown in the proof of Theorem 2) is that the above quantity is equal to the complement of the *total variation* distance $\mathrm{tv}(\mathcal{C}_a, \mathcal{C}_b)$ between the rows $\mathcal{C}_a$ and $\mathcal{C}_b$ of the channel. The total variation distance of two discrete distribution is $1/2$ of their L_1 distance (seen as vectors); hence:

$$\beta(\pi_{ab}, \mathcal{C}) \ = \ 1 - \mathrm{tv}(\mathcal{C}_a, \mathcal{C}_b) \ = \ 1 - \frac{1}{2} \left\| \mathcal{C}_a - \mathcal{C}_b \right\|_1 \ .$$

Then, from Theorem 1, we get that minimizing β is equivalent to finding the rows of the channel that are maximally distant with respect to L_1. This is the well-known *diameter problem* (for L_1): given the set of vectors $\mathcal{C}_{\mathbb{S}}$, find the two that are maximally distant (i.e., find the diameter of the set).

A. Computing β^* with domain knowledge

In practical applications, domain knowledge may enable *a priori* identification of the two leakiest secrets. For example, the smallest and largest webpages users can visit in website fingerprinting (Section VIII); and the smallest and largest exponents in timing side channels against exponentiation algorithms [10]. There are also applications where all the secrets are equally vulnerable; hence β^* is obtained for any pair of distinct secrets. For instance, when the mechanism operates in such a way that all secrets enjoy the same protection (e.g., the Randomized Response mechanism, Section VI).

More generally, if one does not know the exact minimizing secrets, but knows that they belong to a set $\mathbb{S}' \subset \mathbb{S}$, then to determine β^* it suffices measuring β for all $s_1, s_2 \in \mathbb{S}'$.

B. Computing β^* in linear time n

The geometric characterization given by Theorem 2 implies that obtaining β^* requires computing the diameter of a set of $n = |\mathbb{S}|$ vectors of dimension $m = |\mathbb{O}|$. The direct approach is to compute the distance between every pair of vectors, i.e., perform $O(n^2 m)$ operations. This quadratic dependence on n can be prohibitive when the number of secrets grows.

[5]We remark that black-box estimation of the Bayes risk (and, therefore, Bayes security) can be done consistently via distribution-free techniques [10].

We first show that, by using an isometric embedding of L_1^m into $L_\infty^{2^m}$, β^* can be computed in $O(n2^m)$ time. Concretely, each $x \in \mathbb{R}^m$ is translated into a vector $\phi(x) \in \mathbb{R}^{2^m}$, which has one component for every bitstring b of length m, such that $\phi(x)_b = \sum_{i=1}^m x_i (-1)^{b_i}$. Note that the equivalence $\|\phi(x) - \phi(x')\|_\infty = \|x - x'\|_1$ holds for all $x, x' \in \mathbb{R}^m$. The L_∞ diameter problem can be solved in linear time: we only need to find the maximum and minimum value of each component.

This computation is linear in $|\mathbb{S}|$ but exponential in $|\mathbb{O}|$. It outperforms the direct approach when the number of observations is small, but the problem becomes harder as the number of observations grows. When $m = \Theta(n)$ there is no sub-quadratic algorithm for the L_p-diameter problem for any $p \geq 0$ [12] . This suggests that there may not be any sub-quadratic time for computing β^* either.

C. An efficient approximation of β^*

We present an estimation of β^* that can be obtained in $O(nm)$ time. One selects an arbitrary distribution $q \in \mathcal{D}(\mathbb{O})$ and computes the maximal distance d between any channel row and q. The diameter of $\mathcal{C}_{\mathbb{S}}$ is at most $2d$, giving a lower bound on β^*. Furthermore, if q lies within the convex hull of $\mathcal{C}_{\mathbb{S}}$ (denoted by $\mathrm{ch}(\mathcal{C}_{\mathbb{S}})$), then the diameter is at least d, giving also an upper bound:

Proposition 5. *Let $\mathcal{C}$ be a channel, $q \in \mathcal{D}(\mathbb{O})$, and $d = \max_{s \in \mathbb{S}} \|\mathcal{C}_s - q\|_1$. Then $1 - d \ \leq \ \beta^*(\mathcal{C})$. Moreover, if $q \in \mathrm{ch}(\mathcal{C}_{\mathbb{S}})$ then $\beta^*(\mathcal{C}) \ \leq \ 1 - d/2$.*

Good choices for q are distributions that are likely to lie "in-between" the two maximally distant rows, for instance the *centroid* of $\mathcal{C}_{\mathbb{S}}$ (mean of all rows).

Several advanced approximation algorithms exist for the L_2 diameter problem [23]; these could be employed using some embedding of L_1 into L_2. The trivial embedding has distortion $\sqrt{m}$ (since $\|x\|_2 \leq \|x\|_1 \leq \sqrt{m}\|x\|_2$), hence the approximation factor may be too loose as $|\mathbb{O}|$ grows. Low distortion embeddings of L_1 into L_2 exist [3], but it is unclear if they can be applied to the diameter problem. In Section I, we conduct an empirical study of these approximations.

D. Black-box estimation of β^*

The previous sections assume full knowledge of the channel $\mathcal{C}$. In practice, this assumption may fail: systems may be too complex to analyze, or their behavior may be unknown. In such cases, we can estimate the Bayes risk, and therefore β, using black-box estimation tools (i.e., only observing the system's inputs and outputs), such as `F-BLEAU` [10]. As with the white-box case, we need to reduce the number of priors π for which we estimate $\beta(\pi, \mathcal{C})$.

Bounds. A first approach is to use the bounds given by Proposition 5, which can be computed in a black-box setting. One can interact with the system to obtain observations for q. For instance observe: the *mean row*, by drawing observations from the channel with secrets that are chosen uniformly at random; the *any row of the channel*, by drawing observation for

a secret chosen arbitrarily; or *a row with arbitrary distribution*, e.g., by sampling q uniformly at random from the set $\mathbb{O}$.

Building upon R^* black-box estimators [10]. If domain constraints do not enable identifying the pair of leakiest secrets (Section VII-A), we can try to reduce the search space. For instance, we can exploit the triangle inequality on the total variation distance to discard some solutions before computing them. E.g., given the Bayes security for the priors π_{ac} and π_{bc}:

$$\begin{aligned}\beta(\pi_{ac},\mathcal{C}) + \beta(\pi_{bc},\mathcal{C}) - 1 \leq \beta(\pi_{ab},\mathcal{C}) \\ \leq 1 - |\beta(\pi_{ac},\mathcal{C}) - \beta(\pi_{bc,\mathcal{C}})|\,.\end{aligned}$$

Thus, if $\beta(\pi_{ab},\mathcal{C})$ is larger than some already-known $\beta(\pi_{ij},\mathcal{C})$ there is no need to compute it. Conversely, if it is upper bounded by a small quantity, we can compute it earlier aiming at discarding other combinations.

VIII. Discussion and Conclusions

This paper provides building blocks for studying complex algorithms on the basis of Bayes security, a metric that generalizes the cryptographic advantage. Bayes security inherits benefits from both average-case metrics, such as advantage and Bayes risk, and worst-case metrics, such as DP. Similarly to the advantage, Bayes security is threat-specific: it captures the risk for the users in a specified threat model (e.g., what's the probability that a user's data record is leaked). Like DP, Bayes security is easily composable, and it reflects the *worst-case* for the two most vulnerable secrets (e.g., data records). Yet, Bayes security is a weaker worst-case notion than DP, which may enable utility gains in high-security regimes (Section VI).

Applications. The above characteristics make Bayes security suitable for a broad range of security and privacy settings. Below, we discuss some particularly fitting examples.

Website fingerprinting. In website fingerprinting (WF), an adversary with access to an encrypted network tunnel (e.g., VPN or Tor) aims to infer the websites being visited by a user. The *success rate* (or *accuracy*) of an attacker has been used for years as a way of evaluating an attack's goodness. However, this metric suffers from some drawbacks [24], [36]. First, comparing success rate across studies is meaningless, as the number of websites the user can visit strongly affects it: the attack is very simple is the user is only allowed to visit 2 websites as opposed to 100. Second, the prior probability of each website being visited highly skews the success rate; if a website is easy to distinguish from the others and it is very likely to be visited, then the attacker's accuracy would be largely inflated. The use of Mutual Information was suggested as an alternative metric [28]. However, Smith showed that this metric does not capture the standard threat model used in WF, and it may be misleading if we are ultimately interested in learning about an attacker's success probability [32].

β was introduced for WF evaluation [11], although without any theoretical justification. In this work, we developed a theory for β, and we showed that its minimizer, the Bayes security metric, is particularly suited for WF: i) it is prior independent; ii) it measures the risk for the two leakiest secrets (i.e., the two websites that are the easiest to tell apart); iii) as shown in Figure 5, it is captures particularly well the case of sparse prior – in WF, the prior over websites is highly sparse. Overall, this suggests Bayes security is an appropriate choice evaluating the user's risks against WF and, similarly, the information leakage of WF defenses. Future work may study if Bayes security implies bounds w.r.t. other metrics of interest, such as True/False positives or Precision and Recall (Section V).

PPML. We suspect privacy preserving ML (PPML) algorithms can be easily studied by using Bayes security. Its strengths for this kind of analysis are: i) it is easy to derive it analytically (e.g., as the total variation of the posterior for the two leakiest secrets) (Section III); ii) for a large secret space (e.g., data records in a dataset), it characterizes the risk for the most vulnerable ones; this, we argue, gives an easy interpretation of its guarantees; iii) its prior independence helps studying mechanisms irrespective of the adversary's prior knowledge; and, once the attacker's prior is known, it can be plugged in to better capture the risk (Section V); iv) where an analytical study is not possible, Bayes security can be easily estimated in a black-box manner (Section VII). Overall, we expect future work can provide Bayes security-style guarantees for complex ML training pipelines. For example, by exploiting our results on the Gaussian mechanism (Section VI), it may be possible to study the security of DP-SGD against common attacks such as membership inference [31], attribute inference [20], and reconstruction [4], [7]. This will enable bypassing bounds relating ε and the advantage [22], [40], by computing the advantage (or Bayes security) directly. One immediate implication of Theorem 1 is that evaluating membership inference attacks via the cryptographic advantage (which, in this case, matches Bayes security), gives guarantees for any prior probability that "members" may have.

Data release mechanisms. Our analysis in Section VI suggests that, when defending large datasets, Bayes security may help getting better utility than DP in high-privacy regimes.

Fairness. Bayes security captures the risk for the most vulnerable pair of users (Theorem 1). We suspect this characteristic can be adapted for evaluating privacy fairness (e.g., whether some population subgroups enjoy better privacy than others).

Further extensions. In this paper, we discussed various extensions that may further improve Bayes security's suitability to tackle complex algorithms. For example, proving a form of the *miracle theorem* (Section V-E) would give analysts even further flexibility when defining threat models for real-world attacks. Moreover, given the equivalence between Bayes security and total variation (Theorem 2), it may be possible to exploit research on total variation estimation to improve black-box leakage estimation techniques.

In conclusion, Bayes security opens a new space in the security metrics space, offering designers the opportunity to obtain different trade-offs than previous metrics. As we showed in Section VI, these trade-offs enable the choice of security parameters that provide strong protection and potentially with less utility impact under the threat model one chooses.

ACKNOWLEDGMENT

The work of Catuscia Palamidessi has been funded by the European Research Council (ERC) grant Hypatia, grant agreement N. 835294. We are grateful to Boris Köpf, Andrew Paverd, and Santiago Zanella-Beguelin for useful discussion. We are especially thankful to Borja Balle and Lukas Wutschitz for proof-reading our manuscript, and for spotting the equivalence between Bayes security and a special case of approximate differential privacy.

REFERENCES

[1] Mário S. Alvim, Miguel E. Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. On the information leakage of differentially-private mechanisms. *J. of Comp. Security*, 23(4):427–469, 2015.

[2] Mário S. Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Geoffrey Smith. Measuring information leakage using generalized gain functions. In *Proc. of CSF*, pages 265–279. IEEE, 2012.

[3] Sanjeev Arora, James R. Lee, and Assaf Naor. Euclidean distortion and the sparsest cut. In *STOC*, pages 553–562. ACM, 2005.

[4] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. *arXiv preprint arXiv:2201.04845*, 2022.

[5] Mihir Bellare and Phillip Rogaway. Introduction to modern cryptography. *Ucsd Cse*, 207:207, 2005.

[6] Christelle Braun, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Quantitative notions of leakage for one-try attacks. *ENTCS*, 249:75–91, 2009.

[7] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[8] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Prakash Panangaden. On the Bayes risk in information-hiding protocols. *J. of Comp. Security*, 16(5):531–571, 2008.

[9] Giovanni Cherubin. Bayes, not naïve: Security bounds on website fingerprinting defenses. *Proceedings on Privacy Enhancing Technologies*, 4:215–231, 2017.

[10] Giovanni Cherubin, Kostantinos Chatzikokolakis, and Catuscia Palamidessi. F-BLEAU: Fast black-box leakage estimation. In *Proc. of S&P*, pages 1307–1324. IEEE, 2019.

[11] Giovanni Cherubin, Jamie Hayes, and Marc Juarez. Website fingerprinting defenses at the application layer. *Proceedings on Privacy Enhancing Technologies*, 2:165–182, 2017.

[12] Roee David, Karthik C. S., and Bundit Laekhanukit. The curse of medium dimension for geometric problems in almost every norm. *CoRR*, abs/1608.03245, 2016.

[13] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *Proc. of FOCS*, pages 429–438. IEEE Computer Society, 2013.

[14] Cynthia Dwork. Differential privacy. In *Proc. of ICALP*, volume 4052 of *LNCS*, pages 1–12. Springer, 2006.

[15] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Proc. of EUROCRYPT*, volume 4004 of *LNCS*, pages 486–503. Springer, 2006.

[16] Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. of TCC*, volume 3876 of *LNCS*, pages 265–284. Springer, 2006.

[17] Kevin P. Dyer, Scott E. Coull, Thomas Ristenpart, and Thomas Shrimpton. Peek-a-Boo, I Still See You: Why Efficient Traffic Analysis Countermeasures Fail. In *Proc. of S&P*, pages 332–346. IEEE, 2012.

[18] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In *Proc. of CCS*, pages 1054–1067. ACM, 2014.

[19] Barbara Espinoza and Geoffrey Smith. Min-entropy leakage of channels in cascade. In *Proc. of FAST*, volume 7140 of *LNCS*, pages 70–84. Springer, 2011.

[20] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[21] Jamie Hayes and George Danezis. k-fingerprinting: a Robust Scalable Website Fingerprinting Technique. In *25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016.*, pages 1187–1203. USENIX Association, 2016.

[22] Thomas Humphries, Matthew Rafuse, Lindsey Tulloch, Simon Oya, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. Differentially private learning does not bound membership inference. *arXiv preprint arXiv:2010.12112*, 2020.

[23] Mahdi Imanparast, Seyed Naser Hashemi, and Ali Mohades. An efficient approximation for point-set diameter in higher dimensions. In *CCCG*, pages 72–77, 2018.

[24] Marc Juarez, Sadia Afroz, Gunes Acar, Claudia Diaz, and Rachel Greenstadt. A critical evaluation of website fingerprinting attacks. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 263–274, 2014.

[25] Paul C. Kocher. Timing attacks on implementations of diffie-hellman, rsa, dss, and other systems. In *Advances in Cryptology - CRYPTO '96*, 1996.

[26] Boris Köpf, Laurent Mauborgne, and Martín Ochoa. Automatic quantification of cache side-channels. In *Computer Aided Verification - 24th Int. Conf., CAV 2012, Berkeley, CA, USA, July 7-13, 2012 Proceedings*, volume 7358 of *LNCS*, pages 564–580. Springer, 2012.

[27] Jerry Li. Lecture 2 - Lecture notes in Robustness in Machine Learning. https://jerryzli.github.io/robust-ml-fall19/lec2.pdf, 2019.

[28] Shuai Li, Huajun Guo, and Nicholas Hopper. Measuring information leakage in website fingerprinting attacks and defenses. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1977–1992, 2018.

[29] Takao Murakami, Hideitsu Hino, and Jun Sakuma. Toward distribution estimation under local differential privacy with small samples. *PoPETs*, 2018(3):84–104, 2018.

[30] Takao Murakami and Yusuke Kawamoto. Utility-optimized local differential privacy mechanisms for distribution estimation. In *USENIX Security.*, 2019.

[31] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proc. of S&P*, 2017.

[32] Geoffrey Smith. On the foundations of quantitative information flow. In *Proc. of FOSSACS*, volume 5504 of *LNCS*, pages 288–302. Springer, 2009.

[33] François-Xavier Standaert and Cédric Archambeau. Using subspace-based template attacks to compare and combine power and electromagnetic information leakages. In *Cryptographic Hardware and Embedded Systems (CHES)*, 2008.

[34] Raphael R. Toledo, George Danezis, and Ian Goldberg. Lower-cost $\in$-private information retrieval. *PoPETs*, 2016.

[35] Huandong Wang, Chen Gao, Yong Li, Gang Wang, Depeng Jin, and Jingbo Sun. De-anonymization of mobility trajectories: Dissecting the gaps between theory and practice. In *Proc. of NDSS*, 2018.

[36] Tao Wang. High precision open-world website fingerprinting. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 152–167. IEEE, 2020.

[37] Tao Wang, Xiang Cai, and Ian Goldberg. Website fingerprinting. https://crysp.uwaterloo.ca/software/webfingerprint/, 2013.

[38] Charles V. Wright, Lucas Ballard, Fabian Monrose, and Gerald M. Masson. Language identification of encrypted voip traffic: Alejandra y roberto or alice and bob? In *Proceedings of the 16th USENIX Security Symposium, Boston, MA, USA, August 6-10, 2007*, 2007.

[39] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proc. of CSF*, pages 268–282. IEEE Computer Society, 2018.

[40] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proc. of CSF*, pages 268–282. IEEE, 2018.

APPENDIX

A. In general, β is not minimized by the uniform prior

We start with the following lemma.

Lemma 1. *Suppose that $\beta(\pi,\mathcal{C}) = 0$, for some system $(\pi,\mathcal{C})$, and that π has k non-zero components. Let $\pi' = (1/k, ..., 1/k, 0, ..., 0)$, where the non-zero components are in correspondence of the non-zero components of π. Then we have $\beta(\pi',\mathcal{C}) = 0$.*

Proof. Consider the k-dimensional simplex $Simp$ determined by the k non-zero components of π. Since π' has at most the same k non-zero components, it is an element of $Simp$. Consider imaginary lines from π' to each of the vertices of $Simp$. A vertex of $Simp$ is a vector of the form $(0, ..., 0, 1, 0, ..., 0)$, i.e., one component is 1 and all the others are 0. Furthermore, the 1 must be in correspondence of a non-zero component of π. These lines determine a partition of $Simp$ in convex subspaces, and π must belong to one of them. Hence π can be expressed as a convex combination of π' and some vertices of $Simp$, say π_1, ..., π_h. Namely, $\pi = c\pi' + c_1\pi_1 + ... + c_h\pi_h$ for suitable convex coefficients $c, c_1, ..., c_h$. Furthermore, since π has k non-zero components, it is an internal point of $Simp$, and therefore c must be non-zero. Hence, we have:

$$\begin{aligned} 0 &= R^*(\pi,\mathcal{C}) \\ &= R^*(c\pi' + c_1\pi_1 + ... + c_h p_h, \mathcal{C}) \\ &\geq cR^*(\pi',\mathcal{C}) + c_1 R^*(\pi_1,\mathcal{C}) + ... + c_h R^*(\pi_h,\mathcal{C}) \\ &= cR^*(\pi',\mathcal{C})\,, \end{aligned}$$

where the third step comes from the concavity of R^*, and the last one is because $R^*(\pi_j,\mathcal{C}) = 0, \quad \forall j$, since π_j is a vertex. Therefore, since c is not 0, $R^*(\pi',\mathcal{C})$ must be 0. □

We can now prove our result.

Theorem 7. *Let $n = |\mathbb{S}|$, and let υ denote the uniform prior on $\mathbb{S}$. For any prior π with k non-zero components, if $\beta(\pi,\mathcal{C}) = 0$ then*

$$\beta(\upsilon,\mathcal{C}) \leq \frac{1 - k/n}{1 - 1/n}\,.$$

Moreover, there exists a channel $\mathcal{C}$ for which equality is reached.

Proof. Let $m = |\mathbb{O}|$, and let $\mathbb{S}'$ be the set of the non-zero components of π. It is sufficient to note that:

$$\begin{aligned} &\sum_{o\in\mathbb{O}} \mathcal{C}_{s,o}\upsilon(s) \\ &= 1/n \sum_{o\in\mathbb{O}} \max_{s\in S} \mathcal{C}_{s,o} \\ &\geq 1/n\,(\mathcal{C}_{s_1,o_1} + ... + \mathcal{C}_{s_m,o_m}) \quad \text{where } s_i = \arg\max_{s\in S'} \mathcal{C}_{s,o_i} \\ &= k/n\,; \end{aligned}$$

the last equality is due to the fact that, by definition of s_i,

$$\sum_o \max_{s\in S'} \mathcal{C}_{so} = (\mathcal{C}_{s_1,o_1} + ... + \mathcal{C}_{s_m o_m})\,.$$

Therefore, for π' defined as in Lemma 1, $\beta(\pi', C) = 0$ implies $1/k(\mathcal{C}_{s_1,o_1} + ... + \mathcal{C}_{s_m o_m}) = 1$, from which we derive $(\mathcal{C}_{s1o1} + ... + \mathcal{C}_{s_m o_m}) = k$. This proves the first statement.

The second claim of the theorem states the existence of a channel $\mathcal{C}'$ for which equality is reached. We define $\mathcal{C}'$ so that it coincides with $\mathcal{C}$ in the rows corresponding to the non-zero components of π. Define all the other rows identical to the previous ones (it does not matter which ones are chosen). Then:

$$\sum_o \max_{s\in\mathbb{S}} \mathcal{C}'_{s,o} \;=\; \sum_{o\in\mathbb{O}} \max_{s\in\mathbb{S}'} \mathcal{C}'_{s,o} \;=\; \sum_{o\in\mathbb{O}} \max_{s\in\mathbb{S}'} \mathcal{C}_{s,o} \;=\; k\,,$$

therefore proving the second part of the theorem. □

B. Proof of Theorem 1

Let $\mathcal{U}^{(k)}$, for $k = 1, ..., n$, be the set of priors with exactly k non-zero components, and such that the distribution on those components is uniform. In the following we indicate with $\mathcal{U}$ the set $\mathcal{U} = \mathcal{U}^{(1)} \cup \mathcal{U}^{(2)} \cup ... \cup \mathcal{U}^{(n)}$.

We start by recalling the following definition from [8] (Definition 3.2, simplified).

Definition 1. Let S be a subset of a vector space, let $g : S \mapsto \mathcal{R}$, and let $S' \subseteq S$. We say that g is *convexly generated* by S' if for all $v \in S$ there exists $S'' \subseteq S'$ such that there exists a set of convex coefficients $\{c_u\}_{u\in S''}$ (i.e., satisfying $\sum_u c_u = 1$ and $c_u \geq 0\ \forall u \in S''$) such that:

$$1)\ v = \sum_{u\in S''} c_u u\ , \qquad 2)\ g(v) = \sum_{u\in S''} c_u g(u)\,.$$

The following results were also proven in the same reference (Proposition 3.9 in [8]).

Proposition 6. *G is convexly generated by $\mathcal{U}$.*

Proposition 7. *$R^*(\pi,\mathcal{C})$ is concave on π.*

The elements of $\mathcal{U}$ are called *corner points* of G, and the elements of each set $\mathcal{U}^{(k)}$ are the *corner points of order k*.

We now prove that if a function is defined as the ratio of a concave function and a convexly generated one, then its minimum is attained on one of the corner points of the function in the denominator. This will be important to characterize the minimum of the Bayes security metric, which is indeed defined as the ratio of the Bayes risk and the guessing error.

Lemma 2. *Let S be a subset of a vector space. Let $f : S \mapsto \mathcal{R}_{\geq 0}$ be a concave function, and let $g : S \mapsto \mathcal{R}_{\geq 0}$ be a function that is convexly generated by a finite $S' \subseteq S$, and which is positive in at least some of the elements of S. Then there exists $u \in S'$ such that $u = \arg\min_{v:g(v)>0} f(v)/g(v)$.*

Proof. Assume by contradiction that $\exists v \in S$ such that $g(v) > 0$ and $\forall u \in \mathcal{U}$ with $g(u) > 0$

$$\frac{f(v)}{g(v)} < \frac{f(u)}{g(u)}\,. \tag{8}$$

Since g is convexly generated by S', and $g(v) > 0$, $\exists S'' \subseteq S'$ such that $g(v) = \sum_{u\in S''} c_u g(u)$, where c_u are suitable

convex coefficients, and $\forall u \in S''\ g(u) > 0$. Therefore:

$$\begin{aligned}\frac{f(v)}{g(v)} &= \frac{f(v)}{\sum_{u\in S''} c_u g(u)} \\ &\geq \frac{\sum_{u\in S''} c_u f(u)}{\sum_{u\in S''} c_u g(u)} && \text{(by concavity of } f\text{)} \\ &> \frac{\sum_{u\in S''} c_u g(u)\frac{f(v)}{g(v)}}{\sum_{u\in S''} c_u g(u)} && \text{(by Equation 8)} \\ &= \frac{f(v)}{g(v)}\end{aligned}$$

which is impossible. Furthermore, S' is finite, hence $\left\{\frac{f(u)}{g(u)} \mid u \in S', g(v) > 0\right\}$ has a minimum. □

Corollary 3. *The minimum of* $\{\beta(\pi,\mathcal{C}) | \pi \in \mathcal{D}(\mathbb{S}), G(\pi) > 0\}$ *exists, and it is in one of the corner points of* G.

Proof. The statement follows from the definition $\beta(\pi,\mathcal{C}) = \frac{R^*(\pi,\mathcal{C})}{G(\pi)}$, and from Proposition 6, Proposition 7, and Lemma 2. □

Furthermore, note that because G in its corner points of order 1 takes value 0, the corner point of G on which β is minimized must have order $k \geq 2$.

We now can prove Theorem 1. It remains to show that the corner points on which β is minimized have order $k = 2$.

Theorem 1. *Consider a channel* $\mathcal{C}$ *on a secret space with* $|\mathbb{S}| \geq 2$. *There exists a prior vector* $\pi^* \in \mathcal{D}(\mathbb{S})$ *of the form*

$$\pi^* = \{0, ..., 0, 1/2, 0, ..., 0, 1/2, 0, ..., 0\}$$

such that

$$\beta^*(\mathcal{C}) = \beta(\pi^*,\mathcal{C}) = \min_{\pi\in\mathcal{D}(\mathbb{S})} \beta(\pi,\mathcal{C})\,.$$

Proof. We show this result by induction over n, the cardinality of $\mathbb{S}$, where we assume $n \geq 2$.

Base case ($n = 2$). Since $\forall \pi \in \mathcal{U}^{(1)}$ the guessing error is $G(\pi) = 0$, the minimizer has to have order 2.

Inductive step. Assuming we proved the result for n, we prove it for $n+1$. By Corollary 3, it is sufficient to show that:

$$\forall \pi \in \mathcal{U}^{(n+1)}\ \exists \pi' \in \mathcal{U}^{(n)}\ \ \frac{R^*(\pi,\mathcal{C})}{G(\pi)} \geq \frac{R^*(\pi',\mathcal{C})}{G(\pi')}\,. \quad (9)$$

Consider the $(n+1)\times m$ channel matrix $\mathcal{C}$. For each row i, we define p_i as the sum of elements of the row which are the maximum in their column. (Ties are broken arbitrarily.) I.e.,

$$p_i \stackrel{\text{def}}{=} \sum_o \mathcal{C}_{i,o}\, I(\mathcal{C}_{i,o} = \max_s \mathcal{C}_{s,o})\,.$$

where $I(S)$ is the indicator function, i.e., the function that gives 1 if the statement S is true, and 0 otherwise.

Similarly, we define q_i as the sum of elements which are the second maximum in the columns that have maximum in column i. More precisely, let $\mathrm{smax}(A)$ be the function returning the second maximum in a set A; for instance, if $a_1 \geq a_2 \geq a_3 \geq \ldots$, then $\mathrm{smax}(\{a_i\}) = a_2$. Again, ties are broken arbitrarily. Then:

$$q_i \stackrel{\text{def}}{=} \sum_o \mathcal{C}_{j,o}\, I(\mathcal{C}_{i,o} = \max_s \mathcal{C}_{s,o} \text{ and } \mathcal{C}_{j,o} = \mathrm{smax}_s \mathcal{C}_{s,o})\,.$$

Note that the elements that compose q_i are in rows different from i and possibly different from each other.

Without loss of generality, assume that we have:

$$p_{n+1} - q_{n+1} = \min_i (p_i - q_i)\,. \quad (10)$$

We further denote by r_o, for $o = 1, \ldots, k$, the elements of the $(n+1)$-th row that are not the components of p_{n+1}, namely

$$\{r_o \,|\, o = 1, \ldots, k\} \stackrel{\text{def}}{=} \{\mathcal{C}_{n+1,o} \,|\, \mathcal{C}_{n+1,o} \neq \max_s \mathcal{C}_{s,o}\}$$

The following observation is immediate:

Fact 1. *For all* $i \in \{1, ..., n+1\}$, *we have* $q_i \geq r_i$.

We can now prove Equation 9. We will prove it for $\pi' = (1/n, ..., 1/n, 0)$, while $\pi = (1/(n+1), ..., 1/(n+1))$ necessarily.

Observe that:

$$\begin{aligned}R^*(\pi,\mathcal{C}) &= 1 - \frac{1}{n+1}\sum_{i=1}^{n+1} p_i = \frac{n+1-\sum_{i=1}^{n+1} p_i}{n+1} \\ G(\pi) &= \frac{n}{n+1} \\ R^*(\pi',\mathcal{C}) &= 1 - \frac{1}{n}\sum_{i=1}^{n} p_i - q_{n+1} = \frac{n-\sum_{i=1}^{n} p_i - q_{n+1}}{n} \\ G(\pi') &= \frac{n-1}{n}\,.\end{aligned}$$

Therefore, to prove Equation 9 we need to demonstrate that:

$$(n-1)\left(n+1-\sum_{i=1}^{n} p_i - p_{n+1}\right) \geq n\left(n-\sum_{i=1}^{n} p_i - q_{n+1}\right)\,.$$

By simplifying and rearranging:

$$\sum_{i=1}^{n} p_i - n\,p_{n+1} + n\,q_{n+1} + p_{n+1} \geq 1\,.$$

By the assumption in Equation 10, we have:

$$\begin{aligned}&\sum_{i=1}^{n} p_i - n\,p_{n+1} + n\,q_{n+1} + p_{n+1} \\ \geq\ &\sum_{i=1}^{n} p_i - \sum_{i=1}^{n} p_i + \sum_{i=1}^{n} q_i + p_{n+1} \\ =\ &\sum_{i=1}^{n} q_i + p_{n+1} && \text{“Fact 1”} \\ \geq\ &\sum_{i=1}^{n} r_i + p_{n+1} = 1\,. && \text{“}\mathcal{C}\text{ is stochastic”}\end{aligned}$$

□

C. Proofs of Section VII

Theorem 2. *For any channel $\mathcal{C}$, it holds that*

$$\beta^*(\mathcal{C}) \;=\; 1 - \frac{1}{2}\max_{a,b\in\mathbb{S}} \|\mathcal{C}_a - \mathcal{C}_b\|_1 \;=\; 1 - \max_{a,b\in\mathbb{S}} \mathrm{tv}(\mathcal{C}_a, \mathcal{C}_b)\ .$$

Proof. Denote by π_{ab} the prior assigning probability $1/2$ to $a, bin\mathbb{S}, a \neq b$. We show that

$$\beta(\pi_{ab}, \mathcal{C}) = 2 - \sum_o \max_{s\in a,b} \mathcal{C}_{s,o} = 1 - \frac{1}{2}\|\mathcal{C}_a - \mathcal{C}_b\|_1\ , \quad (11)$$

Denote $\mathcal{C}_{\uparrow,o} = \max_{s\in a,b} \mathcal{C}_{s,o}$ and $\mathcal{C}_{\downarrow,o} = \min_{s\in a,b} \mathcal{C}_{s,o}$. The fact that $\beta(\pi_{ab}, \mathcal{C}) = 2 - \sum_o \mathcal{C}_{\uparrow,o}$ comes directly from the definition of β.

Since $\mathcal{C}_a$ and $\mathcal{C}_b$ are probability distributions, it holds that

$$\sum_o (\mathcal{C}_{\uparrow,o} + \mathcal{C}_{\downarrow,o}) = 2$$

Hence, we have that

$$\|\mathcal{C}_a - \mathcal{C}_b\|_1 = \sum_o (\mathcal{C}_{\uparrow,o} - \mathcal{C}_{\downarrow,o}) = 2\sum_o \mathcal{C}_{\uparrow,o} - 2$$

and (11) follows directly. We conclude by Theorem 1. □

We now show that taking convex combinations of vectors cannot increase the diameter of a set, which will be useful for both Proposition 5 and Theorem 4. Denote by $\mathrm{diam}(S)$, $\mathrm{ch}(S)$ the diameter and the convex hull of S respectively.

Lemma 3. *For any $S \subseteq \mathbb{R}^n$, it holds that*

$$\mathrm{diam}(\mathrm{ch}(S)) \;=\; \mathrm{diam}(S),$$

where distances are measured wrt any norm $\|\cdot\|$.

Proof. Let $d = \mathrm{diam}(S)$. Since $S \subseteq \mathrm{ch}(S)$ we clearly have $d \leq \mathrm{diam}(\mathrm{ch}(S))$, the non-trivial part is to show that $d \geq \mathrm{diam}(\mathrm{ch}(S))$.

We first show that

$$\forall a \in S, b \in \mathrm{ch}(S) : \|a - b\| \leq d\ . \quad (12)$$

Let $a \in S, b \in \mathrm{ch}(S)$ and denote by $B_d[a]$ the closed ball of radius d centered at a. The diameter of S is d, hence

$$B_d[a] \supseteq S\ , \qquad \text{and since balls are convex}$$
$$B_d[a] = \mathrm{ch}(B_d[a]) \supseteq \mathrm{ch}(S)\ ,$$

which implies $\|a - b\| \leq d$.

Finally we show that

$$\forall b, b' \in \mathrm{ch}(S) : \|b - b'\| \leq d\ .$$

Let $b, b' \in \mathrm{ch}(S)$, from (12) we know that $B_d[b] \supseteq S$, and since balls are convex we have that $B_d[b] \supseteq \mathrm{ch}(S)$, which implies $\|b - b'\| \leq d$. □

Proposition 5. *Let $\mathcal{C}$ be a channel, $q \in \mathcal{D}(\mathbb{O})$, and $d = \max_{s\in\mathbb{S}} \|\mathcal{C}_s - q\|_1$. Then $1 - d \;\leq\; \beta^*(\mathcal{C})$. Moreover, if $q \in \mathrm{ch}(\mathcal{C}_{\mathbb{S}})$ then $\beta^*(\mathcal{C}) \;\leq\; 1 - d/2$.*

Proof. Let $s, s' \in \mathbb{S}$, from the triangle inequality we have that

$$\|\mathcal{C}_s - \mathcal{C}'_s\|_1 \;\leq\; \|\mathcal{C}_s - q\|_1 + \|\mathcal{C}'_s - q\|_1\ ,$$

hence $\mathrm{diam}(\mathcal{C}_{\mathbb{S}}) \leq 2d$; Theorem 2 implies the lower bound.

Moreover, assume that $q \in \mathrm{ch}(\mathcal{C}_{\mathbb{S}})$. Since $\mathcal{C}_s \in \mathrm{ch}(\mathcal{C}_{\mathbb{S}})$ for all $s \in \mathbb{S}$, it holds that

$$\mathrm{diam}(\mathrm{ch}(\mathcal{C}_{\mathbb{S}})) \;\geq\; \max_{s\in\mathbb{S}} \|\mathcal{C}_s - q\|_1 \;=\; d\ .$$

From Lemma 3 we get that

$$\mathrm{diam}(\mathcal{C}_{\mathbb{S}}) \;=\; \mathrm{diam}(\mathrm{ch}(\mathcal{C}_{\mathbb{S}})) \;\geq\; d\ ,$$

which gives us an upper bound from Theorem 2. □

D. Proofs of Section IV

Theorem 3. *For all channels $\mathcal{C}^1, \mathcal{C}^2$ it holds that*

$$\beta^*(\mathcal{C}^1||\mathcal{C}^2) \;\geq\; \beta^*(\mathcal{C}^1)\cdot\beta^*(\mathcal{C}^2)\,.$$

Proof. Recall that $\pi_{ab} \in \mathcal{D}(\mathbb{S})$ denotes the prior that assigns probability $1/2$ to both $a, b \in \mathbb{S}, a \neq b$. We will use the fact that for such priors, $\beta(\pi_{ab}, \mathcal{C})$ can be written as

$$\beta(\pi_{ab}, \mathcal{C}) \;=\; \sum_{o\in\mathbb{O}} \min_{s\in\{a,b\}} \mathcal{C}_{s,o}\ . \quad (13)$$

This comes from the definition of β and the fact that the rows $\mathcal{C}_a$ and $\mathcal{C}_b$ are probability distributions, hence

$$\Big(\sum_{o\in\mathbb{O}} \min_{s\in\{a,b\}} \mathcal{C}_{s,o}\Big) + \Big(\sum_{o\in\mathbb{O}} \max_{s\in\{a,b\}} \mathcal{C}_{s,o}\Big) = \sum_{o\in\mathbb{O}} \sum_{s\in\{a,b\}} \mathcal{C}_{s,o} = 2\ .$$

We also use the basic fact that for non-negative $\{q_i, r_i\}_i$:

$$(\min_i q_i)\cdot(\min_i r_i) \;\leq\; \min_i (q_i \cdot r_i) \quad (14)$$

The proof proceeds as follows:

$$\begin{aligned}
& \beta^*(\mathcal{C}^1)\beta^*(\mathcal{C}^2) \\
=\;& \Big(\min_{a,b\in\mathbb{S}} \beta(\pi_{ab}, \mathcal{C}^1)\Big)\Big(\min_{a,b\in\mathbb{S}} \beta(\pi_{ab}, \mathcal{C}^2)\Big) && \text{“Theorem 1”} \\
\leq\;& \min_{a,b\in\mathbb{S}} \big(\beta(\pi_{ab}, \mathcal{C}^1)\cdot\beta(\pi_{ab}, \mathcal{C}^2)\big) && \text{“(14)”} \\
=\;& \min_{a,b\in\mathbb{S}} \Big(\sum_{o_1\in\mathbb{O}^1} \min_{s\in\{a,b\}} \mathcal{C}^1_{s,o_1}\Big)\Big(\sum_{o_2\in\mathbb{O}^2} \min_{s\in\{a,b\}} \mathcal{C}^2_{s,o_2}\Big) && \text{“(13)”} \\
=\;& && \text{“Rearranging sums, distributively”} \\
& \min_{a,b\in\mathbb{S}} \sum_{o_1\in\mathbb{O}^1} \sum_{o_2\in\mathbb{O}^2} \Big(\min_{s\in\{a,b\}} \mathcal{C}^1_{s,o_1}\Big)\Big(\min_{s\in\{a,b\}} \mathcal{C}^2_{s,o_2}\Big) \\
\leq\;& \min_{a,b\in\mathbb{S}} \sum_{o_1\in\mathbb{O}^1} \sum_{o_2\in\mathbb{O}^2} \min_{s\in\{a,b\}} \mathcal{C}^1_{s,o_1}\mathcal{C}^2_{s,o_2} && \text{“(14)”} \\
=\;& \min_{a,b\in\mathbb{S}} \sum_{o\in\mathbb{O}^1\times\mathbb{O}^2} \min_{s\in\{a,b\}} (\mathcal{C}^1||\mathcal{C}^2)_{s,o} && \text{“Def. of } \mathcal{C}^1||\mathcal{C}^2\text{”} \\
=\;& \min_{a,b\in\mathbb{S}} \beta(\pi_{ab}, \mathcal{C}^1||\mathcal{C}^2) && \text{“(13)”} \\
=\;& \beta^*(\mathcal{C}^1||\mathcal{C}^2)
\end{aligned}$$

This bound is tight. E.g., $\beta^*(\mathcal{C}||\mathcal{C}) = \beta^*(\mathcal{C})\cdot\beta^*(\mathcal{C})$ for

$$\mathcal{C} = \begin{bmatrix} 0.4 & 0.6 \\ 0 & 1 \end{bmatrix}\ .$$

□

Theorem 4. *For all channels $\mathcal{C}^1, \mathcal{C}^2$ it holds that*

$$\beta^*(\mathcal{C}^1\mathcal{C}^2) \;\geq\; \max\{\beta^*(\mathcal{C}^1), \beta^*(\mathcal{C}^2)\}\,.$$

Proof. The $\beta^*(\mathcal{C}^1\mathcal{C}^2) \geq \beta^*(\mathcal{C}^1)$ part is easy, and comes from the fact that

$$R^*(\pi, \mathcal{C}^1\mathcal{C}^2) \;\geq\; R^*(\pi, \mathcal{C}^1)$$

for all priors π, hence also for the one achieving $\beta^*(\mathcal{C}^1\mathcal{C}^2)$.

The more interesting part is to show that $\beta^*(\mathcal{C}^1\mathcal{C}^2) \geq \beta^*(\mathcal{C}^2)$. The key observation is that the rows of $\mathcal{C}^1\mathcal{C}^2$ are convex combinations of those of $\mathcal{C}^2$:

$$(\mathcal{C}^1\mathcal{C}^2)_{s_1} = \sum_{s_2 \in \mathbb{S}^2} \mathcal{C}^1_{s_1,s_2}\mathcal{C}^2_{s_2} \qquad \text{for all } s_1 \in \mathbb{S}^1$$

Denote by $\mathcal{C}_\mathbb{S} = \{C_s \mid s \in \mathbb{S}\}$ the set of $\mathcal{C}$'s rows, we have:

$$\begin{aligned} (\mathcal{C}^1\mathcal{C}^2)_{\mathbb{S}^1} \;&\subseteq\; \mathrm{ch}(\mathcal{C}^2_{\mathbb{S}^2})\;, && \text{hence} \\ \mathrm{diam}((\mathcal{C}^1\mathcal{C}^2)_{\mathbb{S}^1}) \;&\leq\; \mathrm{diam}(\mathrm{ch}(\mathcal{C}^2_{\mathbb{S}^2}))\;. \end{aligned}$$

Finally, from Lemma 3 we get that

$$\mathrm{diam}((\mathcal{C}^1\mathcal{C}^2)_{\mathbb{S}^1}) \;\leq\; \mathrm{diam}(\mathrm{ch}(\mathcal{C}^2_{\mathbb{S}^2})) \;=\; \mathrm{diam}(\mathcal{C}^2_{\mathbb{S}^2})\;.$$

Theorem 2 concludes, since $\beta^*(\mathcal{C}) = 1 - \frac{1}{2}\mathrm{diam}(\mathcal{C}_\mathbb{S})$. □

E. Proofs of Section V

Theorem 5.

1) *If $\mathcal{C}$ is ε-LDP, then for every π we have $\beta(\pi, \mathcal{C}) \geq \frac{2}{1+\exp(\varepsilon)}$.*

2) *For every $n, m \geq 2$ there exists a $n \times m$ ε-LDP channel $\mathcal{C}^*$ such that $\beta^*(\mathcal{C}^*) = \frac{2}{1+\exp(\varepsilon)}$. Examples of such $\mathcal{C}^*$'s are illustrated in Figure 3.*

Proof. From Theorem 1 we know that for every $\mathcal{C}$ there exists a π^* with a support containing only two secrets, and uniformly distributed on them, such that $\beta^*(\pi^*) = \min_\pi \beta(\pi, \mathcal{C})$. Given $\mathcal{C}$, let us assume, without loss of generality, that the two secrets are s_1 and s_2, and that for each o in the first k columns we have $\mathcal{C}_{s_1,o} \geq \mathcal{C}_{s_2,o}$, and in the last $m-k$ columns we have $\mathcal{C}_{s_1,o} < \mathcal{C}_{s_2,o}$. Then, if we define

$$a \stackrel{\text{def}}{=} \sum_{j=1}^{k} \mathcal{C}_{s_1,o_j} \quad \text{and} \quad b \stackrel{\text{def}}{=} \sum_{j=k+1}^{m} \mathcal{C}_{s_2,o_j}, \tag{15}$$

we have $1-b \leq a$ and $1-a < b$. From the constraints (2), we also know that $a \leq \exp(\varepsilon)(1-b)$ and $b \leq \exp(\varepsilon)(1-a)$. Hence there exists x, y with $1 \leq x \leq \exp(\varepsilon)$ and $1 < y \leq \exp(\varepsilon)$ such that $a = x(1-b)$ and $b = y(1-a)$. The figure below illustrates the situation in the first two rows of the matrix:

$a \;=\; x\,(1-b)$	$1-a$
$1-b$	$b \;=\; y\,(1-a)$

From $a = x(1-b)$ and $b = y(1-a)$ we derive

$$a = \frac{xy-x}{xy-1} \quad \text{and} \quad b = \frac{xy-y}{xy-1}, \tag{16}$$

from which we can compute the Bayes risk of $\mathcal{C}$ in π^*, as a function of x and y:

$$\begin{aligned} f(x,y) \;&\stackrel{\text{def}}{=}\; R^*(\pi^*, \mathcal{C}) && (17)\\ &=\; 1 - \pi^*_{s_1} a - \pi^*_{s_2} b && (18)\\ &=\; 1 - \frac{1}{2}(a+b) && (19)\\ &=\; \frac{\frac{1}{2}(x+y)-1}{xy-1}. && (20) \end{aligned}$$

In order to find the minimum of $f(x,y)$, we compute its partial derivatives:

$$\frac{\partial f}{\partial x}(x,y) = \frac{-\frac{1}{2}y^2 + y - \frac{1}{2}}{(xy-1)^2}\;, \quad \frac{\partial f}{\partial y}(x,y) = \frac{-\frac{1}{2}x^2 + x - \frac{1}{2}}{(xy-1)^2}\;.$$

Since $1 \leq x \leq \exp(\varepsilon)$ and $1 < y \leq \exp(\varepsilon)$, we can easily see that both partial derivatives are negative, hence the minimum value of $R^*(\pi^*, \mathcal{C})$ is obtained for the highest possible values of x and y, namely $x = y = \exp(\epsilon)$. Therefore, using (20):

$$R^*(\pi^*, \mathcal{C}) \;\geq\; \frac{\frac{1}{2}(\exp(\epsilon)+\exp(\epsilon))-1}{\exp(\epsilon)\exp(\epsilon)-1} \;=\; \frac{1}{\exp(\epsilon)-1}. \tag{21}$$

Finally, since $G(\pi^*) = 1/2$, we conclude:

$$\begin{aligned} \beta(\pi, \mathcal{C}) \;&=\; \frac{R^*(\pi, \mathcal{C})}{G(\pi)} \\ &\geq\; \frac{R^*(\pi^*, \mathcal{C})}{G(\pi^*)} && \text{(by Theorem 1)} \\ &\geq\; \frac{2}{1+\exp(\varepsilon)} && \text{(by (21)).} \end{aligned}$$

2 Consider the values a, b defined in (15). Any $n \times m$ matrix $\mathcal{C}^*$ for which these values are

$$a \;=\; b \;=\; \frac{\exp(\epsilon)}{\exp(\epsilon)+1}$$

achieves the above lower bound for β:

$$\beta^*(\mathcal{C}^*) \;=\; \beta(\pi^*, \mathcal{C}^*) \;=\; \frac{2}{1+\exp(\varepsilon)}$$

Hence the bound is a minimum. Figure 3 shows two examples of such matrices. □

F. Generalization using gain/loss functions

We describe here the generalizations of $\mathcal{L}^\times$ and β, parameterized by a *gain function* g (for vulnerability) or a *loss function* ℓ (for risk), discussed in Section V.

Let $\mathbb{W}$ be the set of *guesses* the adversary can make *about* the secret; a natural choice is $\mathbb{W} = \mathbb{S}$, but other choices model a variety of adversaries, (e.g., guessing a part or property of the secret, or making an approximate guess). A *gain* function $g(w, s)$ models the adversary's gain when guessing $w \in \mathbb{W}$ and the actual secret is $s \in \mathbb{S}$. Prior and posterior g-vulnerability [2] are the *expected gain of an optimal guess*:

$$V_g(\pi) = \max_w \textstyle\sum_s \pi_s g(w,s)\;, \quad V_g(\pi, \mathcal{C}) = \textstyle\sum_o p(o) V_g(\delta^o)\;,$$

where δ^o is the posterior distribution on $\mathbb{S}$ produced by the observation o. Then g-leakage expresses how much vulnerability increases due to the channel: $\mathcal{L}^\times_g(\pi, \mathcal{C}) = V_g(\pi, \mathcal{C}) / V_g(\pi)$.

Similarly, we use a *loss* function $\ell(w, s)$, modelling how the adversary's loss in guessing w when the secret is s. Prior and posterior ℓ-risk are the expected loss of an optimal guess:

$$R_\ell(\pi) = \min_w \sum_s \pi_s \ell(w, s) \ , \quad R_\ell(\pi, \mathcal{C}) = \sum_o p(o) R_\ell(\delta^o) \ .$$

Then β_ℓ can be defined by comparing prior and posterior risk: $\beta_\ell(\pi, \mathcal{C}) = R_\ell(\pi, \mathcal{C}) / R_\ell(\pi)$.

Clearly, $\mathcal{L}^\times = \mathcal{L}_g^\times$ for the *identity* gain function, given by $g(w, s) = 1$ iff $w = s$ and 0 otherwise. Similarly, $\beta = \beta_\ell$ for the 0-1 loss function, $\ell(w, s) = 0$ iff $w = s$ and 0 otherwise.

G. Proofs of Section VI

We first prove results on the Bayes security of Gaussian mechanisms, which serves as a template for the security derivation for the Laplace distribution.

Proposition 4. *$\mathcal{G}$ is β^*-secure with $\beta^* = 1 - (\Phi(\alpha) - \Phi(-\alpha))$, where Φ is the CDF of $\mathcal{N}(0, 1)$, and $\alpha = \max_{s_i, s_j} \frac{|s_i - s_j|}{2\sigma}$.*

Before determining the Bayes security of this mechanism, we prove a simple lemma:

Lemma 4 (Total variation of two Gaussians)**.**

$$\mathrm{tv}(\mathcal{N}(\mu_p, \sigma^2), \mathcal{N}(\mu_q, \sigma^2)) = \Phi(\alpha) - \Phi(-\alpha)$$

where Φ is the CDF of $\mathcal{N}(0, 1)$ and $\alpha = \frac{|\mu_p - \mu_q|}{2\sigma}$.

Proof. This proof follows closely the proof of Theorem 3.1 in [27]. Let us first calculate the total variation between the following two Gaussians: $P \sim \mathcal{N}(\mu_p, 1)$ and $Q \sim \mathcal{N}(\mu_q, 1)$. Without any loss of generality, let $\mu_p \leq \mu_q$; hence, $P(x) \geq Q(x)$ for $x \leq \frac{\mu_p + \mu_q}{2}$.

$$\begin{aligned}
\mathrm{tv}(P, Q) &= \int_{-\infty}^{\frac{\mu_p + \mu_q}{2}} P(x) - Q(x) dx \\
&= \int_{-\frac{\mu_p - \mu_q}{2}}^{\frac{\mu_p - \mu_q}{2}} P(x) - Q(x) dx \\
&= P_{A \sim \mathcal{N}(0,1)} \left(A \in \left[-\frac{\mu_p - \mu_q}{2}, \frac{\mu_p - \mu_q}{2} \right] \right) \\
&= \Phi\left(\frac{\mu_p - \mu_q}{2} \right) - \Phi\left(-\frac{\mu_p - \mu_q}{2} \right)
\end{aligned}$$

Now, observe that the total variation is scale-independent. Hence we have: $\mathrm{tv}(\mathcal{N}(\mu_p, \sigma^2), \mathcal{N}(\mu_q, \sigma^2)) = \mathrm{tv}(\mathcal{N}(\mu_p/\sigma, 1), \mathcal{N}(\mu_q/\sigma, 1))$. Applying the result obtained above to the scaled means concludes the proof. □

Proposition 3. *$\mathcal{L}$ is β^*-secure with*

$$\beta^* = \exp\left(- \max_{s_i, s_j \in \mathbb{S}} \frac{|s_i - s_j|}{2\lambda} \right) ,$$

Sketch proof. We first compute the total variation distance between $\Lambda(\mu_p, 1)$ and $\Lambda(\mu_q, 1)$. By applying similar arguments to the one used for Lemma 4, we obtain:

$$\mathrm{tv}(\Lambda(\mu_p, 1), \Lambda(\mu_q, 1)) = F(\alpha) - F(-\alpha) \, ,$$

where F is the CDF of $\mathcal{L}(0, 1)$, and $\alpha = \frac{|\mu_p - \mu_q|}{2}$.

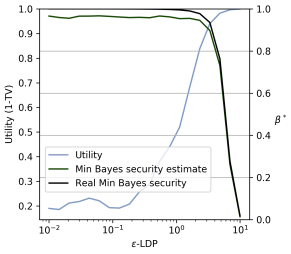


Fig. 7: Randomized Response obfuscation mechanism on the US `Census1990` dataset. Utility (blue line) vs security measured in Min Bayes security (green line) and DP (x-axis).

Via explicit computation of F:

$$\mathrm{tv}(\Lambda(\mu_p, 1), \Lambda(\mu_q, 1)) = 1 - \exp(-\alpha) \, .$$

We apply the above result to compute the total variation between $\Lambda(\mu_p/\lambda, 1)$ and $\Lambda(\mu_q/\lambda, 1)$ (which is equal to the total variation between $\Lambda(\mu_p, \lambda)$ and $\Lambda(\mu_q, \lambda)$), and use the equivalence between total variation and Bayes security metric to conclude the proof. □

H. Empirical evaluation of Randomized Response

Dataset. The US 1990 Census dataset (`Census1990`) comprises 2,458,285 records with a number of attributes for each record. As Murakami and Kawamoto [30], we reduced the attributes to those we judged potentially sensitive: age (8 values), income (5 values), marital status (5 values), and sex (2 values). Overall, the number of values that can be taken by the vectors describing an individual is $8 \times 5 \times 5 \times 2 = 400$. This is the size of both the secret and output spaces for RR.

Methodology. First, we use RR to obfuscate the `Census1990` dataset to guarantee different levels of ε-LDP, and measure the resulting utility by computing the total variation distance between the empirical estimation $\hat{p}$ and the true distribution p.

Second, we compute the Bayes security for RR, both analytically using Equation 7 for different number of secrets $n = |\mathbb{S}|$, and empirically using `fbleau` [10]. Because the Bayes security between any pair of secrets obfuscated by RR is identical, one computation for one arbitrary pair suffices to obtain β^*.

Results. We show in Figure 7 the security of the empirical estimation $\hat{p}$ (Bayes security, empirical and estimation, in green, and DP in the x-axis), and the utility after applying RR to obtain ε-LDP for $\varepsilon \in [0.01, 10]$. We observe that utility is low for values $\varepsilon < 2$. Concretely, utility reaches 95% for $\varepsilon = 3.3$. While this is weak protection in terms of differential privacy, we obtain $\beta^* = 0.96$ `fbleau` estimate ($\beta^* = 0.99999$ from Equation 7), which means that the adversary's probability of success is small. Even for $\varepsilon = 4.8$, which yields utility of 99%, β^* is above 0.9 ($\beta^* = 0.99995$ analytic value).

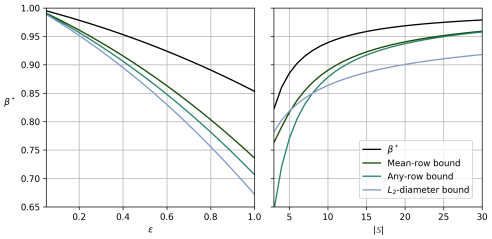


Fig. 8: β^* and bounds for Rand. Response. Left: $|\mathbb{S}| = |\mathbb{O}| = 10$ while varying ε. Right: $\varepsilon = 0.5$ while varying $|\mathbb{S}| = |\mathbb{O}|$.

I. Bayes security approximation via Proposition 5

In Figure 8, we show β^* and various lower bounds for the Randomized Response mechanism (RR) (Section VI). Two of the bounds are obtained via Proposition 5, by setting q to be the mean row (the uniform distribution for RR) and any row of $\mathcal{C}$ (all rows are equal for RR). The third bound is obtained via the L_2-diameter by using the trivial embedding. The mean-row bound is the most accurate in this case.