PAC Privacy and Black-Box Privatization

Hanshen Xiao[1] and Srinivas Devadas[2]

[1]Department of Computer Science, Purdue University / NVIDIA, hsxiao@purdue.edu
[2]CSAIL, MIT, devadas@mit.edu

1 Abstract

This article discusses a new privacy definition called *Probably Approximately Correct (PAC) Privacy*, which offers a formal probabilistic characterization of privacy risk aligned with general privacy concerns. PAC Privacy enables automated, universal and provable privatization for any (possibly black-box) computation. While PAC Privacy shows promise both conceptually and operationally, further research is needed to evaluate its theoretical and practical advantages, as well as its potential synergies with other privacy frameworks.

2 Introduction

Any processing of sensitive data carries the risk of leakage. In practice, leakage manifests in various forms, including physical characteristics, such as timing, power consumption, and cache behaviors in cryptographic implementations [1]; or releases, such as aggregated census data (e.g., average salary or age) [2], AI models trained on sensitive data [3] or their inference results [4]. However, mitigating information leakage often comes at a cost—privacy protection typically necessitates trading off the efficiency of information propagation against obscuring correlations with sensitive features. Thus, a fundamental challenge in privacy research is designing the most efficient leakage control mechanisms, ensuring that, given the observed leakage, an adversary cannot easily recover the underlying secret while minimizing performance or utility degradation. To achieve an optimal utility-privacy tradeoff, a crucial step is to rigorously define privacy and quantitatively assess privacy risk.

2.1 Defining Privacy through Belief Changes

The first formal definition of privacy was established by Shannon in 1949 [5], known as *perfect secrecy*. To illustrate, consider a leakage or processing function $\mathcal{F}(\cdot)$, where the input X represents a secret containing sensitive features to be protected, and the output $\mathcal{F}(X)$ represents the corresponding leakage during or after processing X. Perfect secrecy requires that the leakage $\mathcal{F}(X)$ is *statistically* independent of the secret X. For example, if $\mathcal{F}(\cdot)$ always returns a constant, this provides no useful information to assist adversary's inference on X.

Although perfect secrecy is prohibitively expensive, which effectively prevents any information propagation, it establishes a fundamental semantic notion of privacy: regardless of an adversary's prior belief about the secret X, observing $\mathcal{F}(X)$ does not alter their posterior belief. For instance, if X is a six-digit passcode of a mobile phone and an adversary initially believes $X = 123456$, then under perfect secrecy after observing $\mathcal{F}(X)$, the adversary's belief remains unchanged. Mathematically, this implies that for any assumed prior distribution of X, the posterior distribution of X conditional on $\mathcal{F}(X)$ is identical to the prior.

This principle—limiting changes in adversarial belief—forms the foundation of many subsequent works, including modern cryptography [6] and Differential Privacy (DP) [7]. Notably, from Goldwasser and Micali's pioneering work on probabilistic encryption in 1982, this concept has also been equivalently framed as a challenge for an adversary to distinguish the leakage produced by two *arbitrary* secret candidates $\bar{X}$ and $\bar{X}'$, namely *Input-Independent Indistinguishability* (III). The exact definition of distinguishability depends on the adversary model and the sensitive features

in X that require protection. For example, in DP, where the goal is to obscure the participation of an individual record in a sensitive dataset X, the III requirement ensures that there does not exist a computationally-unbounded adversary that can distinguish the distributions of $\mathcal{F}(\bar{X})$ and $\mathcal{F}(\bar{X}')$ for *any* adjacent $\bar{X}$ and $\bar{X}'$ differing by a single data point [1]. The distinguishability challenge has been quantified through various divergence metrics, leading to different DP security parameters, including pure ϵ DP, approximate (ϵ, δ) DP and Rényi $(\alpha, \epsilon_\alpha)$ DP; they reduce to perfect indistinguishability with security parameters being zeroes.

It is also worth emphasizing the *necessity* of III, the indistinguishability of two inputs *arbitrarily* differing in the sensitive features, if we expect a *worst-case* bound on the impact of leakage for *arbitrary* adversarial belief. However, in many practical applications, achieving tight III analysis remains a significant challenge.

2.2 Fundamental Challenges in Worst-Case Analysis

We begin with some intuition behind the challenges in deriving III analysis. Consider a scenario where X is a 512-bit secret key used in an encryption program executed on a complex circuit/processor. The leakage function $\mathcal{F}$ captures the timing required to perform the encryption with key X on this circuit. To preserve privacy, we aim to randomize or obfuscate the execution time by introducing noise. Most existing noise mechanisms [2] rely on an upper bound of the sensitivity, defined as the maximal difference between the leakage produced by any two secret selections $\bar{X}$ and $\bar{X}'$ differing in the sensitive features to protect

$$\sup_{\bar{X}, \bar{X}'} \|\mathcal{F}(\bar{X}) - \mathcal{F}(\bar{X}')\|$$

under some metric $\|\cdot\|$. However, without additional knowledge or assumptions about the circuit and program to execute, determining the worst-case sensitivity through exhaustive evaluations of $\mathcal{F}$ over all possible secret key selections is computationally prohibitive requiring exponential time [2]. Currently, tight sensitivity bounds are only available for a limited class of simple mechanisms, such as averaging or linear queries.

To establish provable III guarantees for more complex mechanisms, existing methods—particularly in the DP literature—primarily follow two approaches. The more widely used one is *decompose-then-compose*: a mechanism is (artificially) decomposed into multiple simpler steps with easily bounded sensitivity, and each step's output is perturbed to satisfy III before being passed to the next step. The III guarantees of individual steps are then composed. This approach ensures privacy even if all intermediate results are released. However, it often results in loose privacy bounds, requiring excessive perturbation to achieve a provable guarantee. A notable example is DP Stochastic Gradient Descent (DP-SGD) [3, 9].

An alternative approach is *subsample-then-aggregate* [10], as seen in methods like Private Aggregation of Teacher Ensembles (PATE) [4]. Here, the input data is partitioned into multiple disjoint subsets, and processing is performed separately on each before aggregation. These artificial modifications approximate the original algorithm, using aggregation as a building block to ensure tractable III guarantees. However, such modifications come at a cost, imposing constraints at both the input and algorithmic levels [11].

3 PAC Privacy – Impossibility of Customized Adversarial Inference

Originating from the III framework, Probably Approximately Correct (PAC) [3] Privacy employs a more intuitive probabilistic language to quantify the advantage an adversary gains from leakage when inferring X. Compared to III, PAC Privacy introduces two additional assumptions regarding secret entropy and adversarial knowledge.

a) Secret Entropy: Entropy measures the randomness of a variable. PAC Privacy focuses on scenarios where the secret X is randomly drawn from some distribution D. Unlike an adversarial belief, which is a *subjective* assumption

[1] Compared to DP that focuses on individual privacy, cryptographic applications typically aim to protect the entire input X but against a weaker, computationally-bounded adversary. In such cases, the III requirement is that no polynomial-time algorithm can distinguish the distributions of $\mathcal{F}(\bar{X})$ and $\mathcal{F}(\bar{X}')$ for arbitrary $\bar{X}$ and $\bar{X}'$. In this article, we focus on general statistical protection against computationally-unbounded adversaries.

[2] It is known that computing the sensitivity of a general function $\mathcal{F}$ is NP-hard [8].

[3] The definition of PAC Privacy borrows the idea from the PAC learning theory, which models privacy preservation as an impossible learning task for an adversary.

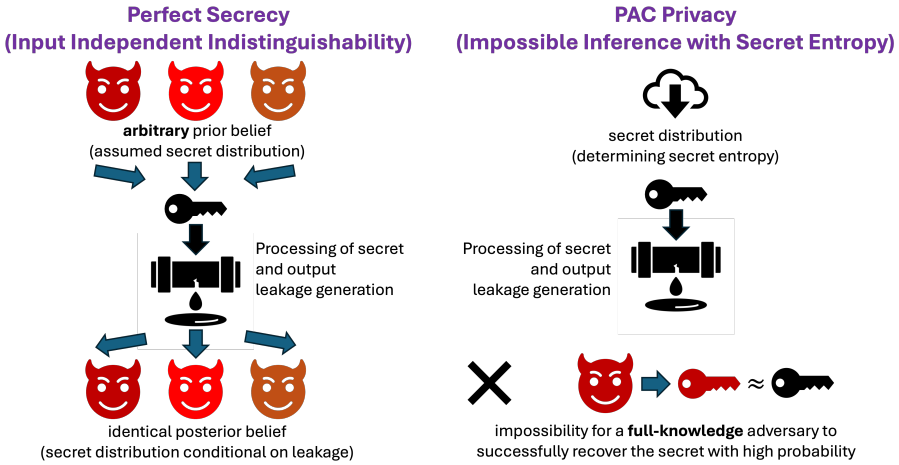

Figure 1: Illustrations of Input-Independent Indistinguishability and PAC Privacy

about X, the entropy of X is an *objective* property determined by the underlying distribution D. In some cases, such as side-channel leakage from cryptographic systems, the distribution D of a secret key X enjoys a closed form, e.g., a uniform distribution. However, in many practical scenarios, e.g., a collection of training image samples for a machine learning task, characterizing D is challenging. For these cases, we may artificially create a data distribution. One commonly-used strategy is subsampling: we may generate our secret data X by subsampling the data pool, as is also adopted in membership inference attacks and DP auditing [12].

b) Full-Knowledge Adversary: Rather than considering arbitrary adversarial beliefs, PAC Privacy assumes a computationally-unbounded adversary with full knowledge of both the secret distribution D and the leakage function $\mathcal{F}$. The only elements unknown to the adversary are the randomness inherent in secret generation and the randomness in the processing function $\mathcal{F}$, which are typically controlled by the user or secret holder.

Under these assumptions, PAC Privacy introduces a criterion ρ, allowing a general expression of privacy concerns based on events where a full-knowledge adversary, after observing the leakage, can reconstruct a satisfactory estimate $\tilde{X}$ such that $\rho(\tilde{X}, X) = 1$. The choice of ρ reflects the level of leakage deemed unacceptable by the secret holder. We list several examples: imagine X is a secret key, and ρ captures a full reconstruction where $\rho(\tilde{X}, X) = 1$ iff $\tilde{X} = X$; or X is a dataset of medical records, and ρ captures membership identification where $\rho(\tilde{X}, X) = 1$ iff $\tilde{X}$ correctly recovers the patient's identity; X can also be a personal record, and ρ captures some approximation where $\rho(\tilde{X}, X) = 1$ iff $\tilde{X}$ estimates the salary attribute within a error margin of $1,000$.

Accordingly, PAC Privacy quantifies risk as the adversary's *best* possible success probability, $(1-\delta_\rho)$, in producing such a satisfactory estimate. The probability $\Pr(\rho(\tilde{X}, X) = 1)$ here accounts for the randomness in both the secret X and the leakage function $\mathcal{F}$. Thus, a PAC Privacy guarantee essentially characterizes an adversarial inference task that is *provably impossible*, as formally defined below.

Definition 1 (**$(\delta_\rho, \rho, \mathsf{D})$ PAC Privacy [13])** *For a leakage/processing function $\mathcal{F}$, some data distribution D, and an inference criterion function $\rho(\cdot, \cdot)$, we say $\mathcal{F}$ satisfies $(\delta_\rho, \rho, \mathsf{D})$-PAC Privacy if the following experiment is impossible:*

A user generates data X from distribution D and sends $\mathcal{F}(X)$ to an informed adversary. A full-knowledge adversary who knows D and $\mathcal{F}$ is asked to return an estimation $\tilde{X}$ on X such that with probability at least $(1 - \delta_\rho)$, $\rho(\tilde{X}, X) = 1$.

We illustrate the comparison between the III framework and PAC Privacy in Fig. 1 and have two important remarks.

1. **III Can (Loosely) Imply PAC Privacy**: As discussed earlier, III provides a global upper bound on the difference between an arbitrary prior and its corresponding posterior belief. By setting the prior belief to that of a full-knowledge adversary, III guarantees can be used to derive an upper bound of PAC Privacy guarantees [14]. However, this reduction can be loose due to two main reasons. First, worst-case leakage—defined independently

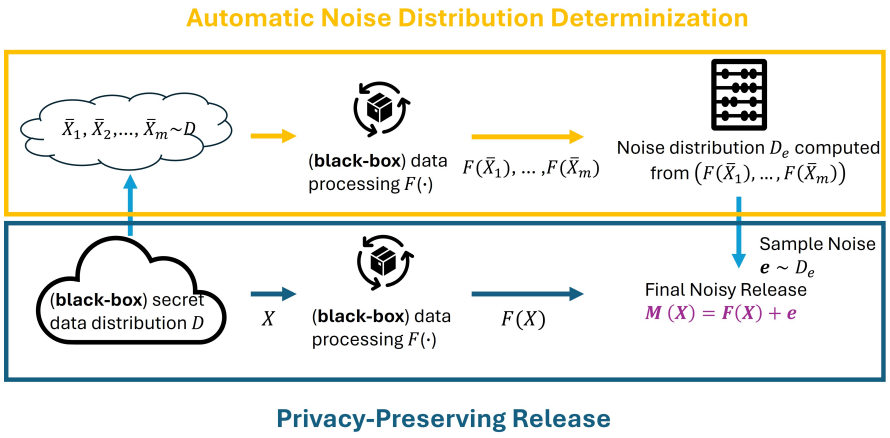

Figure 2: Automated Noise Solution in PAC Privacy

of the secret distribution—can be overly conservative when applied to the actual secret distribution. Second, distinguishability hardness, typically expressed through binary hypothesis testing, may not tightly capture more general adversarial inference tasks (as illustrated by the different ρ selections discussed earlier).

2. **Immunity to Biased Adversaries**: While PAC Privacy explicitly targets a full-knowledge adversary rather than arbitrary adversarial beliefs, its probabilistic guarantees remain valid for adversaries with biased beliefs or incomplete knowledge of the secret distribution D or the leakage function $\mathcal{F}$. For any inference task ρ, those adversaries can never achieve a success rate exceeding that of a full-knowledge adversary.

4 Black-Box Privacy Analysis and Solution

In this section, we provide an overview of black-box PAC Privacy analysis, which automatically determines a perturbation strategy to randomize $\mathcal{F}$ into a satisfactory private version $\mathcal{M}$ with required privacy guarantees. Before proceeding, it is important to clarify what we mean by "black-box". To establish a universal framework for deriving privacy solutions, we aim to minimize algorithmic assumptions about the leakage function $\mathcal{F}$. Specifically, we assume that the secret holder can perform privatization without prior knowledge of the secret distribution D or the leakage function $\mathcal{F}$ but only relying on samples from D and evaluations of $\mathcal{F}$ on these samples to construct the perturbation strategy. However, this black-box restriction applies only to the secret holder; for the full-knowledge adversary, both D and $\mathcal{F}$ are fully known.

With this understanding, the framework can be naturally divided into two key components:

a) **Automated Noise Distribution Determination** — The secret holder samples m instances $\{\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_m\}$ from D and evaluates $\mathcal{F}$ on each instance. An algorithm then takes these evaluations $\{\mathcal{F}(\bar{X}_i) \mid i = 1, 2, \ldots, m\}$, along with the required PAC Privacy parameters (δ_ρ, ρ) and confidence level $(1-\gamma)$, and outputs a noise distribution D_e.

b) **Privacy-Preserving Release** — The secret holder samples the true secret X from D, samples noise e from D_e, and releases the noisy output $\mathcal{M}(X) = \mathcal{F}(X) + e$.

The entire procedure is illustrated in Fig. 2. Intuitively, Part (a) serves as a calibration phase, where the secret holder gains insight into the leakage distribution through empirical observations. After sufficiently many evaluations m, a noise scheme is constructed such that, with high confidence $(1-\gamma)$, the perturbed mechanism $\mathcal{M}(\cdot) = \mathcal{F}(\cdot) + e$ satisfies the required PAC Privacy guarantees. The noisy leakage $\mathcal{M}(X)$ is then used in Part (b) for final release. Technical details on noise determination can be found in Sections 3 and 4 of [13]. Below, we highlight several key insights.

1. **Inference Hardness Becomes Learnable with Proper Perturbation**: One clear role of noise is to control leakage: intuitively, stronger noise increases the challenge of adversarial inference. However, beyond merely limiting leakage, well-structured noise plays a crucial role in smoothing the leakage distribution, thereby enabling a tractable formulation of inference hardness in a black-box setting. While determining the exact (black-box) distribution of $\mathcal{F}(X)$ remains fundamentally infeasible with finite samples—similar to the worst-case analysis challenges in the III framework—an insightful observation is that, under proper perturbation (e.g., Gaussian noise), the adversary's posterior success rate itself can be provably bounded based on empirical evaluations.

2. **Anisotropic (Non-Uniform) Noise**: In many practical applications, the leakage function $\mathcal{F}(X)$ is not uniformly distributed, and the required noise should adapt to its structure. In high-dimensional spaces, it is optimal to introduce noise selectively—adding just enough noise to match the variation of $\mathcal{F}(X)$ along each direction. [13] presents a method for determining the optimal noise under mild assumptions on the covariance spectrum, while a follow-up work [12] proposes a more efficient hybrid approach that approximates the optimal noise distribution in linear time.

3. **Trade-off among Simulation Budget, Confidence, and Noise**: There exists a fundamental trade-off among three factors: the number of simulations m, the confidence level $(1-\gamma)$, and the required noise level. A larger simulation budget enables a tighter noise solution with higher confidence, while a smaller budget necessitates either lower confidence or more perturbation.

4. **Win-Win Between Privacy, Stability, and Algorithmic Co-Design**: One of the key contributions of the black-box analysis is fostering a *win-win* scenario between provable privacy and algorithmic stability. Under PAC Privacy, a more stable leakage function—one exhibiting less variation across different secret inputs in an average sense—is inherently easier to privatize, requiring less noise. Stability is desirable in many practical contexts, contributing to properties such as adversarial robustness and generalization in machine learning. Furthermore, the black-box analysis framework facilitates flexible exploration of algorithmic structures, allowing privacy-preserving designs to be co-optimized with other trustworthy guarantees, such as fairness, backdoor defense, and copyright protection.

5 Future Directions of PAC Privacy

We would like to point out several interesting directions to advance black-box privatization stemming from PAC Privacy. From a theoretical standpoint, for many complicated adversarial inference tasks, existing PAC Privacy results are still conservative upper bounds on the optimal posterior success rate. Narrowing this gap with tight characterization remains an open challenge.

From a noise solution perspective, most existing approaches focus on optimizing and injecting zero-mean (Gaussian) noise. However, in many practical scenarios—especially in side-channel leakage mitigation—only one-sided (positive) noise is feasible. For instance, extending processing time or sending and storing dummy messages/data are typically implementable, whereas modifications in the opposite direction are constrained by hardware limitations, communication protocols, and data management systems. This motivates an interesting generalization: optimizing noise under such constraints.

From an efficiency standpoint, end-to-end evaluation of complex processing functions, such as deep learning model, can be computationally expensive. Treating every procedure as a black box in PAC Privacy analysis may still incur high costs. A promising direction is to explore more efficient privacy analysis in a gray-box manner by decomposing a processing function into multiple black-box components.

Finally, from a win-win perspective, stability in practical processing may not always manifest in an absolute sense, but rather in a distributional or geometric form. For example, Stochastic Gradient Descent (SGD) exhibits distributional stability, while principal component analysis (PCA) demonstrates geometric stability. Efficiently and provably leveraging privacy benefits from such more involved forms of stability remains an open challenge. We refer interested readers to [12, 15] for further intuition and examples.

References

[1] Xiaoxuan Lou, Tianwei Zhang, Jun Jiang, and Yinqian Zhang. A survey of microarchitectural side-channel vulnerabilities, attacks, and defenses in cryptography. *ACM Computing Surveys (CSUR)*, 54(6):1–37, 2021.

[2] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[3] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[4] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2016.

[5] Claude E Shannon. Communication theory of secrecy systems. *The Bell system technical journal*, 28(4):656–715, 1949.

[6] Shafi Goldwasser and Silvio Micali. Probabilistic encryption & how to play mental poker keeping secret all partial information. In *Proceedings of the fourteenth annual ACM symposium on Theory of computing*, pages 365–377, 1982.

[7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[8] Xiaokui Xiao and Yufei Tao. Output perturbation with query relaxation. *Proceedings of the VLDB Endowment*, 1(1):857–869, 2008.

[9] Hanshen Xiao, Jun Wan, and Srinivas Devadas. Geometry of sensitivity: Twice sampling and hybrid clipping in differential privacy with optimal gaussian noise and application to deep learning. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 2636–2650, 2023.

[10] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

[11] Hanshen Xiao, Zihang Xiang, Di Wang, and Srinivas Devadas. A theory to instruct differentially-private learning via clipping bias reduction. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2170–2189. IEEE, 2023.

[12] Mayuri Sridhar, Hanshen Xiao, and Srinivas Devadas. PAC-Private Algorithms. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2025.

[13] Hanshen Xiao and Srinivas Devadas. Pac Privacy: Automatic privacy measurement and control of data processing. In *Annual International Cryptology Conference*, pages 611–644. Springer, 2023.

[14] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pages 635–658. Springer, 2016.

[15] Hanshen Xiao, G. Edward Suh, and Srinivas Devadas. Formal privacy proof of data encoding: The possibility and impossibility of learnable encryption. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1834–1848, 2024.