Curator Attack: When Blackbox Differential Privacy Auditing Loses Its Power

Shiming Wang
Shanghai Jiao Tong University
Shanghai, China
my16wsm@sjtu.edu.cn

Liyao Xiang[†]
Shanghai Jiao Tong University
Shanghai, China
xiangliyao08@sjtu.edu.cn

Bowei Cheng
Shanghai Jiao Tong University
Shanghai, China
chengbowei@sjtu.edu.cn

Zhe Ji
Shanghai Jiao Tong University
Shanghai, China
ji_zhe@sjtu.edu.cn

Tianran Sun
Shanghai Jiao Tong University
Shanghai, China
Seriousss@sjtu.edu.cn

Xinbing Wang
Shanghai Jiao Tong University
Shanghai, China
xwang8@sjtu.edu.cn

Abstract

A surge in data-driven applications enhances everyday life but also raises serious concerns about private information leakage. Hence many privacy auditing tools are emerging for checking if the data sanitization performed meets the privacy standard of the data owner. Blackbox auditing for differential privacy is particularly gaining popularity for its effectiveness and applicability to a wide range of scenarios. Yet, we identified that blackbox auditing is essentially flawed with its setting — small probabilities/densities are ignored due to inaccurate observation. Our argument is based on a solid false positive analysis from a hypothesis testing perspective, which is missed out by prior blackbox auditing tools. This oversight greatly reduces the reliability of these tools, as it allows malicious or incapable data curators to pass the auditing with an overstated privacy guarantee, posing significant risks to data owners. We demonstrate the practical existence of such threats in classical differential privacy mechanisms against four representative blackbox auditors with experimental validations. Our findings aim to reveal the limitations of blackbox auditing tools, empower the data owner with the awareness of risks in using these tools, and encourage the development of more reliable differential privacy auditing methods.

CCS Concepts

• **Security and privacy** → **Privacy-preserving protocols**.

Keywords

Differential privacy, Blackbox auditing

ACM Reference Format:
Shiming Wang, Liyao Xiang[†], Bowei Cheng, Zhe Ji, Tianran Sun, and Xinbing Wang. 2024. Curator Attack: When Blackbox Differential Privacy Auditing Loses Its Power. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3658644.3690367

1 Introduction

The rise of data-driven applications, e.g., healthcare data and software usage analysis, hinges upon the massive scale of data collected from abundant sources. Yet this data usually lacks close inspection and is at risk of severe (personal) information leakage. A notable instance is the potential leakage of sensitive personal data in large language model-based services, as reported by Kim et al. [18]. Since ordinary *data owners* lack powerful means to safeguard their data independently, they entrust their data to a *data curator.* These curators are responsible for sanitizing the data prior to its release or removing private information in a preprocessing step, as outlined in several studies [3, 12, 17, 27]. In particular, if the data is protected by differential privacy (DP) — the golden standard for restricting record-level privacy leakage — the curator is expected to apply DP mechanisms by the privacy level designated by the data owner.

Hereby data owners often raise concerns about whether the sanitized data truly meets the required privacy level, as the curator can be malicious or incapable, failing to deliver the promised privacy and putting the data owners at risk. Hence, a third-party *DP auditor* plays an important role in ascertaining whether the curator properly implants DP mechanisms at a sufficient level of privacy. This motivates a line of auditing solutions including [5–11, 13, 14, 19, 25, 28, 30] for ϵ-DP, as well as (ϵ, δ)-DP [6–9, 19, 20].

Among various auditing solutions, *blackbox auditing* is becoming the dominant approach for its effectiveness and extensive applicability. Blackbox auditing refers to a scheme that seeks the most powerful witness with mere blackbox access to the DP mechanism. However, lacking an internal view of the audited DP mechanism, blackbox auditing can hardly find the optimal witness in practice. It often relies on the output samples of the queries to find a surrogate witness. Such an auditing imposes the least requirement on the data curator as no raw data or the internal workings of the mechanism are revealed, which is particularly desirable when the mechanism runs on proprietary or complex algorithms.

Representative blackbox auditing tools [5, 11, 20] typically make a non-parametric estimation of the privacy level given collected output samples. The sampling-based approach inevitably leads to

[†] Corresponding author.

CCS '24, October 14–18, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3690367

significant error in the estimation of small probabilities or densities. In response, DP-Sniper [11] and MPL [5] proactively constrain the estimation within the large probability (density) regions and ignore small probabilities (densities). Delta-Siege [20] samples a surrogate variable that largely circumvents the small probability events. However, we observe the error in small probability estimation persists and eventually causes blackbox auditing schemes to fail to catch the overclaimed privacy of malicious curators.

In this paper, we re-examine the pitfall of existing blackbox auditing tools. Our analysis does not entirely negate blackbox auditing; rather it systematically identifies the inherent weakness of blackbox auditing, and empowers the data owners with the awareness of the risk of using existing auditing services. Our argument is based on a solid evaluation of these tools through the lens of hypothesis testing: the null hypothesis H_0 is "the data curator's ϵ_c-DP claim is valid," whereas the alternative hypothesis H_1 is "the curator's provided privacy is weaker than claimed." This novel formulation describes our contribution as *auditing the effectiveness of auditing tools*, which should not be confused with the canonical hypothesis testing interpretation of differential privacy [15] that *audits the effectiveness of DP mechanisms*, as elaborated in §3.

High FPs indicate the auditor fails to catch curators with exaggerated privacy claims, posing privacy leakage threats to data owners. By our hypothesis testing, we find that false positives (FPs) widely exist in most blackbox auditing tools against both benchmark and adapted curator mechanisms. We conclude that *false positives are ubiquitous in blackbox auditing because it ignores small probabilities (densities).* As blackbox auditors limit their computation within large probabilities (densities), they can inspect only a segment of the curator's type II error - type I error tradeoff curve on the canonical hypothesis testing interpretation of DP. In the remaining portion beyond the auditors' scrutiny, curators can manipulate their curves at will. This unchecked flexibility permits undetected privacy violations, leading to a surge in FPs.

Beyond FPs, false negatives (FNs) are also spotted in certain blackbox auditors, which suggest the auditor would wrongly refute a valid privacy claim. This leads to mistrust in curators who offer adequate privacy protection. The frequent occurrence of FPs and FNs should raise significant concerns in blackbox auditing, even if the auditors reliably validate some correct DP claims (high true positives). Our findings are summarized in Table 1.

False positives (FPs) in auditing are particularly fatal, as they compromise the privacy of data owners directly. To demonstrate the privacy impact of FPs in real-world applications, we translate them into realistic *curator attacks*. In these attacks, curators aim to secretly divulge the owners' private data, by providing deceptive DP mechanisms to curate the data while evading the auditors' detection. Hereby we introduce three executable conditions to create such attacks, instantiate these conditions by specific benchmark/adapted DP mechanisms, and show how they evade the state-of-the-art blackbox auditing tools including DP-sniper [11], MPL [5], and Delta-siege [20]. We observe that this FP issue of uncaught overstated privacy guarantees also exists for differentially-private stochastic gradient descent (DPSGD) [23] [1].

Table 1: Presence of FPs and FNs in real-world DP mechanisms. Each FP represents a successful curator attack that compromises the data owners' privacy.

Auditors	**DP-Sniper** [11]	**MPL** [5]	**Delta-Siege** [20]	**DPSGD-Audit** [23]
Guarantee	ϵ-DP		(ϵ, δ)-DP	
Audited DP Mechanism	Laplace SVT RAPPOR	Laplace SVT RAPPOR	Laplace Gaussian	DP-SGD
Errors	FP	FP	FP & FN	FP

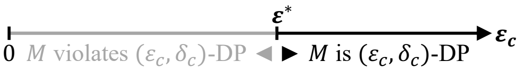

$\boldsymbol{\varepsilon}^*$
$\boldsymbol{\varepsilon_c}$
0
M violates $(\varepsilon_c, \delta_c)$-DP ◀
▶ M is $(\varepsilon_c, \delta_c)$-DP

Figure 1: Illustration of the mechanism's strongest achievable DP guarantee ϵ^* and the its privacy claim ϵ_c.

To sum up, current blackbox auditing for DP is caught in an inherent dilemma between applicability and reliability. As its primary advantage, wide applicability comes at the cost of inevitable false positives. Attempts to reduce FPs may require compromising the blackbox principle thus harming its applicability. We hope to warn data owners of approaching existing blackbox auditors with caution, and to inspire future DP auditors to strive towards resolving the current dilemma.

2 Preliminaries

2.1 Differential Privacy Auditing

Differential Privacy. DP is a principled formulation of privacy-preserving mechanisms to prevent privacy leakage in data analysis. It requires that the probability ratio of the mechanism generating the same output on two arbitrary adjacent datasets is bounded by e^{ϵ}. Formally, a mechanism $M : \mathbb{A} \to \mathbb{B}$ is (ϵ_c, δ_c)-differentially private if for any pair of adjacent datasets $(a, a') \in \mathbb{A}^2$ and all outcome sets $S \subset \mathbb{B}$, $\Pr[M(a) \in S] \leq e^{\epsilon_c} \Pr[M(a') \in S] + \delta_c$.

Definition 1 (True Privacy Level ϵ^*). *Given δ_c, the strongest achievable DP guarantee of a mechanism M is defined by*

$$\epsilon^* := \min\{\epsilon | \forall (a, a', S), \Pr[M(a) \in S] \leq e^{\epsilon} \Pr[M(a') \in S] + \delta_c\}.$$

A smaller ϵ^* indicates stronger privacy. Hence M cannot achieve any (ϵ_c, δ_c)-DP where $\epsilon_c < \epsilon^*$, as illustrated in Figure 1.

DP Auditing. Consider the realistic scenario where a data collector releases a mechanism M that claims to satisfy (ϵ_c, δ_c)-DP. A crucial question arises:

Does M actually afford (ϵ_c, δ_c)-DP as claimed?

A series of auditing tools are devised to validate the privacy claim ϵ_c if possible and discard it where not [5, 10, 11, 13, 20, 28]. Auditing tools with various access assumptions approach this problem by computing the *power*

$$\xi(a, a', S, \delta_c) := |\ln(\Pr[M(a) \in S] - \delta_c) - \ln(\Pr[M(a') \in S])|$$

[1] Our code is available at [2].

of a certain *witness* (a, a', S), given the claimed parameter δ_c. The power suggests to what extent the mechanism output is distinguishable. If an auditing tool finds a witness with power exceeding ϵ_c, the claimed (ϵ_c, δ_c)-DP does not hold. The *maximal power* is obtained on the *theoretical optimal outcome set* S^*, defined as

$$\forall S \subset \mathbb{B},\ \xi(a, a', S, \delta_c) \le \xi(a, a', S^*, \delta_c). \tag{1}$$

The auditing schemes can typically identify the optimal adjacent (a, a') using established pattern heuristics [13]. However, they are not guaranteed to pinpoint S^*, and it is impossible to enumerate all permissible witnesses and check their corresponding power individually. Hence various auditing algorithms [5, 24, 28, 31] are devised to locate a triple $(a, a', \hat{S})$ that approximately maximizes the power, referred to as the auditing tool's *empirical witness.*

Definition 2 (Maximal Power ξ^*). *Given δ_c, an auditing tool's maximal attainable power against a mechanism M is defined by $\xi^* := \xi(a, a', \hat{S}, \delta_c)$.*

The auditing scheme either discards or confirms the privacy claim by comparing ξ^* with ϵ_c. If $\xi^* > \epsilon_c$, the tool asserts that M violates ϵ_c-DP, suggesting either faulty mechanism designs or incorrect code implementations. Otherwise, the privacy claim is deemed valid as no ϵ_c-DP violation is detected.

It further follows from Def 1 that $\epsilon^* = \xi(a, a', S^*)$. Therefore, as marked blue in Figure 1, the ξ^* attained by an auditing tool should always lower bound the theoretical ϵ^*. The goal of auditing is to approach ϵ^* as closely as possible.

2.2 Sampling-Based Blackbox Auditing for DP

Auditing tool's assumption. A range of auditors has access to the curator mechanism's inner structure to facilitate determining the optimal witness. In the absence of the information, sampling-based blackbox auditing tools come into play. They are free to choose the adjacent datasets a and a', and collect the corresponding mechanism's output samples, but they have no knowledge of the mechanism's design or code implementation. Relying solely on mechanism outputs, blackbox auditing is free from most scenario constraints, and is thus adaptable to a wide range of auditing user cases.

User cases of blackbox auditing. To facilitate understanding, we discuss two concrete scenarios from [5]: 1) When skeptical users request third parties to audit a mechanism, but the data collector is reluctant to reveal its proprietary mechanism design and code implementation. Blackbox assumption safeguards intellectual property while enabling effective auditing. 2) When the mechanism under audit is so complex that considering its design is infeasible, for example involving intricate hash functions as in RAPPOR [27] or multiple compositions as in SVT [22]. In these cases, auditing schemes that require additional access are either impossible or overly cumbersome to implement, while blackbox tools remain easy to deploy.

However, due to their limited queries and absence of knowledge about the inner workings of the mechanism M, blackbox auditing schemes face inherent limitations: the distributions of $M(a)$ and $M(a')$ are not analytically known. To compute ξ^*, they have to make nonparametric estimations about the probabilities from collected samples instead, which inevitably induces estimation errors. Below we introduce the common intuitions of representative blackbox auditing schemes and then formalize how they handle this limitation respectively.

Representative blackbox auditing schemes. DP-Sniper [11] and MPL [5] target only the ϵ-DP guarantee, whereas Delta-Siege audit the general (ϵ, δ)-DP mechanisms. We present a sketch of their algorithms in Table 2. First, the auditor queries the targeted mechanism on adjacent datasets multiple times to collect the output samples. Then the auditor computes the likelihood ratio of the output samples:

$$r(b) := \frac{p(b|a)}{p(b|a')}, \tag{2}$$

where $p(b|a)$ and $p(b|a')$ are probability density (or mass) functions for continuous (or discrete) outputs. Third, to achieve high power, the outcome set should pick outputs that are most likely to originate from a rather than a', i.e., b with a large $r(b)$. A simple exercise shows that for ϵ-DP, the theoretical optimal outcome set S^* defined in Eq. (1) consists only of outputs with the largest ratio. A detailed analysis is provided in Appendix A [2]. Formally,

$$S^* = \begin{cases} \{b^* | \forall b \in \mathbb{B}, r(b) \le r(b^*)\} \text{ for } \delta_c = 0; \\ \arg\max\limits_{S} \left\{ \ln \frac{\Pr[M(a) \in S] - \delta_c}{\Pr[M(a') \in S]}, \ln \frac{\Pr[M(a) \in S] - \delta_c}{\Pr[M(a') \in S]} \right\} \text{ for } \delta_c \ne 0. \end{cases} \tag{3}$$

For the relaxed (ϵ, δ)-DP, the parameter δ suggests that the outcome set achieving the largest ξ does not strictly have to be the smallest possible set. However, since the magnitude of δ is typically very low, the optimal S^* most likely remains within the small probability region.

Omitting small probabilities (densities). As the auditor moves toward a smaller S, the probabilities $\Pr[M(a) \in S]$ and $\Pr[M(a') \in S]$ decrease, and the difficulties in assessing these small probabilities emerge. This is because given limited queries, an adversary is almost impossible to observe output samples in the small probability regime. Even if such query outputs are observed, estimating small probabilities from samples introduces large variance to the result, making the estimation highly unreliable.

To avoid errors from unreliable estimation, DP-Sniper proactively ignores small probability outcomes altogether. Specifically, it picks a pre-set threshold c and uses

$$\Pr^{\ge c}[M(A) \in S] := \max\{\Pr[M(A) \in S], c\}$$

in place of $\Pr[M(A) \in S]$. Only large probabilities above c are used as is, thus eliminating the effect of small probabilities when computing ξ^*. The optimal outcome set no longer reduces to a single point, but contains outputs with the largest ratio and satisfies $\Pr[M(a') \in \hat{S}] = c$. To ensure its existence, $\hat{S}$ in DP-Sniper is randomized, i.e., it includes each outcome with a probability $q \in [0, 1]$. The detailed auditing procedure is in Table 2 [2].

MPL [5] uses kernel density estimation to directly estimate the densities per single output point. ξ^* is unstable when the true density is close to 0 because small errors in the density estimation will

[2] We use ratio estimation instead of the original posterior estimation in [11]. This is only a representational adjustment to align with MPL. It does not change the underlying technical procedure. By definition, the posterior is $p(a|b) = p(b|a)/(p(b|a) + p(b|a'))$, assuming the prior is $p(a) = p(a)' = 1/2$. Then $p(a|b) = 1/(1 + 1/r(b))$, and hence the posterior and the ratio are equivalent.

Table 2: Detailed procedure of existing sampling-based blackbox auditing tools. Operations of ignoring small probabilities (DP-Sniper [11]) and ignoring small densities (MPL [5]) are marked yellow.

	Step 1: Collect samples	**Step 2**: Estimate likelihood	**Step 3**: Construct empirical outcome set $\hat{S}$	**Step 4**: Compute ξ^*
MPL [5]	Query mechanism M with a and a' for N times respectively, collect the output samples $b = M(a)$ and $b = M(a')$.	Use KDE to estimate densities $p[b\|a]$ and $p[b\|a']$.	$\hat{S} = \hat{b} = \arg\max_b \left\| \frac{p^{\geq\tau}(b\|a)}{p^{\geq\tau}(b\|a')} \right\|$	$\xi^* = \left\| \ln\left(\frac{p^{\geq\tau}(\hat{S}\|a)}{p^{\geq\tau}(\hat{S}\|a')} \right) \right\|$
DP-Sniper [11]		Train classifer to estimate the likelihood ratio $r(b)$.	$\Pr[b \in \hat{S}] = \begin{cases} 1 & \text{if } r(b) > \hat{t}, \\ q & \text{if } r(b) = \hat{t}. \end{cases}$ (t, q) chosen $s.t.$ $\Pr[M(a') \in \hat{S}] = c$.	$\xi^* = \ln\left(\frac{\Pr^{\geq c}[M(a) \in \hat{S}]}{\Pr^{\geq c}[M(a') \in \hat{S}]} \right)$
Delta-Siege [20]			$\hat{S} = \{b\|r(b) > t\}$. t chosen to minimize $\rho^* = \min_{\epsilon,\delta} \left\{ \rho(\epsilon,\delta) \middle\| \frac{\Pr[M(a) \in \hat{S}] - \delta}{\Pr[M(a') \in \hat{S}]} \geq e^\epsilon \right\}$, where $\rho(\epsilon, \delta)$ is any function non-increasing in both ϵ and δ.	Given claimed δ_c, solve $\rho(\xi^*, \delta_c) = \rho^*$.
DPSGD-Audit [23]	Run one-step DPSGD N times on a and a', compute samples b according to Alg 2 in [23].	Estimate the likelihood ratio $r(b)$.	$\hat{S} = \{b\|r(b) > t\}$. t chosen to maximize $\xi(\hat{S}) = \max\{\ln((\Pr[M(a) \in \hat{S}] - \delta_c)/\Pr[M(a') \in \hat{S}]),$ $\ln(1 - \Pr[M(a') \in \hat{S}] - \delta_c)/(1 - \Pr[M(a) \in \hat{S}])\}$.	$\xi^* = \xi(\hat{S})$.

translate into great errors in the logarithm. Therefore, similar to DP-Sniper, MPL [5] ignores small densities by using a pre-set threshold τ to truncate its estimated density $p^{\geq\tau}(b|A) := \max\{p(b|A), \tau\}$ for any output b and dataset A. It selects its $\hat{S}$ as the individual output that maximizes the truncated likelihood ratio.

Delta-Siege [20] also prioritizes output samples with a large ratio $r(b)$. It does not voluntarily cut off small probabilities, yet it still hardly observes small probability outputs. It further introduces a privacy surrogate $\rho(\epsilon, \delta)$, which can be any non-increasing function in both ϵ and δ. All (ϵ, δ) guarantees sharing the same ρ value are considered equivalent, and a smaller ρ signifies weaker privacy. The auditor then varies its ratio threshold t to obtain multiple pairs $\Pr[M(a) \in \hat{S}]$ and $\Pr[M(a') \in \hat{S}]$. Each probability pair pinpoints a set of infeasible (ϵ, δ)-DP guarantees, from which the auditor selects the particular (ϵ, δ) with the smallest ρ. Finally, the auditor chooses the smallest ρ across all probability pairs as its ultimate privacy evaluation, and computes the distinguishability ξ^* at the specified δ_c accordingly.

DPSGD-Audit is mostly similar to DP-Sniper. The only differences are that the collected samples are DPSGD outputs instead of statistical query outputs, and that its search for the optimal outcome set S^* involves non-zero δ_c as in Eq. (3).

To sum up, with the truncated probabilities, the auditing tool's maximal power in the final step is unified by

$$\xi^* = \begin{cases} \ln(\Pr^{\geq c}[M(a) \in \hat{S}] / \Pr^{\geq c}[M(a') \in \hat{S}]), \text{ DP-Sniper;} \\ |\ln(p^{\geq\tau}(\hat{S}|a) / p^{\geq\tau}(\hat{S}|a'))|, \text{ MPL;} \\ \epsilon \text{ s.t. } \rho(\epsilon, \delta_c) = \rho^*, \text{ Delta-Siege;} \\ \max\left\{ \ln\frac{\Pr[M(a) \in \hat{S}] - \delta_c}{\Pr[M(a') \in \hat{S}]}, \ln\frac{1 - \Pr[M(a') \in \hat{S}] - \delta_c}{1 - \Pr[M(a) \in \hat{S}]} \right\}, \text{DPSGD-Audit.} \end{cases} \quad (4)$$

3 Threat Model and Formulation

Curator attack. In data analysis scenarios, the data owner holds the raw data while the curator analyzes it and releases the outcomes. To comply with privacy regulations, the data is typically sanitized in one of two means: either the data owner applies DP mechanisms provided by the curator and sends the sanitized data to the curator for analysis, or the curator collects the raw data and applies DP algorithms themselves. In both cases, the curator is responsible for devising a sanitization mechanism that meets the required privacy level. Meanwhile, the data owner relies on blackbox auditors to inspect if the provided mechanism truly achieves its privacy claim.

We assume a malicious curator who runs DP mechanisms not conforming to its privacy claim, while escaping the auditors' detection by exploiting the loopholes in the auditing tool. It directly compromises the privacy of data owners, hence termed a *curator attack*. Clearly, the attack is only feasible due to the inadequacy of the auditor. Therefore, it is essential to get an in-depth understanding of the current auditing schemes and be aware of the conditions under which they fail.

In the upcoming sections, we will explore the criteria for effective auditing through the lens of hypothesis testing. We will show that prior evaluations of blackbox auditors are incomplete and fail to reveal significant vulnerabilities within the tools, namely the false positive (FP) errors in hypothesis testing. Each FP constitutes a successful curator attack, as elaborated later in this section. This oversight occurs because these existing tools wrongly consider ignoring small probabilities harmless, as demonstrated against some classical well-constructed and flawed DP mechanisms.

We emphasize that this risk is not just theoretical; it is prevalent in real scenarios. Curator attacks almost always happen whenever the data owner resorts to an existing blackbox auditor, as malicious curators are *both tempted and able* to deploy an FP mechanism. Specifically, a curator has two incentives: (1) to divulge the owners' private data; or (2) to reduce privacy protection in exchange for better data utility. It is able to do so by using an existing FP or, if none exists, by tweaking the mechanism to craft a new FP, as detailed in §4. Either way, the curator is able to deploy a false positive mechanism that fulfills the attack. We then instantiate the construction of the attack with concrete real-life mechanisms and auditors in §5-8. This real-world implication indicates that

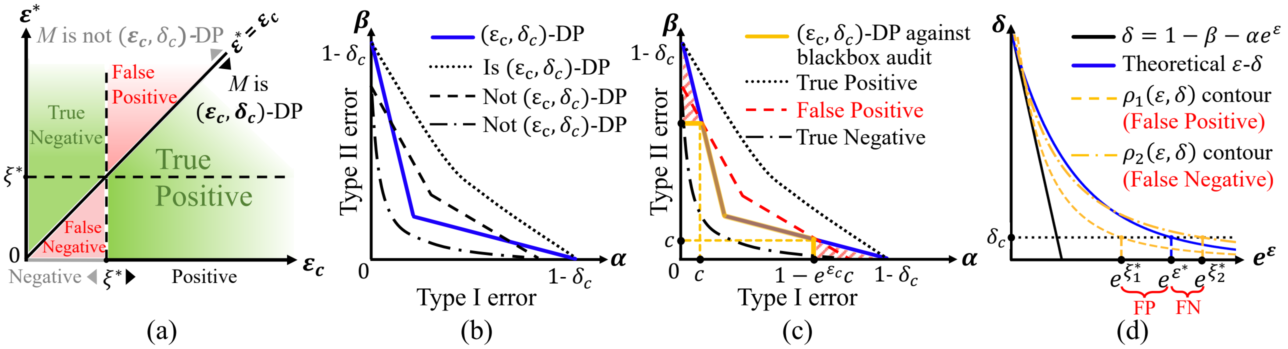


Figure 2: (a) An illustration of the FP, FN, TP, and TN regions divided by the true privacy level ϵ^*, the claimed level ϵ_c, and maximal power ξ^*. (b)(c) DP vs. DP against blackbox audit. FPs exist since the audit ignores small probabilities (densities). (d) Delta-Siege uses a privacy surrogate ρ to seek the optimal power ξ^*. Different ρs lead to FPs or FNs depending on the position with the theoretical $\epsilon - \delta$ DP curve. The illustrative privacy surrogates are $\rho_1(\epsilon, \delta) = e^{-3\epsilon}/\delta$ and $\rho_2(\epsilon, \delta) = e^{-2\epsilon}/\delta$.

the efficacy of blackbox auditing is questionable and should be reassessed.

A hypothesis test. We formally state our problem of auditing the blackbox auditing tool as a hypothesis testing problem $\mathcal{P}_1$ where the auditing target is mechanism M with privacy claim ϵ_c:

$$H_0: \ M \text{ is } (\epsilon_c, \delta_c)\text{-DP}.$$
$$H_1: \ M \text{ violates } (\epsilon_c, \delta_c)\text{-DP}.$$

Confirming or discarding the privacy claim based on ξ^* in §2.1 corresponds to retaining or rejecting the null hypothesis H_0, i.e., if $\xi^* \leq \epsilon_c$, we retain H_0 which is the positive case; otherwise, we reject H_0 which is the negative case. The auditing performance consists of four outcomes: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). We illustrate their definition and practical interpretations in Figure 2(a).

Each mechanism M under scrutiny corresponds to a distinct point (ϵ_c, ϵ^*) in the two-dimensional plane of Figure 2(a). It is also associated with a unique auditing power ξ^* against a designated auditing tool. Following from Eq. (1), whether M is in fact ϵ_c-DP hinges on the comparison between ϵ^* and ϵ_c, and is thus separated by the solid diagonal line in Figure 2(a). Meanwhile, the audit's assertion of "Positive" or "Negative" rests upon ξ^* versus ϵ_c, hence the vertical dashed line. This division segments the entire plane into four regions, representing TP, TN, FP, and FN respectively.

A fresh look at FP. Prior evaluations typically focus on the TP and TN regions while neglecting the FP analysis: for mechanisms that are correctly implemented, the auditing tools confirm the privacy claim with ξ^* being a nearly tight lower bound of ϵ_c (TP); for mechanisms that are wrongly devised, the auditing tools report a ξ^* that significantly exceeds ϵ_c (TN). FP was mostly overlooked: few addresses the case where a mechanism is wrongly devised ($\epsilon^* > \epsilon_c$) without the auditor catching it ($\xi^* \leq \epsilon_c$). Specifically, DPSGD-Audit [23] mentions the intuition behind the blackbox auditor's limitation with a toy example, but is confined to the DPSGD mechanism and lacks a systematic quantitative analysis; MPL [5] briefly addresses the limitation but only with one special case, which downplays the problem's severity; DP-Sniper [11] and Delta-Siege [20] do not mention it at all.

However, FP analysis is critical because a false positive directly results in a successful curator attack by definition. More importantly, we point out that FPs always exist since blackbox auditing schemes cannot handle small probabilities (densities). We will use the canonical f-DP formulation [15] to explain the point. By [15], DP can be completely characterized by the type I error α and type II error β of the following hypothesis testing $\mathcal{P}_2$:

$$H_0^{\text{DP}}: \ \text{the input dataset is } a'.$$
$$H_1^{\text{DP}}: \ \text{the input dataset is } a.$$

Note that the testing problem $\mathcal{P}_2$ is different from $\mathcal{P}_1$. $\mathcal{P}_1$ assesses whether an auditor adheres to its privacy claim with a focus on evaluating the auditing capability: a more powerful auditing tool should distinguish H_0 from H_1 with a higher accuracy. In contrast, $\mathcal{P}_2$ describes how well a mechanism protects individual data records: being harder to differentiate H_0^{DP} and H_1^{DP} means a higher level of privacy of the mechanism.

As visualized in Figure 2(b), (ϵ_c, δ_c)-DP is parametrized by the following tradeoff of $\mathcal{P}_2$ (the blue solid curve) [15]

$$\inf_{\phi}\{\beta_\phi : \alpha_\phi \leq \alpha\} = \max\{0, 1 - \delta_c - e^{\epsilon_c}\alpha, e^{-\epsilon_c}(1 - \delta_c - \alpha)\}$$

for $\alpha \in [0, 1]$. S represents the rejection region for H_0^{DP}, so the type I error is $\alpha = \Pr[M(a') \in S]$ and the type II error is $\beta = 1 - \Pr[M(a) \in S]$. A mechanism is (ϵ_c, δ_c)-DP only if its entire tradeoff curve lies above the blue line. However, when considering DP-Sniper which ignores probabilities below threshold c, satisfying (ϵ_c, δ_c)-DP against blackbox auditing becomes looser:

$$\forall (a, a', S), \ \Pr^{\geq c}[M(a) \in S] \leq e^{\epsilon_c}\Pr^{\geq c}[M(a') \in S] + \delta_c.$$

This transformation results in a *pseudo tradeoff curve*, represented by the yellow solid line in Figure 2(c). The curve is unaltered on $\alpha \in [c, 1 - \delta_c - e^{\epsilon_c}c]$ while requiring

$$\beta \geq 1 - \delta_c - e^{\epsilon_c}\Pr^{\geq c}[M(a') \in S] = 1 - \delta_c - e^{\epsilon_c}c,$$

if α falls below c, indicated by the horizontal yellow line segment. The symmetric holds true where $\beta \leq c$. This difference in the small probability regime (the red-shaded region in Figure 2(c)) allows an infinite number of mechanisms to pass the audit without actually satisfying (ϵ_c, δ_c)-DP. MPL's auditing [5] follows a similar paradigm, except that it permits DP violations within small densities rather than probabilities.

Delta-Siege is more complicated as it relies on the specific choice of the surrogate function $\rho(\epsilon, \delta)$. For the general case of auditing (ϵ, δ)-DP, we could write the surrogate in Table 2 as

$$\rho^* = \min_{\epsilon,\delta} \left\{ \rho(\epsilon, \delta) \middle| \frac{1-\beta-\delta}{\alpha} = e^{\epsilon} \right\}. \tag{5}$$

The observed samples could lead to different pairs of (α, β) and each (α, β) pair corresponds to a distinct line in the e^{ϵ}-δ plane of Figure 2(d) (marked as the black solid line), which represents the set of feasible (ϵ, δ) values at each (α, β). Each ρ value corresponds to a distinct contour marked as the yellow dashed curve, the shape of which depends on the specific function we choose, and the contour moves strictly towards the upper-right corner with a decreasing ρ. Hence the minimal ρ^* is reached at the contour closest to the upper-right corner while intersecting the black solid line. As an example, for convex ρ contours, this occurs where the contour is tangent to the solid line. The auditing result ξ^* lies in the intersection of the ρ^* contour and $\delta = \delta_c$. Clearly, $\xi^* = \epsilon^*$ only if the chosen ρ contour is identical to the theoretical e^{ϵ}-δ curve of M (shown by blue solid line), or the chosen contour happens to intersect $\delta = \delta_c$ at exactly $\epsilon = \epsilon^*$. Both cases are virtually impossible for the blackbox auditor, as it is unaware of the mechanism details. Thus different choices of ρ would lead to the discrepancies between ξ^* and ϵ^*, which turn out to be the FP and FN cases, as shown in Figure 2(d).

With the above theoretical intuition, we mainly discuss the practical existence of FPs in existing auditing tools in the upcoming sections.

4 Roadmap for False Positive Analysis

We begin with a formal definition of the FP in our hypothesis testing problem $\mathcal{P}_1$:

Definition 3 (False Positives). *A mechanism M with privacy claim (ϵ_c, δ_c) is an FP against an auditing tool if*

$$\xi^* \leq \epsilon_c < \epsilon^*, \tag{6}$$

i.e., M violates (ϵ_c, δ_c)-DP ($\epsilon_c < \epsilon^*$) but passes the auditing ($\xi^* \leq \epsilon_c$).

False negatives are defined similarly, i.e., $\xi^* > \epsilon_c \geq \epsilon^*$. Given the above definition, our ultimate goal is to seek the feasible region of the DP parameter(s) satisfying Eq. (6). There are two possibilities: existing DP mechanisms with the parameters in the above range are naturally FPs, and we call them *benchmark mechanisms in loose audit region*; in the case where such parameters are non-existent, we tweak DP mechanisms to make the feasible region non-empty, and we refer to them as *adapted mechanisms in tight audit region.* We unify the two cases in Alg. 1 where line 1-5 represents the first case and line 6-12 shows the second one. We denote the mechanism parameter(s) as ϑ, (e.g., the scale of the added noise) and write ϵ^* and ξ^* as functions of ϑ. For clarity, we represent the parameter of benchmark and adapted mechanisms as $\vartheta = \theta$ and $\vartheta = \widetilde{\theta}$ respectively. Other notations are summarized in Table 3.

Table 3: Notations used for FP analysis.

Notation	Definition
M_ϑ	Mechanism with parameter ϑ
$\vartheta = \theta$	Parameter of the benchmark mechanism M_θ, e.g. Laplacian noise scale θ of SVT [22].
$\vartheta = \widetilde{\theta}$	Parameter of the adapted mechanism $M_{\widetilde{\theta}}$, e.g. crafted noise scale $\widetilde{\theta}$ of adapted SVT.
S^*	Theoretical optimal outcome set.
$\hat{S}$	Empirical outcome set.
$\epsilon^*(\vartheta), \xi^*(\vartheta)$	ϵ^* and ξ^* as functions of mechanism parameter. $\epsilon^*(\vartheta) = \xi(a, a', S^*)$; $\xi^*(\vartheta) \overset{\text{Eq. (4)}}{=} \xi(a, a', \hat{S})$.

For both the benchmark and the adapted mechanisms, FP exists only if $\xi^*(\vartheta) < \epsilon^*(\vartheta)$ holds as a prerequisite, and such a prerequisite also serves as a guideline in search of adaptations to existing DP mechanisms. Thus we formalize the procedure of seeking/constructing FPs for the auditing tools as follows, with the detailed conditions presented in the upcoming sections:

- (*P1-3*): $\xi^*(\vartheta) < \epsilon^*(\vartheta)$. Prerequisite 1-3 are for DP-Sniper, MPL, and Delta-Siege, respectively. DPSGD-Audit shares prerequisite 1 with DP-Sniper. We compute the mechanism's theoretical S^* following Eq. (3), and identify when the benchmark or adapted mechanism conforms to the prerequisites.
- (*R1*): $\epsilon^*(\vartheta) > \epsilon_c$. We derive $\epsilon^* = \xi(a, a', S^*)$ and solve the inequality for θ or $\widetilde{\theta}$. This determines the parameter that violates the DP claim.
- (*R2*): $\xi^*(\vartheta) \leq \epsilon_c$. We compute the auditing tool's empirical $\hat{S}$ following Table 2, derive ξ^* following Eq. (4) and solve the inequality for θ or $\widetilde{\theta}$. This determines the parameter that passes the auditing.

The final false positive θ or $\widetilde{\theta}$ is the intersection of solution sets for the prerequisite, R1 and R2.

The procedure of discovering FNs is constructed similarly by simply reverting the inequalities:

(*P3'*): $\xi^*(\vartheta) > \epsilon^*(\vartheta)$, (*R3*): $\epsilon^*(\vartheta) \leq \epsilon_c$, (*R4*): $\xi^*(\vartheta) > \epsilon_c$.

By definition 2, ξ^* should never exceed ϵ^*. Hence we point out that FNs occur only from the erroneous ξ^* expression used by the auditor. DP-Sniper, MPL, and DPSGD-Audit are free from FNs as they adhere to the primitive ξ^* definition. Discussion of false negatives is limited to Delta-Siege where improper privacy surrogate is used.

5 False Positives Against DP-Sniper

To start with, DP-Sniper was only evaluated at one particular ϵ_c for each benchmark mechanism in [11]. However, its auditing effectiveness may vary across the spectrum of ϵ_c. There may exist some ϑ s.t. $\xi^*(\vartheta) < \epsilon_c \leq \epsilon^*(\vartheta)$, making the benchmark mechanism itself a false positive. Hence we instantiate the prerequisite as

Prerequisite 1 (P1). *Given any mechanism M_ϑ, the iff condition for $\xi^*(\vartheta) < \epsilon^*(\vartheta)$ against DP-Sniper's auditing with probability threshold c is*

$$\Pr[M_\vartheta(a') \in S^*] < c, \tag{7}$$

where S^* is the theoretical optimal outcome set in Eq. (3). That is, M_ϑ must satisfy Eq. (7) to produce a false positive against DP-Sniper.

Algorithm 1: Sketch for FP construction against blackbox auditing tools.

```
Input: Benchmark mechanism $M_\theta$, user-designated privacy level $\epsilon_c$
Output: Mechanisms marked as FP
Derive $\epsilon^*(\theta)$ of $M_\theta$ following Def 1;
Derive $\xi^*(\theta)$ of $M_\theta$ following Eq. (4);
if $\xi^*(\theta) \leq \epsilon_c < \epsilon^*(\theta)$ then
    mark $M_\theta$ as FP;
    /* If the auditing is loose, the benchmark
       mechanism is an FP. */
end
else if $\xi^*(\theta) = \epsilon_c = \epsilon^*(\theta)$ then
    adapt $\theta$ into $\widetilde{\theta}$ following Prerequisite 1 or 2;
    /* If the auditing against benchmark
       mechanism is tight, we adapt it to craft
       an FP. */
    derive $\epsilon^*(\widetilde{\theta})$ of $M_{\widetilde{\theta}}$ following Def 1;
    derive $\xi^*(\widetilde{\theta})$ of $M_{\widetilde{\theta}}$ following Eq. (4);
    solve $\widetilde{\theta}$ s.t. $\xi^*(\widetilde{\theta}) \leq \epsilon_c < \epsilon^*(\widetilde{\theta})$;
    mark $M_{\widetilde{\theta}}$ as FP.
end
```

We reserve the detailed proof for Appendix B [2] and here provide the intuition. Recall that DP-Sniper's empirical witness $\hat{S}$ prioritizes outputs with a high ratio and is determined by $\Pr[M_\vartheta(a') \in \hat{S}] = c$. The set S^* comprises only those outputs with the highest likelihood ratio, that is $\{b^* | \forall b \in \mathbb{B}, r(b) \leq r(b^*)\}$. Therefore, when $\Pr[M_\vartheta(a') \in S^*] > c$, the empirical $\hat{S}$ contains a subset of S^* exclusively, resulting in $\xi^*(\vartheta) = \epsilon^*(\vartheta)$ and eliminating the possibility of false positives. However, if $\Pr[M_\vartheta(a') \in S^*] < c$, to bridge the probability gap between c and $\Pr[M_\vartheta(a') \in S^*]$, the empirical $\hat{S}$ is forced to include additional outputs besides S^*. With a lower ratio, these additional outputs dilute the power on S^*, leading to $\xi^*(\vartheta) < \epsilon^*(\vartheta)$. Only in this case can we solve for ϑ to satisfy the FP condition in Eq. (6).

We now follow Alg. 1 and leverage Prerequisite 1 to construct FP on specific examples. We only consider $c \in [0, 1/2)$ for DP-Sniper; otherwise, c is no longer a small probability. For ease of illustration, we introduce each benchmark mechanism and its adapted counterpart in one pseudocode: adaptations are indicated with strike-though annotations. The example of RAPPOR is moved to Appendix E [2] due to space limit.

5.1 Example 1: Laplace Mechanism

Benchmark Laplace mechanism $M_\theta^{\mathbf{lap}}$ claims to be ϵ_c-DP and operates as in Alg. 2, where $\mathrm{Lap}(\Delta/\theta)$ denotes the Laplace distribution with density

$$p(v|\theta, \Delta) = \frac{\theta}{2\Delta} e^{-\frac{\theta}{\Delta}|v|}.$$

Without loss of generality, let the query output be $q(a) = 0$ and $q(a') = \mu \in [0, \Delta]$.

Algorithm 2: Benchmark/adapted Laplace mechanism

```
Input: Dataset $A$, query $q$, sensitivity $\Delta$, parameter $\theta/\widetilde{\theta}$
Output: $b$
~~$v \sim \mathrm{Lap}(\Delta/\theta)$~~   $v \sim \widetilde{\mathrm{Lap}}(\Delta/\widetilde{\theta})$;
$b = q(A) + v$;
```

Theorem 1. *The benchmark Laplace mechanism M_θ^{lap} is an FP against DP-Sniper's auditing with probability threshold c iff its privacy claim ϵ_c and parameter θ satisfy*

$$\begin{cases} (P1)\ \ \theta > -\ln(2c), & (8a) \\ (R1)\ \ \theta > \epsilon_c, & (8b) \\ (R2)\ \theta \leq -\ln(4c - 4c^2 e^{\epsilon_c}). & (8c) \end{cases}$$

Proof. We show how Eq. (8) corresponds to P1, R1, and R2. Intuitively, the optimal adjacent (a, a') for auditing is obtained when $q(a) = 0$ and $q(a') = \Delta$ as it maximizes the difference between output distributions of a and a'.
(P1): We determine the parameter θ that conforms to prerequisite 1. As illustrated in Figure 3, the likelihood ratio $r(b)$ is non-increasing in b, and the theoretical optimal outcome set S^* with the largest ratio is $S^* = (-\infty, 0]$. Hence $\Pr[M_\theta(a') \in S^*] = e^{-\theta}/2$ and prerequisite 1 becomes Eq. (8a).
(R1): We identify the parameter θ that satisfies $\epsilon^*(\theta) > \epsilon_c$. Following from $S^* = (-\infty, 0]$, the theoretical privacy level of M_θ^{lap} is $\epsilon^*(\theta) = \xi(a, a', S^*) = \theta$. Hence R1 becomes $\theta > \epsilon_c$ as in Eq. (8b).
(R2): To calculate $\xi^*(\theta)$, we observe that DP-Sniper's empirical $\hat{S}$ selects the outputs b in a descending order of $r(b)$ until the probability $\Pr[M_\theta^{\mathrm{lap}}(a') \in \hat{S}]$ reaches c. The ratio is non-increasing in the output, so $\hat{S}$ is a one-sided interval of b; and since $c > \Pr[M_\theta^{\mathrm{lap}}(a') \in (-\infty, 0]]$ according to P1, the empirical $\hat{S}$ is $\Pr[b \in \hat{S}] = \mathbb{1}\left[b \in (-\infty, (\frac{\ln(2c)}{\theta} + 1) \cdot \Delta]\right]$, as shown in Figure 3(a). Therefore, following from Eq. (4), $\xi^*(\theta) = \ln(\Pr^{\geq c}[M_\theta^{\mathrm{lap}}(a) \in \hat{S}] / \Pr^{\geq c}[M_\theta^{\mathrm{lap}}(a') \in \hat{S}]) = \ln(1 - e^{-\theta}/(4c)) - \ln(c)$, and R2 becomes Eq. (8c). □

Alternatively, when $\theta \leq -\ln(2c)$, we have $\Pr[b \in S^*] > c$ and DP-Sniper's $\hat{S}$ contains the optimal S^* exclusively, as illustrated in Figure 3(a). Then $\epsilon^*(\theta) = \xi^*(\theta)$ and the benchmark Laplace mechanism no longer yield false positives. We then adapt the noise distribution to craft a new false positive as below.

Adapted Laplace mechanism $M_{\widetilde{\theta}}^{\mathbf{lap}}$. Following prerequisite 1, a viable adaptation should generate an outcome set where the ratio is high and the probability of a' is low. Then a natural adaptation is to use a bounded noise distribution. It creates outcomes where a' has zero density while a maintains a non-zero density, thereby satisfying the intuition of P1. We denote such a bounded noise distribution as $\widetilde{\mathrm{Lap}}(\Delta/\widetilde{\theta})$, with hyperparameter $\widetilde{\theta} = (\widetilde{\theta_1}, \widetilde{\theta_2})$ and density

$$p(v|\widetilde{\theta}, \Delta) = \begin{cases} \frac{\widetilde{\theta_1}}{2\Delta} e^{-\frac{\widetilde{\theta_1}}{\Delta}|v|}, & |v| \leq \widetilde{\theta_2}, \\ \frac{\widetilde{\theta_1}}{2\Delta} e^{-\frac{\widetilde{\theta_1}\widetilde{\theta_2}}{\Delta}}, & \widetilde{\theta_2} \leq |v| \leq \widetilde{\theta_2} + \frac{\Delta}{\widetilde{\theta_1}}. \end{cases}$$

Figure 3: DP-Sniper's $\hat{S}$ against the benchmark Laplace $M_{\theta}^{\mathbf{lap}}$. The theoretical optimal set is $S^* = (-\infty, 0]$.

This particular noise distribution is not exclusive to crafting an FP. It is selected primarily for the ease of computation.

Theorem 2. *The adapted Laplace mechanism $M_{\widetilde{\theta}}^{lap}$ is an FP against DP-Sniper auditing with probability threshold c, if its privacy claim ϵ_c and parameter $\widetilde{\theta} = (\widetilde{\theta_1}, \widetilde{\theta_2})$ satisfy*

$$(R2)\ \ 1 \leq \widetilde{\theta_1} < \epsilon_c \ \ \& \ \ \frac{1}{2} e^{-\frac{\widetilde{\theta_1}\widetilde{\theta_2}}{\Delta} + \widetilde{\theta_1} - 1} \leq (e^{\epsilon_c} - e^{\widetilde{\theta_1}})c, \tag{9a}$$

$$or\ \ (R2)\ \ \widetilde{\theta_1} < \min\{\epsilon_c, 1\} \ \ \& \ \ \frac{\widetilde{\theta_1}}{2} e^{-\frac{\widetilde{\theta_1}\widetilde{\theta_2}}{\Delta}} \leq (e^{\epsilon_c} - e^{\widetilde{\theta_1}})c. \tag{9b}$$

Please refer to Appendix C [2] for the proof.

5.2 Example 2: SVT

Benchmark SVT mechanism $M_{\theta}^{\mathbf{svt}}$. The SVT mechanism [22] is described in Alg. 3. Compared to the previous Laplace mechanism, SVT is representative of analysis because 1) the outputs are discrete-valued vectors, and 2) the output probabilities are more complicated due to the algorithm's two-step composition. We replicate the settings from [11] where the abort threshold is set as $\bar{t} = 1$, and the query thresholds are $T = 0.5$.

Specifically, Alg. 3 aborts until it outputs a symbol $\top$ or exhausts all N queries, so the possible outputs are $b^j := [\top],\ j = 0; b^j := [\bot^j, \top],\ j = 1, \cdots, N-1$; and $b^j := [\bot^N], j = N$. As an example, the corresponding output probability of $b^j (j \in [1, N-1])$ is $\Pr[M^{\mathrm{svt}}(A) = b^j] = \int_{-\infty}^{\infty} \Pr[\rho = z] \cdot \prod_{i \in [1, j-1]} \Pr[q_i(A) + \nu_i < T_i + z] \cdot \Pr[q_j(A) + \nu_j \geq T_j + z] \cdot \mathrm{d}z$ for $A = a$ or a'. With the integration and product involved, SVT's output distribution is not straightforward as in §5.1.

Theorem 3. *The benchmark SVT mechanism M_{θ}^{svt} is an FP against DP-Sniper's auditing with probability threshold c if its privacy claim ϵ_c and parameter θ satisfy Eq.* (10)*:*

$$\begin{cases} (P1)\ \ \frac{2}{3} e^{-\frac{\theta}{4}} - \frac{1}{6} e^{-\frac{\theta}{2}} < c, & (10a) \\ (R1)\ \ \ln\left(2 + \frac{1}{3} e^{-\theta/2} - \frac{4}{3} e^{-\theta/4}\right) > \epsilon_c, & (10b) \\ (R2)\ \ \ln\left(\frac{c}{2}\left(1 + \frac{c + \frac{1}{6} e^{-\theta/2} - \frac{2}{3} e^{-\theta/4}}{1 + \frac{1}{6} e^{-\theta/2} - \frac{2}{3} e^{-\theta/4}}\right)\right) \leq \epsilon_c. & (10c) \end{cases}$$

The proof is similar to Thm. 1 and is deferred to Appendix D [2].

Algorithm 3: Benchmark/adapted SVT mechanism

```
Input: Dataset $A$, query set $Q$, sensitivity $\Delta$, parameter $\theta/\widetilde{\theta}$,
       thresholds $\mathbf{T} = T_1, \cdots T_N$, abort threshold $\bar{t}$
Output: $b$
~~$\theta_1 = \theta/2, \rho \sim \mathrm{Lap}(\Delta/\theta_1)$~~ $\rho \sim \mathrm{Uniform}(-2\widetilde{\theta_1}, -\widetilde{\theta_1})$;
count = 0;
for each query $q_i \in Q$ do
    ~~$\theta_2 = \theta/2, \nu_i \sim \mathrm{Lap}(2\bar{t}\Delta/\theta_2)$~~ $\nu_i \sim \mathrm{Lap}(\bar{t}\Delta/\widetilde{\theta_2})$;
    if $q_i(A) + \nu_i \geq T_i + \rho$ then
        Output $b_i = \top$;
        count=count+1, Abort if count$\geq \bar{t}$;
    end
    else
        Output $b_i = \bot$;
    end
end
```

Adapted SVT mechanism $M_{\widetilde{\theta}}^{\mathbf{svt}}$. Upon inspecting the expression for $\Pr[M^{\mathrm{svt}}(A) = b^j]$, we can see that simplifying $\Pr[\rho = z]$ facilitates the probability calculation. Hence we modify the noise ρ to follow a uniform distribution. To align with prerequisite 1, we further assign a negative value to ρ so that $\Pr[q_i(A) + \nu_i < T_i + z]$ can attain small values, thereby allowing the mechanism's output probability to be small as well. W.l.o.g., in Alg. 3, we choose noise ρ and ν to follow $\mathrm{Uniform}(-2\widetilde{\theta_1}, -\widetilde{\theta_1})$ and $\mathrm{Lap}(1/\widetilde{\theta_2})$ respectively. These distributions are merely selected for ease of probability calculation. Hence we have

Theorem 4. *The adapted SVT mechanism $M_{\widetilde{\theta}}^{svt}$ is an FP against DP-Sniper's auditing with probability threshold c, if its privacy claim ϵ_c and parameter $\widetilde{\theta} = (\widetilde{\theta_1}, \widetilde{\theta_2})$ satisfy*

$$\begin{cases} (P1)\ \ f(\widetilde{\theta_1}, \widetilde{\theta_2}) < c, & (11a) \\ (R1)\ \ \widetilde{\theta_2} > \epsilon_c, & (11b) \\ (R2)\ \ e^{\widetilde{\theta_2}} f(\widetilde{\theta_1}, \widetilde{\theta_2}) + q \cdot (1 - e^{\widetilde{\theta_2}} f(\widetilde{\theta_1}, \widetilde{\theta_2})) \leq e^{\epsilon_c} c, & (11c) \end{cases}$$

where $f(\widetilde{\theta_1}, \widetilde{\theta_2}) := (e^{-\widetilde{\theta_1}\widetilde{\theta_2}} - e^{-2\widetilde{\theta_1}\widetilde{\theta_2}})/(\widetilde{\theta_1}\widetilde{\theta_2})$ and $q := (c - f(\widetilde{\theta_1}, \widetilde{\theta_2}))/(1 - f(\widetilde{\theta_1}, \widetilde{\theta_2}))$.

The detailed proof is deferred to Appendix D [2].

6 False Positives Against MPL

Different from DP-Sniper, MPL's audit against the benchmark Laplace and SVT has been confirmed to be consistently tight across various ϵ_c regions in [5]. Therefore, these benchmark mechanisms do not have a loose audit region, but can be adapted to others with a loose audit region, as shown in this section.

Prerequisite 2 (P2). *Given an adapted mechanism M_{ϑ} against MPL's auditing with density threshold τ, the iff condition for $\xi^*(\vartheta) < \epsilon^*(\vartheta)$ is*

$$\forall b \in S^*,\ \min\{p(b|a), p(b|a')\} < \tau, \tag{12}$$

where S^* is the theoretical optimal outcome set in Eq. (3)*. Hence M_{ϑ} must first satisfy Eq.* (12) *to produce a false positive against MPL.*

The proof is deferred to Appendix F [2] and the intuition is as follows. If some output b in the theoretical optimal set has both densities $p(b|a)$ and $p(b|a')$ above the density threshold, MPL can precisely estimate these densities, thereby closely approximating the real ϵ^*. False positive arises only when all theoretical optimal outputs conform to Eq. (12), leading to substantial deviations in density estimation. Similar to §5, we now instantiate Alg. 1 and Prerequisite 2 on the Laplace, SVT, and One-time RAPPOR (moved to Appendix G [2]) mechanisms.

6.1 Example 1: Laplace Mechanism

Adapted Laplace mechanism $M^{\mathbf{lap}}_{\widetilde{\theta}}$. Following P2, a straightforward adaptation is to leave the original Laplace distribution unchanged for densities above τ, and adapt the small densities to τ. Formally, the adapted noise distribution is $\widetilde{\mathrm{Lap}}(\Delta/\widetilde{\theta})$, with density

$$p(v|\widetilde{\theta},\Delta) = \begin{cases} \frac{\widetilde{\theta}}{2\Delta} e^{-\frac{\widetilde{\theta}}{\Delta}|v|}, & |v| \le \frac{\Delta}{\widetilde{\theta}} \ln \frac{2\Delta\tau}{\widetilde{\theta}}, \\ \tau, & \frac{\Delta}{\widetilde{\theta}} \ln \frac{2\Delta\tau}{\widetilde{\theta}} \le |v| \le \frac{\Delta}{\widetilde{\theta}} \ln \frac{2\Delta\tau}{\widetilde{\theta}} + \frac{\Delta}{\widetilde{\theta}}. \end{cases}$$

W.l.o.g., let $q(a) = 0$ and $q(a') = \Delta$.

Theorem 5. *The adapted Laplace mechanism $M^{lap}_{\widetilde{\theta}}$ is an FP against MPL auditing with density threshold τ, iff its privacy claim ϵ_c and parameter θ satisfy*

$$(R2)\quad \widetilde{\theta} \le \epsilon_c. \tag{13}$$

Proof. (P2): The theoretical optimal set S^* of this adapted mechanism is where either $p(b|a) = 0$ or $p(b|a') = 0$ and the other is not 0, with a likelihood ratio of ∞. Hence prerequisite 2 is naturally satisfied for any $\widetilde{\theta}$ value.
(R1): The theoretical privacy level of $M^{\mathrm{lap}}_{\widetilde{\theta}}$ is $\epsilon^*(\widetilde{\theta}) = \infty$ for any $\widetilde{\theta}$, so R1 always holds as well.
(R2): The empirical outcome set is now $\hat{S} = \{0\}$. Hence $\xi^*(\widetilde{\theta}) = |\ln(p^{\ge\tau}(0|a)) - \ln(p^{\ge\tau}(0|a'))| = \widetilde{\theta}$, if substituting $q(a) = 0$ and $q(a') = \Delta$ into $\xi^*(\widetilde{\theta})$, we turn the condition of R2 into Eq. (13). □

6.2 Example 2: SVT

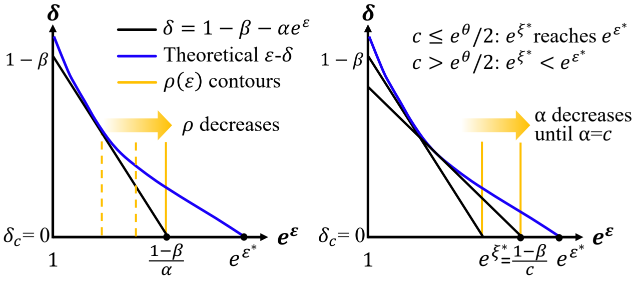
Adapted SVT mechanism $M^{\mathbf{svt}}_{\widetilde{\theta}}$. Following prerequisite 2, the proper adaptation should be able to generate outcomes with small probabilities. The intuition is similar to that in Sec. 5.2, so we leverage the same noise adaptation.

Theorem 6. *The adapted SVT mechanism $M^{svt}_{\widetilde{\theta}}$ is an FP against MPL auditing with probability threshold τ, if its privacy claim ϵ_c and parameter $\widetilde{\theta} = (\widetilde{\theta}_1, \widetilde{\theta}_2)$ satisfy*

$$\begin{cases} (P1)\quad f(\widetilde{\theta}_1, \widetilde{\theta}_2) < \tau, & (14a) \\ (R1)\quad \widetilde{\theta}_2 > \epsilon_c, & (14b) \\ (R2)\quad e^{\widetilde{\theta}_2} f(\widetilde{\theta}_1, \widetilde{\theta}_2) \le e^{\epsilon_c}\tau, & (14c) \end{cases}$$

where $f(\widetilde{\theta}_1, \widetilde{\theta}_2) := (e^{-\widetilde{\theta}_1\widetilde{\theta}_2} - e^{-2\widetilde{\theta}_1\widetilde{\theta}_2})/(\widetilde{\theta}_1\widetilde{\theta}_2)$.

The proof of the theorem is similar to the proof of Thm. 4 and thus is omitted.



(a) Minimize ρ over (ε, δ) (b) Minimize ρ over (α, β)

Figure 4: Laplace mechanism against Delta-Siege's auditing.

7 False Positives/Negatives Against Delta-Siege

The prerequisite of Delta-Siege differs from DP-Sniper or MPL, as the privacy surrogate $\rho(\epsilon, \delta)$ introduces potential false negatives. A detailed proof is in Appendix H [2].

Prerequisite 3 (P3/P3'). *Given any mechanism M_{ϑ} against Delta-Siege's auditing with privacy function $\rho(\epsilon, \delta)$ and smallest achievable probability c, $\xi^*(\vartheta) \ne \epsilon^*(\vartheta)$ iff the $\rho(\epsilon, \delta)$ contour differs from $\delta = \mathcal{T}_{\vartheta}(\epsilon)$ or $\Pr[M(a') \in S^*] < c$, where $\mathcal{T}_{\vartheta}(\epsilon)$ is the mechanism's theoretical δ-ϵ curve defined as*

$$\mathcal{T}_{\vartheta}(\epsilon) := \min\{\delta | M_{\vartheta} \textit{ satisfies } (\epsilon, \delta)\textit{-DP.}\}$$

and S^* is the theoretical optimal set in Eq. (1).

7.1 Example 1: Laplace Mechanism

Benchmark Laplace mechanism $M^{\mathbf{lap}}_{\theta}$ operates as in Alg. 2. Delta-Siege specifically uses $\rho(\epsilon) = \frac{\Delta}{\epsilon}$, while we point out that any non-increasing $\rho(\epsilon)$ follows the same FP analysis as below.

Theorem 7. *The benchmark Laplace mechanism M^{lap}_{θ} is an FP against Delta-Siege's auditing with smallest achievable probability c and any $\rho(\epsilon)$ that is non-increasing and continuous in ϵ, iff its privacy claim ϵ_c and parameter θ satisfy Eq.* (8a), (8b) *and* (8c).

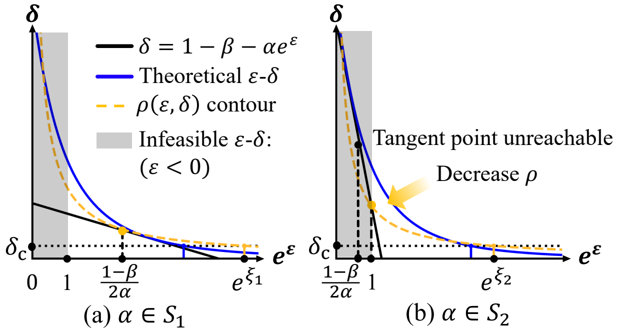
Proof. We reserve the rigorous proof for Appendix H [2] and here visualize the intuition. The theoretical $\delta - e^{\epsilon}$ curve of $M^{\mathrm{lap}}_{\theta}$ is $\delta = \mathcal{T}(\epsilon) = 1 - e^{-\theta/2}\sqrt{e^{\epsilon}}$, and the $\rho(\epsilon)$ contours are vertical lines as shown in Figure 4(a). Given any α and β, the minimal ρ is the contour closest to the upper-right corner while intersecting the line $\delta = 1 - \beta - \alpha e^{\epsilon}$, i.e. when $e^{\epsilon} = (1-\beta)/\alpha$.

(P3): As auditor varies its likelihood threshold, the (α, β) pair changes while conforming to

$$\beta(\alpha) = 1 - e^{\theta}\alpha \text{ if } \alpha < \frac{e^{-\theta}}{2};\ \beta(\alpha) = \frac{e^{-\theta}}{4\alpha} \text{ if } \frac{e^{-\theta}}{2} \le \alpha < \frac{1}{2}, \tag{15}$$

by the definition of tradeoff function [15]. The pair achieves ξ^* when the ρ contour is rightmost, which in turn is when α reaches the minimum c, as in Figure 4(b). Therefore, the prerequisite of $\xi^* < \epsilon^*$ becomes $(1 - \beta(c))/c < e^{\epsilon^*}$, which combined with Eq. (15) is simplified as $c \ge e^{\theta}/2$ (Eq. (8a)).

(R1 & R2): The rest of the proof is omitted as it is identical to DP-Sniper's Theorem 1. □



(a) $\alpha \in S_1$ (b) $\alpha \in S_2$

Figure 5: Gaussian mechanism against Delta-Siege's auditing.

We demonstrate that no FN exists for the benchmark Laplace mechanism. When $c < e^{\theta}/2$, it follows from Eq. (15) that $\forall\alpha \in [c, e^{\theta}/2)$, $(1-\beta)/\alpha = e^{\theta}$, i.e. the e^{ξ^*} point in Figure 4 coincides with the e^{ϵ^*} point. Hence ξ^* never exceeds ϵ^* and FNs do not exist.

7.2 Example 2: Gaussian Mechanism

Benchmark Gaussian Mechanism $M_\theta^{\mathcal{N}}$. The benchmark Gaussian mechanism adds Gaussian noise with standard deviation θ. Below we instantiate the conditions for FP and FN when $\rho(\epsilon,\delta) = e^{-\epsilon}/\delta$. Other choices of ρ follow a similar analysis.

Theorem 8. *Consider a Delta-Siege auditor with* $\rho(\epsilon,\delta) = e^{-\epsilon}/\delta$ *and a smallest achievable probability* c. *Define*

$$\beta(\alpha) := \Phi(\Phi^{-1}(1-\alpha) - \tfrac{\Delta}{\theta}),$$
$$S_1 := \{\alpha | c < \alpha < \tfrac{1-\beta(\alpha)}{2}\} \text{ and } S_2 := \{\alpha | \tfrac{1-\beta(\alpha)}{2} \le \alpha < 1\}.$$

The benchmark gaussian mechanism $M_\theta^{\mathcal{N}}$ *is a false positive (false negative) against such auditor iff* θ *satisfies Eq.* (16a) *(Eq.* (16b)*).*

$$(FP)\quad -\ln(\delta_c\rho^*(\theta)) \overset{R2}{\le} \epsilon_c \overset{R1}{<} \mathcal{T}_\theta^{-1}(\delta_c) \tag{16a}$$

$$(FN)\quad -\ln(\delta_c\rho^*(\theta)) \overset{R4}{>} \epsilon_c \overset{R3}{\ge} \mathcal{T}_\theta^{-1}(\delta_c), \tag{16b}$$

where $\mathcal{T}_\theta^{-1}(\cdot)$ *is the inverse function of* $\mathcal{T}_\theta(\epsilon) = \Phi(-\frac{\epsilon\theta}{\Delta} + \frac{\Delta}{2\theta}) - e^{\epsilon}\Phi(-\frac{\epsilon\theta}{\Delta} - \frac{\Delta}{2\theta})$, *and* $\rho^*(\theta) := \min\{\min_{\alpha\in S_1} \frac{1}{1-\beta(\alpha)-\alpha}, \min_{\alpha\in S_2} \frac{4\alpha}{(1-\beta(\alpha))^2}\}$.

Proof. We first compute $\xi^*(\theta)$ and $\epsilon^*(\theta)$. Then prerequisites (P3/P3') and requirements (R1-4) are directly established. The expression of $\beta(\alpha)$ and $\mathcal{T}_\theta(\epsilon)$ both follow directly from [15].

Given a claimed δ_c, $\epsilon^*(\theta) = \mathcal{T}_\theta^{-1}(\delta_c)$ by definition. The computation of $\xi^*(\theta)$ involves a two-case discussion as illustrated in Figure 5. For any (α,β) pair, the convex ρ contour is tangent to the line $\delta = 1-\beta-\alpha e^{\epsilon}$ on the $\delta - e^{\epsilon}$ plain at $e^{\epsilon} = \frac{1-\beta}{2\alpha}$. This tangent point changes across the $(\alpha, \beta(\alpha))$ pairs and may lie within the infeasible region where $\epsilon < 0$ (the shaded area in Figure 5). Specifically, when $\frac{1-\beta}{2\alpha} > 1$ (i.e. $\alpha \in S_1$), ρ is minimized at the tangent point e^{ϵ}, yielding $\rho = \frac{4\alpha}{(1-\beta)^2}$. When $\frac{1-\beta}{2\alpha} \le 1$ (i.e. $\alpha \in S_2$), the tangent point is no longer valid, and the minimal ρ occurs at $e^{\epsilon} = 1$ instead, resulting in $\rho = \frac{1}{1-\beta-\alpha}$. The minimal ρ^* is found across all $(\alpha, \beta(\alpha))$ pairs. Finally, given $\rho(\epsilon,\delta) = e^{-\epsilon}/\delta$, the auditor's computed power is $\xi^* = -\ln(\delta_c\rho^*(\theta))$. □

8 False Positives in DPSGD-Audit

The one-step DPSGD mechanism selects two adjacent datasets, performs one step of gradient descent, and adds Gaussian noise with standard deviation θ to the gradient. Its FP analysis closely mirrors that of DP-Sniper. The only difference is that with $\delta_c \neq 0$, S^* is calculated using the primitive definition in Eq. (3) instead. Hence we defer the prerequisite and the detailed discussion to Appendix I [2].

Theorem 9. *One-step DPSGD* M_θ^{sgd} *is an FP against blackbox auditor with smallest achievable probability* c *if its privacy claim* (ϵ,δ) *and parameter* θ *satisfy*

$$\ln(1-\delta_c-\beta(c)) - \ln(c) \le \epsilon_c < \mathcal{T}^{-1}(\delta_c),$$

where $\mathcal{T}_\theta^{-1}(\cdot)$ *and* $\beta(\cdot)$ *are as defined in Thm. 8.*

9 Discussion

Constructing a curator attack. We show how to construct a successful curator attack with our previous false positive analysis. As an example, consider a malicious curator who intends to use the SVT mechanism, but pass DP-Sniper's auditing with an overstated ϵ_c-DP claim. To launch this attack, the curator could first solve Eq. (10), select any θ from the solution, and deploy the benchmark SVT mechanism with this θ. If Eq. (10) has no solution, the curator could tweak the mechanism as in Alg. 3, solve Eq. (11) instead, and use any $\tilde{\theta}$ from its solution to deploy the adapted SVT mechanism. The curator could now claim to satisfy ϵ_c-DP without getting caught. Attacks involving other blackbox auditors and mechanisms can be constructed in the same way.

Confidence interval of ξ^***.** When computing ξ^* in our analysis above, we directly use the theoretical $\Pr[M(a) \in S]$ and $\Pr[M(a') \in S]$ for large values. The implicit assumption is that the auditor's estimation of large probabilities above c is accurate. Hence our analytical computation of ξ^* represents the theoretical value of the auditor's empirical power.

In practice, these probabilities are also estimated via randomized sampling, which introduces slight variations across multiple auditing runs due to sampling randomness. Therefore, existing auditors typically report a confidence interval for its power instead of a single value. Specifically, DP-Sniper [1], MPL [5] and Delta-Siege [20] employs a one-sided confidence interval $[\underline{\xi}^*, \infty)$. DPSGD-Audit [23] reports a two-sided interval $[\underline{\xi}^*, \overline{\xi}^*)$ [32].

A privacy claim ϵ_c passes the auditing if and only if it falls within the auditor's confidence interval. Hence for our evaluation in §10, we only need to report the lower bound $\underline{\xi}^*$ for all auditing tools except for DPSGD-Audit.

Dilemma between auditor's reliability & applicability. A natural question following our discussion is the future attempts towards better blackbox auditing. Current tools are unreliable because they fail to examine the small probability region. Hence to eliminate FPs universally, blackbox auditing must inspect the entire type II-type I tradeoff curve of the curator.

One potential move is to "open the blackbox" and leverage the mechanism's inner structure to parametrically estimate the tradeoff, as suggested in [23]. For instance, if the curator uses Gaussian noise, then its tradeoff curve has the parametric form of μ-Gaussian differential privacy [15]. With this knowledge, the auditor could estimate the parameter μ from the collected samples and obtain the entire tradeoff fairly accurately, including the small probability segments, thereby resolving the false positive issue.

However, this approach violates the blackbox auditing assumption, which limits the tool's applicability fundamentally. It cannot audit any curator mechanisms that are proprietary or have intricate tradeoff functions (e.g. the SVT mechanism). Therefore, blackbox auditors are stuck in this dilemma between reliability and applicability for the moment, and perfecting them remains an open problem.

Extending to other auditors. Our analysis of FP and FN can be generalized to future blackbox auditors. We first extend the prerequisites of FP and FN as follows, and the rest of the analysis pipeline is the same as Alg 1.

Step 1: Given any blackbox auditor, we traverse the outcome sets of the curator mechanism and mathematically formulate the auditor's error in estimating the curator's theoretical probability on each outcome set. For current blackbox auditors, this corresponds to concluding the pattern of "omitting small probabilities" in §2.2.

Step 2: With this formulation, given any benchmark DP mechanism, we could explicitly deduce the auditor's observed tradeoff $\mathcal{T}_{audit}$ and identify the "region of discrepancy", where the observed $\mathcal{T}_{audit}$ differs from the curator's theoretical tradeoff $\mathcal{T}_{curator}$. For example, for existing blackbox auditors, this is the small probability region marked in Figure 2 (c)(d).

Step 3: Given a claimed privacy parameter δ_c, we identify the theoretical optimal witness from the theoretical tradeoff (eg. Eq. (3)). Then we tweak the curator mechanism so that the optimal witness falls within the region of discrepancy. For example, in §5.1, we adjust the Laplace mechanism's θ or adapt its noise distribution to $\widetilde{\text{Lap}}(\Delta/\tilde{\theta})$, so that the optimal S^* lies within the small probability region as in Eq. (7). The resulting mechanism will be a false positive or false negative in our analysis.

10 Experiments

In this section, we empirically verify that all identified benchmark and adapted mechanisms in previous sections are indeed false positives (negatives), suggesting possible curator attacks.

10.1 Setup

DP curator mechanisms. We first act as the data curator and implement all benchmark and adapted mechanisms in Python. The DP mechanisms under examination include Laplace, SVT [22], one-time RAPPOR [27], Gaussian, and DPSGD. We move the results on one-time RAPPOR to Appendix J [2].

For each benchmark mechanism M_θ, we traverse the possible values of its parameter θ, derive $\epsilon^*(\theta)$ by definition, and perform corresponding auditing to obtain an empirical $\xi^*(\theta)$. For an intuitive and comprehensive revelation of FPs, each benchmark M_θ is illustrated with a plot of the relationship between $\epsilon^*(\theta)$ and $\xi^*(\theta)$.

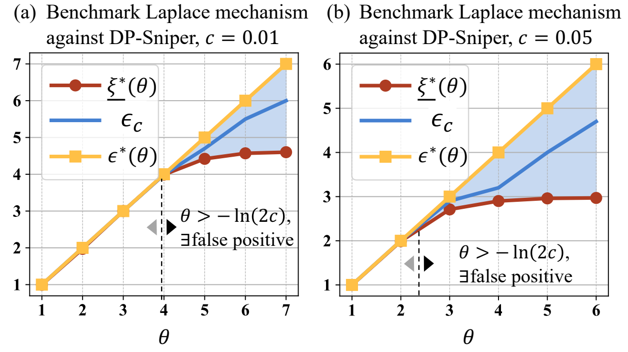


Figure 6: The benchmark Laplace mechanism is an FP against DP-Sniper at $\theta > -\ln(2c)$.

For each adapted mechanism $M_{\widetilde{\theta}}$, we show how the adapted $\widetilde{\theta}$ leads to FPs provided a range of privacy claim ϵ_c.

Auditing tools. For a fair evaluation, we reuse the official code in [11], [5] and [20] to eliminate any potential faulty implementations. For DP-Sniper, MPL, and Delta-Siege, the adjacent datasets are constructed using the `GenerateInputs()` function in [11], which follows the input pattern heuristic as summarized in Table 5. Specifically, the sampling numbers of DP-Sniper and MPL are set according to the instructions in the original work of DP-Sniper and MPL ([11] Theorem 2 and [5] (C3) and (D)). Specifically, the sampling number of DP-Sniper is $10.7 \cdot 10^6$ for $c = 0.01$ and $2.05 \cdot 10^6$ for $c = 0.05$. The sampling number of MPL is $3 \cdot 10^6$. We report the one-sided 0.95 confidence interval of ξ^*. The sampling number of Delta-Siege is 15000, and we report the one-sided 0.9 confidence interval of ξ^*.

For DPSGD-Audit, we evaluate the tasks of image classification and next word prediction on datasets and models as in Table 4. The sampling number is $N = 1000$ for CIFAR10, and $N = 10000$ for SVHN and WikiText. We report the two-sided bayesian interval [32] of ξ^* with siginificance level below 0.03, which is the superior confidence interval for DPSGD auditing so far.

Table 4: Tasks, datasets, models and parameters of the DPSGD-Audit.

DPSGD Task	Dataset	Model	Sampling number N	δ_c
Image classification	CIFAR10 SVHN	ResNet20	1000 10000	10^{-4}
Next word prediction	WikiText-103	T5	10000	10^{-5}

10.2 False Positives against DP-Sniper

Laplace mechanism. We set the probability threshold as $c = 0.01$ or $c = 0.05$. As illustrated in Figure 6, when $\theta > -\ln(2c)$, there is a gap between the mechanism's privacy level and DP-Sniper's empirical power. Therefore, any privacy claim ϵ_c within the blue region makes the benchmark Laplace mechanism a false positive, aligning with Thm. 1. Meanwhile, for $\theta < -\ln(2c)$, the auditing is validated to be tight, reflecting the iff condition.

Table 5: Input patterns adopted from [11] and [5] (using 5-dimensional output as an example).

Pattern	Query $q(a)$	Query $q(a')$
One Above	[1,1,1,1,1]	[2,1,1,1,1]
One Below	[1,1,1,1,1]	[0,1,1,1,1]
One Below Rest Above	[1,1,1,1,1]	[0,2,2,2,2]
Half Half	[1,1,1,1,1]	[0,0,0,2,2]
All Above & All Below	[1,1,1,1,1]	[2,2,2,2,2]
X shape	[1,1,0,0,0]	[0,0,1,1,1]

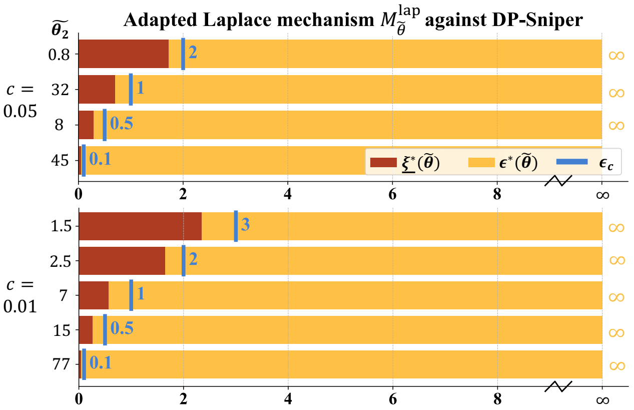


Figure 7: Adapted Laplace mechanism against DP-Sniper's auditing, $c = 0.01$ or 0.05. All adapted mechanisms are FPs.

For the adapted Laplace mechanism, we set $\epsilon_c \in \{0.1, 0.5, 1, 2\}$ for $c = 0.01$ and $c = 0.05$. We first solve $\tilde{\theta}$ following Thm. 2. To satisfy $\tilde{\theta}_1 < \epsilon_c$ in Eq. (9), w.l.o.g., we let $\tilde{\theta}_1 = \epsilon_c/2$. We then arbitrarily choose an $\tilde{\theta}_2$ from the solution set as shown in Figure 7. The theoretical privacy level is $\epsilon^*(\tilde{\theta}) = \infty$ as marked yellow, indicating the adapted Laplace mechanisms do not satisfy any DP. Yet they all pass the auditing as DP-Sniper's power (marked red) is lower than the ϵ_c-DP privacy claim (marked blue), making them false positives.

SVT. For the benchmark SVT mechanism, we set $T = 1$, $N = 1$ and $\bar{t} = 1$. As illustrated in Figure 8, the benchmark SVT behaves similarly to the Laplace mechanism. The curator can choose any privacy claim within the blue region without being detected.

For our adapted SVT mechanism, we set $T = 1$, $N = 10$ and $\bar{t} = 1$. We select a wide range of ϵ_c values, encompassing weak, moderate, and strong privacy claims shown in Figure 9. Notice that the yellow bar is not the precise $\epsilon^*(\tilde{\theta})$ but one of its lower bounds, as introduced in §5.2. Therefore, the actual gap between DP-Sniper's auditing and the adapted mechanism's actual privacy level can be even larger than illustrated. Hence all adapted mechanisms are false positives.

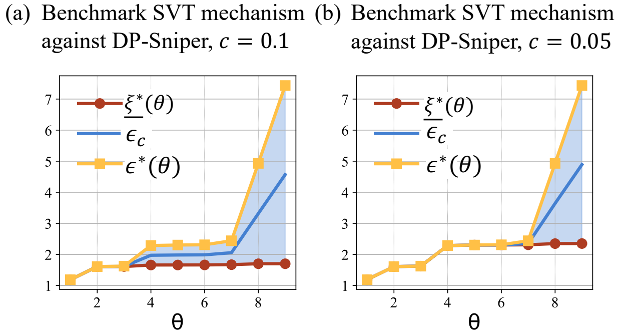


Figure 8: The benchmark SVT against DP-Sniper's auditing.

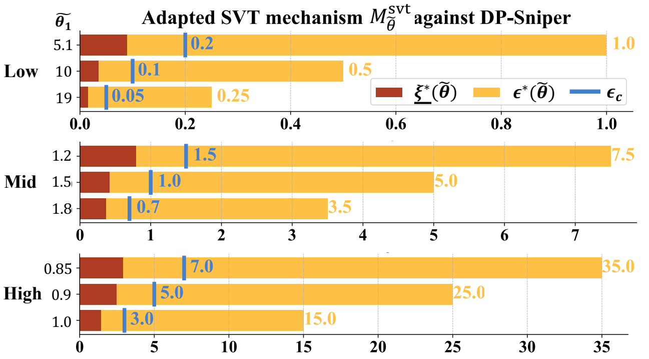


Figure 9: Adapted SVT mechanism against DP-Sniper's auditing, $c = 0.01$.

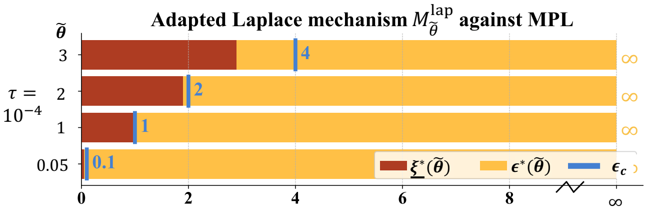


Figure 10: Adapted Laplace mechanism against MPL's auditing, $\tau = 10^{-4}$.

10.3 False Positives against MPL

As discussed in §6, we only need to discuss FP's occurrence in MPL against the adapted Laplace mechanism, the adapted SVT mechanism, and the benchmark RAPPOR mechanism. The conclusions are similar to that of DP-Sniper, and the results are in Figure 10, 11 and Appendix J [2].

10.4 FPs & FNs against Delta-Siege

The auditing results of Delta-Siege are empirically unstable between multiple audit runs. Hence we follow its original work and run five independent audit runs for each θ. The reported $\underline{\xi}^*(\theta)$ is the maximum among all runs. As shown in Table 6, for the same privacy surrogate $\rho(\epsilon, \delta)$, the Gaussian mechanism can be a false positive for some θ and a false negative for others. Hence Delta-Siege's auditing is unreliable under a blackbox setting. Results of the Laplace mechanism are omitted as they are similar to that of DP-Sniper.

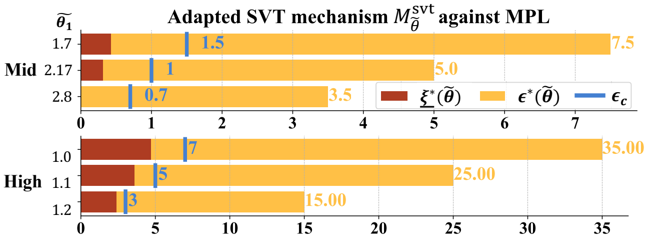


Figure 11: Adapted SVT against MPL's auditing, $\tau = 10^{-4}$. Results in the low parameter regions are omitted because MPL's power ξ^* are almost 0.

Table 6: Gaussian mechanism M_θ^N against Delta-Siege auditing, with a privacy surrogate $\rho(\epsilon, \delta) = 1/(e^\epsilon \delta)$.

$\rho(\epsilon, \delta) = \frac{1}{e^\epsilon \delta}$	Empirical optimal (α, β)	$\epsilon^*(\theta)$	$\xi^*(\theta)$	Error
$\delta_c = 0.005$	$(0.055, 0.921)$	0.30	1.56	FN
	$(0.0006, 0.975)$	3.5	3.9	FN
$\delta_c = 0.05$	$(0.005, 0.675)$	5.1	4.7	FP

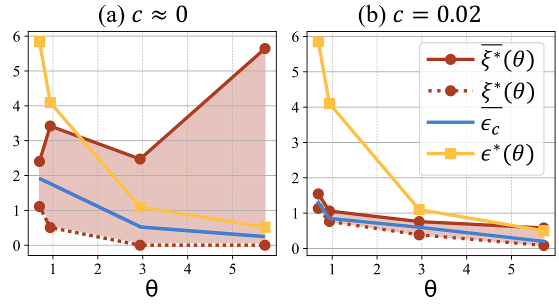


Figure 12: Blackbox DPSGD auditing, two-sided bayesian confidence interval ξ^*. $\delta = 10^{-4}$. Image classification task on CIFAR-10 with ResNet-20.

10.5 False Positives against DP-SGD Audit

We report the two-sided confidence interval $(\underline{\xi^*}, \overline{\xi^*})$ of the DPSGD auditor's power with a significance level below 0.03 [32]. As illustrated in Figure 12, 13 and 14, any ϵ_c claim that falls within the red confidence interval and below the yellow $\xi^*(\theta)$ serves as a false positive.

Selecting extreme threshold values allows the auditor to obtain very small probabilities, i.e. $c \approx 0$. However, as discussed in §2.2, such probability estimations are highly unstable, which leads to a confidence interval too wide to be useful because it allows numerous false positive ϵ_c claims (Figure 12(a), 13(a) and 14(a)). For a comprehensive analysis, we also report the results when the auditor uses moderate thresholds instead, resulting in $c = 0.02$. As shown in Figure 12(b), 13(b) and 14(b), the confidence interval is now significantly narrower and hence more informative. However, the ϵ_c claims are also false positives as the confidence interval is still much lower than $\epsilon^*(\theta)$.

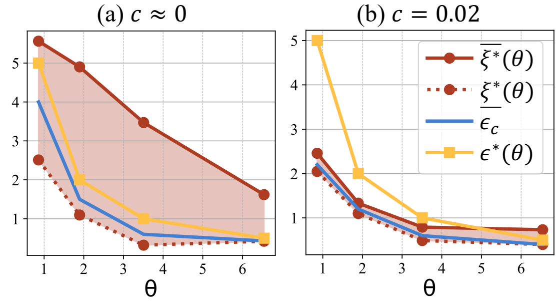


Figure 13: Blackbox DPSGD auditing, two-sided bayesian confidence interval ξ^*. $\delta_c = 10^{-4}$. Image classification task on SVHN with ResNet-20.

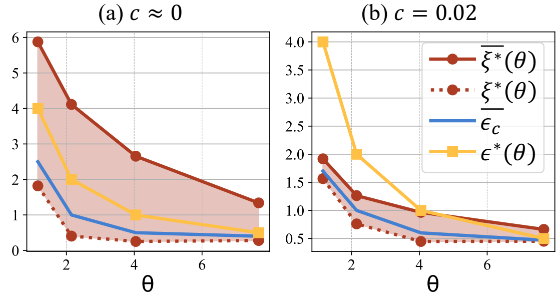


Figure 14: Blackbox DPSGD auditing, two-sided bayesian confidence interval ξ^*. $\delta = 10^{-5}$. Next word prediction task on WikiText-103 with T5.

11 Related Works

We discuss existing approaches to auditing the DP claim of a curator mechanism.

Blackbox auditing for DP mechanisms. Blackbox auditing requires only query access to the curator's mechanism. From the collected output samples of the mechanism, it determines the optimal witness, makes nonparametric estimations on the output probabilities/densities and computes the power accordingly. The primary strength of blackbox auditing lies in its minimal access requirements, thus highly applicable to a vast array of curator mechanisms.

Representative blackbox auditing tools against DP statistical queries include DP-Sniper [11], MPL [5], Delta-Siege [20], etc. DP-Sniper and MPL both leverage a well established heuristic to locate the optimal adjacent dataset pair [13]. Their main focus and contribution lies in determining the optimal outcome set, which was previously intractable to solve. DP-Sniper addresses this challenge by training classifiers to estimate the likelihood ratio and selecting the outputs with the highest ratios as the optimal outcome set. However, this method implicitly induces a parametric assumption on posterior distributions, potentially limiting DP-Sniper's effectiveness if the posterior distribution significantly deviates from the assumed class in the classifier. To improve this, MPL [5] removes this parametric assumption by leveraging the kernel density estimator (KDE), which directly estimates the density function based on interpolation. The performance of MPL [5] and DP-Sniper [11] are close, while MPL [5] is more computationally efficient. Delta-Siege [20] targets the more general (ϵ, δ)-DP and attempts to unify the two parameters with a privacy surrogate $\rho(\epsilon, \delta)$. It then transforms the problem of maximizing the power ξ to minimizing ρ. However, we point out that this method is inherently flawed under the blackbox auditing setting, as blindly choosing ρ without any knowledge of the audited mechanism induces numerous false positives/negatives.

Another line of work focuses on auditing the DPSGD training pipeline [16, 21, 24, 26, 32] and deriving tighter confidence intervals of their estimates. These work typically achieve auditing via launching membership inference attacks under various attacker access. Notice that our blackbox auditing access does not suggest the auditor only has API access to the trained model. Rather, a blackbox auditor can obtain the intermediate gradients of the DPSGD training. Blackbox auditing only means the auditor does not leverage the parametric information of the added noise.

Non-blackbox auditing for DP mechanisms. Compared to the auditors above, non-blackbox auditing [4, 10, 13, 14, 25, 28–30] requires knowledge of the inner structure of the curator's mechanism, such as mechanism design or code implementation. As such, As such, compared to blackbox auditing, they are more suited for mechanism developers rather than general data owners. In addition, they are not applicable to proprietary curator mechanisms and are cumbersome to deploy when dealing with complex mechanisms.

Specifically, a line of these schemes leverage formal verification to perform DP auditing [4, 25, 28, 29]. Their applicability is highly limited by the symbolic solver they use. For example, CheckDP [28] and DiPC [25] uses solvers that are non-compatible with hash functions. Therefore, unlike blackbox auditors, they cannot audit any curator mechanism involving hash functions, for example the RAPPOR mechanism.

Other auditing methods make additional assumptions about the curator mechanism beyond just obtaining output samples, limiting their applicability to specific mechanisms or scenarios that meet their requirements. For instance, DP-Finder [10] requires the mechanisms to have a differentiable power on the witness. Therefore, it does not support any mechanism involving arbitrary loops or hash functions, such as RAPPOR [27]. DP-Stochastic-Tester [30] requires the mechanism's output space to be $\mathbb{R}$, or its confidence interval computation would fail. $\mathcal{T}_{\text{priv}}$ [14] requires oracle access to the specific output probability, which is hardly possible for mechanisms in reality. StatDP [13] needs to run the curator's mechanism without any noise in one of its auditing steps, and is already confirmed to be unsound and unstable in [11]. In contrast, blackbox auditors do not require any such assumptions and are thereby universally applicable.

Another non-blackbox auditor for DPSGD leverages parametric estimation. Unlike blackbox auditors' non-parametric estimations, this auditor [23] knows the functional form of the curator's type II-type I error tradeoff, and applies parametric estimations on the collected samples to determine its power ξ. Compared to blackbox auditing, such auditors are limited to mechanisms with definable parametric tradeoff models, e.g. the Gaussian mechanism. They fail when the tradeoff is hard to express in a closed form, e.g. the SVT mechanism.

12 Conclusion

While blackbox auditing for differential privacy has become popular for its wide applicability, our research reveals a critical flaw: the neglect of small probabilities/densities due to imprecise observation. This gap, highlighted through our false positive analysis, compromises the reliability of these tools, allowing data curators with exaggerated privacy claims to go undetected. Our practical experiments with classical differential privacy mechanisms further underscore the risks posed by such curators. Ultimately, our work sheds light on the vulnerabilities in current blackbox auditing practices, aims to heighten data owners' awareness of these risks, and calls for the advancement of more dependable differential privacy auditing solutions.

Acknowledgments

The research was supported in part by National Science and Technology Major Project 2021ZD0112801, NSF China (62272306, 62032020, 62136006).

References

[1] [n.d.]. DP-Sniper Code. https://github.com/eth-sri/dp-sniper. Accessed: 2023-02-07.

[2] [n.d.]. Our Curator Attack. https://github.com/ShimingWang98/Curator-Attack-When-Blackbox-DP-Auditing-Loses-Its-Power.

[3] John M Abowd. 2018. The US Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2867–2867.

[4] Aws Albarghouthi and Justin Hsu. 2017. Synthesizing coupling proofs of differential privacy. *Proceedings of the ACM on Programming Languages* 2, POPL (2017), 1–30.

[5] Önder Askin, Tim Kutta, and Holger Dette. 2022. Statistical quantification of differential privacy: a local approach. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 402–421.

[6] Gilles Barthe, Rohit Chadha, Vishal Jagannath, A Prasad Sistla, and Mahesh Viswanathan. 2020. Deciding differential privacy for programs with finite inputs and outputs. In *Proceedings of the 35th Annual ACM/IEEE Symposium on Logic in Computer Science*. 141–154.

[7] Gilles Barthe, George Danezis, Benjamin Grégoire, César Kunz, and Santiago Zanella-Beguelin. 2013. Verified computational differential privacy with applications to smart metering. In *2013 IEEE 26th Computer Security Foundations Symposium*. IEEE, 287–301.

[8] Gilles Barthe, Noémie Fong, Marco Gaboardi, Benjamin Grégoire, Justin Hsu, and Pierre-Yves Strub. 2016. Advanced probabilistic couplings for differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 55–67.

[9] Gilles Barthe, Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, César Kunz, and Pierre-Yves Strub. 2014. Proving differential privacy in Hoare logic. In *2014 IEEE 27th Computer Security Foundations Symposium*. IEEE, 411–424.

[10] Benjamin Bichsel, Timon Gehr, Dana Drachsler-Cohen, Petar Tsankov, and Martin Vechev. 2018. Dp-finder: Finding differential privacy violations by sampling and optimization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 508–524.

[11] Benjamin Bichsel, Samuel Steffen, Ilija Bogunovic, and Martin Vechev. 2021. Dp-sniper: Black-box discovery of differential privacy violations using classifiers. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 391–409.

[12] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. 2017. Collecting telemetry data privately. *Advances in Neural Information Processing Systems* 30 (2017).

[13] Zeyu Ding, Yuxin Wang, Guanhong Wang, Danfeng Zhang, and Daniel Kifer. 2018. Detecting violations of differential privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 475–489.

[14] Kashyap Dixit, Madhav Jha, Sofya Raskhodnikova, and Abhradeep Thakurta. 2013. Testing the Lipschitz property over product distributions with applications to data privacy. In *Theory of Cryptography: 10th Theory of Cryptography Conference, TCC 2013, Tokyo, Japan, March 3-6, 2013. Proceedings*. Springer, 418–436.

[15] Jinshuo Dong, Aaron Roth, and Weijie J Su. 2019. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383* (2019).

[16] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems* 33 (2020), 22205–22216.

[17] Noah Johnson, Joseph P Near, and Dawn Song. 2018. Towards practical differential privacy for SQL queries. *Proceedings of the VLDB Endowment* 11, 5 (2018), 526–539.

[18] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2024. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems* 36 (2024).

[19] Xiyang Liu and Sewoong Oh. 2019. Minimax optimal estimation of approximate differential privacy on neighboring databases. *Advances in neural information processing systems* 32 (2019).

[20] Johan Lokna, Anouk Paradis, Dimitar I Dimitrov, and Martin Vechev. 2023. Group and Attack: Auditing Differential Privacy. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. 1905–1918.

[21] Fred Lu, Joseph Munoz, Maya Fuchs, Tyler LeBlond, Elliott Zaresky-Williams, Edward Raff, Francis Ferraro, and Brian Testa. 2022. A general framework for auditing differentially private machine learning. *Advances in Neural Information Processing Systems* 35 (2022), 4165–4176.

[22] Lyu Min, Su Dong, and Li Ninghui. 2016. Understanding the sparse vector technique for differential privacy.

[23] Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. 2023. Tight Auditing of Differentially Private Machine Learning. *arXiv preprint arXiv:2302.07956* (2023).

[24] Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. 2021. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*. IEEE, 866–882.

[25] Vishal Jagannath Ravi. 2019. Automated methods for checking differential privacy. (2019).

[26] Thomas Steinke, Milad Nasr, and Matthew Jagielski. 2024. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems* 36 (2024).

[27] Erlingsson Ulfar, Pihur Wasyl, and Korolova Aleksandra. 2014. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*.

[28] Yuxin Wang, Zeyu Ding, Daniel Kifer, and Danfeng Zhang. 2020. Checkdp: An automated and integrated approach for proving differential privacy or finding precise counterexamples. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. 919–938.

[29] Yuxin Wang, Zeyu Ding, Guanhong Wang, Daniel Kifer, and Danfeng Zhang. 2019. Proving differential privacy with shadow execution. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*. 655–669.

[30] Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. 2020. Differentially private SQL with bounded user contribution. *Proceedings on privacy enhancing technologies* 2020, 2 (2020), 230–250.

[31] Jungang Yang, Liyao Xiang, Ruidong Chen, Weiting Li, and Baochun Li. 2021. Differential privacy for tensor-valued queries. *IEEE Transactions on Information Forensics and Security* 17 (2021), 152–164.

[32] Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Ahmed Salem, Victor Rühle, Andrew Paverd, Mohammad Naseri, Boris Köpf, and Daniel Jones. 2023. Bayesian estimation of differential privacy. In *International Conference on Machine Learning*. PMLR, 40624–40636.