Plausible Deniability Guarantees for Whistleblowers

Anonymous Authors[1]

Abstract

Whistleblowers are a key safeguard against organizational wrongdoing, but the threat of retaliation deters reporting. Existing whistleblower-protection proposals lack formal privacy guarantees, and existing differential privacy mechanisms do not directly target the natural threat model — one in which the audited organization itself observes auditor selection decisions and uses them to identify reporters. We formalize this setting as per-report $(0, \delta)$-differential privacy on the transcript of audit selections under a strong-adversary threat model. Within this framework we prove that the standard approach — randomized response applied at the selection step — must approach uniform random auditing at any fixed $(0, \delta)$ level as the horizon grows. We then give a generic mechanism that reduces private auditing to private continual counting: any $(0, \delta)$-DP continual counter plugs in by post-processing, and the audit transcript inherits the same per-report guarantee. Instantiating the reduction with a recent work in continual counting yields per-report $(0, \delta)$-DP with noise scaling as $O(\sqrt{\log T})$ across a horizon of T audit decisions. A utility theorem shows that the selection error vanishes whenever the noisy report gap between the most-reported organization and the runner-up grows faster than $\sqrt{\log T}$. Simulations show a substantial improvement over randomized response. Code to reproduce all experiments is available in the anonymized supplement.

1. Introduction

Whistleblowers are a crucial safeguard against organizational wrongdoing, from small businesses (Clarke, 2020) to national governments (Scheuerman, 2014), but they frequently face retaliation from the organizations they expose (Mesmer-Magnus and Viswesvaran, 2005). In a 2020 survey, 61% of employees across government, for-profit, and not-for-profit sectors experienced retaliation after reporting misconduct (Initiative, 2020). This threat deters potential whistleblowers (Cassematis and Wortley, 2013) and reduces the effectiveness of those who do, as withholding information can reduce the risk of detection (Powar and Beresford, 2023). Accordingly, anonymity is widely treated as central to effective whistleblower protection (Sharma et al., 2018; Baljija and Min, 2023). However, there is still no widely accepted formal model for protecting whistleblower anonymity (Kutyłowski and Wechta, 2025).

To see why whistleblower anonymity is difficult, consider a setup where whistleblowers can submit anonymous reports to an independent auditor about an organization they work for. A whistleblower can still be identified by an organization if an audit happens shortly after an internal scandal, even if the audit is generic (Wallmeier and Promann, 2025). This is because a whistleblower may be one of the few people who have access to internal information that motivates the audit in the first place.

The need for such protection is acute in AI governance. Recent calls for responsible reporting frameworks for frontier AI development (Kolt et al., 2024) and end-to-end internal algorithmic auditing (Raji et al., 2020; Anderljung et al., 2023) all envision insider information — and hence whistleblowers — as a central signal. For these mechanisms to be credible, employees willing to file reports must be able to do so without their identity leaking through the very audits their reports trigger. Our work supplies the missing technical primitive.

The key insight to preserve anonymity is this: if the auditor introduces randomness into the organizations selected for audit, the organization cannot be sure whether the audit was due to a whistleblower report, or to randomness. This gives whistleblowers plausible deniability: the ability to deny responsibility due to other alternative possibilities. One classic approach to achieve plausible deniability is using randomized response (Warner, 1965): the mechanism sometimes follows the true signal and sometimes outputs a random response.

Prior work (Chassang and Miquel, 2019) adapts this idea to

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

whistleblowing.

In a practical setup, we would like to maintain a report stream of all whistleblower reports that can be used to select which organization to audit, compiled into an audit transcript. Ideally, the mechanism should protect any single report while still allowing the auditor to make useful audit decisions. Between organizations selected for audit, we would like to impose no restrictions on the number of whistleblower reports. Unfortunately, for this setup, randomized response cannot guarantee a fixed privacy level over a long horizon without making the audit decision increasingly random. In particular, maintaining a fixed privacy guarantee over horizon T requires the probability of random selection to approach one as T grows.

Our contributions are:

1. **Formalization (Sections 2–3).** We formalize whistleblower auditing as per-report $(0, \delta)$-differential privacy on the transcript of audit selection decisions, under a strong-adversary threat model in which the audited organization itself is the adversary. To our knowledge, no prior work targets this exact granularity — per-report adjacency, per-decision observable, fixed horizon — under this threat model.

2. **Negative result for the standard approach (Section 4, Theorem 2).** We prove that randomized response applied at the selection step — the canonical adaptation of Warner's (Warner, 1965) primitive used in the economics literature on whistleblowing (Chassang and Miquel, 2019) — cannot maintain a fixed $(0, \delta)$-DP guarantee over a horizon T without degenerating to uniform random auditing as T grows.

3. **Generic reduction (Section 5, Proposition 3).** We give a mechanism that reduces private auditing to private continual counting: any $(0, \delta)$-DP continual counter plugs in by post-processing, and the audit transcript inherits the same per-report guarantee. Future improvements in continual counting can be swapped in easily.

4. **Concrete instantiation and utility analysis (Sections 5.4 and 6).** We instantiate the reduction with the Toeplitz-factorization counter of Fichtenberger et al. (2023), which was recently shown to be near-optimal for continual counting by Dvijotham et al. (2024). This obtains per-report $(0, \delta)$-DP with noise scaling as $O(\sqrt{\log T})$. A utility theorem (Theorem 6) shows the error decays whenever the noisy report-count gap between the organization with the most reports and the runner-up organization grows faster than $\sqrt{\log T}$ — an asymptotic improvement over randomized response.

5. **Empirical validation (Section 7).** Simulations show that the resulting mechanism substantially outperforms randomized response in both a static gap sweep and a dynamic online auditing setting.

2. An Auditing Framework

We propose the following auditing setup. At each timestep, the auditor collects any new whistleblower reports and selects an organization for audit. The collected reports form a *report stream* $\mathbf{R}_t = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_t) \in \mathbb{N}^{C \times t}$. This is a sequence of vectors where the c-th element of $\mathbf{r}_t$, denoted $r_{t,c}$, counts the number of new reports about organization c at time t. Together, the sequence of audit decisions forms an *audit transcript* $\mathbf{a}_t = (a_1, a_2, \ldots, a_t) \in [C]^t$. The auditing setup must satisfy the following conditions.

Condition 1 (Privacy)**.** *The auditor must guarantee any whistleblower a fixed level of privacy, regardless of report time and report organization.*

We formalize Condition 1 using the following definitions of adjacency and privacy:

Definition 1 (Adjacent Report Streams)**.** *Two report streams $\mathbf{R}_T, \mathbf{R}'_T$ are* adjacent*, written $\mathbf{R}_T \sim \mathbf{R}'_T$, if there exists a unique $(t', c') \in [T] \times [C]$ such that $r_{t',c'} = r'_{t',c'} + 1$ and $r_{s,c} = r'_{s,c}$ for all $(s, c) \neq (t', c')$.*

Definition 2 ((ϵ, δ)-Differential Privacy)**.** *Let $\mathcal{M} : \mathcal{R} \to \mathcal{A}$ be a randomized mechanism, mapping the space of report streams $\mathcal{R}$ to the space of audit decisions $\mathcal{A}$. The mechanism $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy with respect to the adjacency notion of Definition 1 if for all adjacent $\mathbf{R}_T \sim \mathbf{R}'_T \in \mathcal{R}$ and all measurable $A \subseteq \mathcal{A}$,*

$$\Pr[\mathcal{M}(\mathbf{R}_T) \in A] \;\leq\; \exp(\epsilon) \cdot \Pr[\mathcal{M}(\mathbf{R}'_T) \in A] + \delta.$$

Equivalently, $(0, \delta)$-DP holds iff $d_{\mathrm{TV}}(\mathcal{M}(\mathbf{R}_T), \mathcal{M}(\mathbf{R}'_T)) \leq \delta$ for every adjacent pair (a standard fact; proof in Appendix B).

Let there be C possible organizations to audit. Audits based on whistleblower reports are often costly (Kuang et al., 2021) so it is reasonable for an auditor to audit less frequently than whistleblowers can issue reports, motivating the following practical condition:

Condition 2 (Multiple Reporting)**.** *The auditor should be able to receive multiple reports about a single organization between any two timesteps. Within this time-period, an individual whistleblower can only make one report (but can report at other timesteps).*

Under Condition 2), the above notion of adjacency in Definition 1 corresponds to the inclusion or exclusion of a single whistleblower's contribution.

Condition 3 (Report Reset)**.** *After an organization is audited, previous reports about that organization are no longer counted toward future audit decisions.*

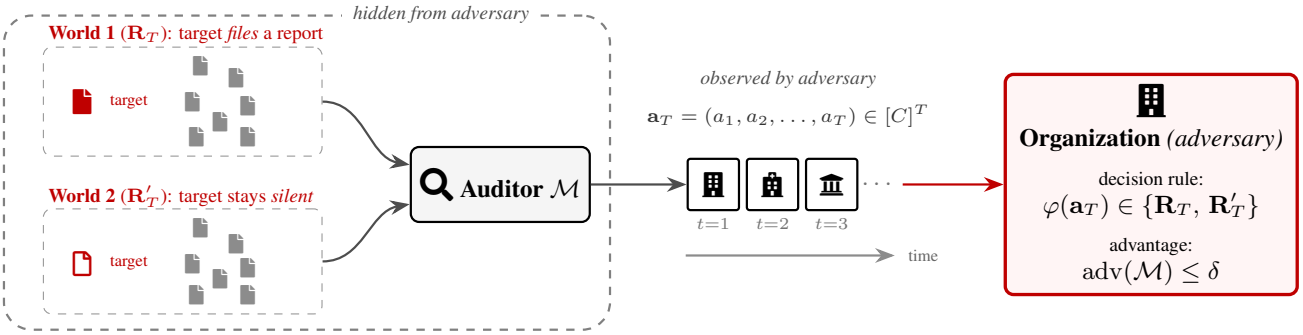


Figure 1. **Threat model — two possible worlds.** In World 1 the target whistleblower files a report (filled red star), yielding stream $\mathbf{R}_T$; in World 2 they remain silent (hollow), yielding $\mathbf{R}'_T$, with all other reporters identical. The auditor $\mathcal{M}$ processes the stream confidentially and releases the audit transcript $\mathbf{a}_T \in [C]^T$. The adversary — the organization itself — observes only $\mathbf{a}_T$ and applies a decision rule $\varphi : [C]^T \to \{\mathbf{R}_T, \mathbf{R}'_T\}$ to identify whistleblowers. The auditor's design bounds the adversary's advantage $\mathrm{adv}(\mathcal{M}) \leq \delta$.

This condition reflects the interpretation of an audit as resolving the currently pending concerns about that organization. Thus, at time t, the relevant count for organization c is not the cumulative number of reports over the entire horizon, but the number of *active* reports: reports received since the most recent audit of c.

Because of Condition 3, reports about an organization are counted only until that organization is audited. We therefore define the *active count* of organization c at time t as the number of reports about c received since the most recent previous audit of c, or since time 1 if c has not yet been audited.

3. A Threat Model

The adversary. Following the strong-adversary threat model of Kulynych et al. (2025), we take the natural adversary in our setting to be the audited organization itself: an entity with near-complete knowledge of its internal state and reporting history, who wishes to identify whistleblowers in order to retaliate. The adversary knows (i) the auditing mechanism $\mathcal{M}$ with all parameters; (ii) two adjacent report streams $\mathbf{R}_T \sim \mathbf{R}'_T$ differing only at one (t^*, c^*); and (iii) the realized audit transcript $\mathbf{a}_T$. Formally, an adversary is any (possibly randomized) decision rule $\phi : [C]^T \to \{\mathbf{R}_T, \mathbf{R}'_T\}$, and may attack via: (1) randomly selecting one of the streams; (2) generating $\hat{\mathbf{a}}_T$ from $\mathcal{M}$ on it; (3) running a singling-out/re-identification attack (Cohen and Nissim, 2020; Kulynych et al., 2025) on $\{\hat{\mathbf{a}}_T, \mathbf{R}_T, \mathbf{R}'_T, \mathcal{M}\}$.

This is the right primitive because retaliation typically follows a binary did-they-report determination, the organization plausibly possesses near-complete internal knowledge, and bounding the strongest adversary automatically bounds all weaker ones.

The advantage of the adversary. The adversary's **baseline** success — guessing $\mathbf{R}_T$ without observing $\mathbf{a}_T$ — is $\Pr[\phi(\emptyset) = \mathbf{R}_T]$. Their **post-observation** success is $\Pr[\phi(\mathbf{a}_T) = \mathbf{R}_T \mid \mathbf{a}_T \sim \mathcal{M}(\mathbf{R}_T)]$. The *additive advantage* of the adversary is the difference between their success rate after observing the output of the mechanism $\mathcal{M}$ and their baseline success rate,

$$\mathrm{adv}(\mathcal{M}) := \Pr\big[\phi(\mathbf{a}_T) = \mathbf{R}_T \big| \mathbf{a}_T \sim \mathcal{M}(\mathbf{R}_T)\big] - \Pr[\phi(\emptyset) = \mathbf{R}_T].$$

Kulynych et al. (2025) show that if $\mathcal{M}$ satisfies $(0, \delta)$-differential privacy, then the maximum value of the advantage, over all possible baseline values, is upper bounded by δ.

Theorem 1 (adapted from Kulynych et al. (2025))**.** *If $\mathcal{M}$ satisfies $(0, \delta)$-differential privacy, then $\mathrm{adv}(\mathcal{M}) \leq \delta$.*

The bound is **baseline-independent**: it applies regardless of any prior the adversary may hold over $\mathbf{R}_T$ vs. $\mathbf{R}'_T$. This is the appropriate guarantee for high-stakes scenarios such as whistleblowing, where assumptions about what the adversary already knows are dangerous to make. We adopt $(0, \delta)$-DP as the operational privacy definition for the rest of the paper, and the auditor offers each whistleblower a fixed δ such that no strong adversary — including the organization itself — can identify their report from the audit transcript with advantage exceeding δ.

Why this setting is novel and technically subtle. Our setting differs from prior work along three axes. First, the privacy unit is a single report but the observable is a sequence of discrete audit decisions in $[C]$ — much sparser than the noisy histograms or aggregate counts of classical continual-observation DP (Dwork et al., 2010; Chan et al., 2011; Fichtenberger et al., 2023; Dvijotham et al., 2024), so off-the-shelf utility analyses do not apply. Second, prior whistleblower mechanisms (Warner, 1965; Chassang and Miquel, 2019) garble each report at submission time under an untrusted-auditor model; we instead trust the auditor with reports but treat the audit transcript as public and observed by a powerful insider adversary, shifting randomization

from the worker to the auditor's selection policy and from per-message to per-report-on-the-transcript over horizon T. Third, matrix-factorization counters (Fichtenberger et al., 2023; Dvijotham et al., 2024) require the horizon to be fixed in advance to certify a privacy level — recent streaming variants (Andersson and Pagh, 2025) partially relax this — and the mechanism's own past audit decisions, themselves driven by noisy counts, determine future counter resets, so the privacy proof must track which counter instance carries the extra report through mechanism-dependent restart histories. Standard composition tools (Dwork et al., 2014; Rogers et al., 2016) do not directly suffice; we handle this via an explicit coupling argument (Appendix B).

Utility. If privacy were not a concern, we assume the auditor would like to select the organization with the largest number of whistleblower reports to audit next. To quantify this, we define the error of an auditing mechanism $\mathcal{M}$ at any time t as the probability it does not select the organization with the largest number of reports at that time.

Definition 3 (Error)**.** *For any time t, let c_t^* be the organization with the largest number of active reports. The error of a mechanism $\mathcal{M}$ given report stream $\mathbf{R}_t$ is the mis-selection probability* $\Pr[\mathcal{M}(\mathbf{R}_t) \neq c_t^*]$.

We now have a concrete way to compare mechanisms. For any mechanism $\mathcal{M}$ that satisfies Conditions 2 and 1, the auditor offers any potential whistleblower a fixed privacy level δ corresponding to an upper bound on the adversary's advantage. For this privacy level, a better mechanism has strictly lower error, as defined in Definition 3. The next natural questions are what guarantees are there for existing mechanisms, and can we do better?

4. The Current Approach: Randomized Response

Existing research on private mechanisms for whistleblower anonymity (Chassang and Miquel, 2019) is inspired by a classic mechanism for plausible deniability, *randomized response* (Warner, 1965). The mechanism is described in Appendix Algorithm 2. The overall idea is to flip a biased coin and, depending on the outcome, to respond either randomly (to increase privacy) or with the desired output (to reduce error).

In order to bound the additive advantage of the adversary, we give a $(0, \delta)$-differential privacy guarantee for the above mechanism.

We show that Randomized Response achieves $(0, 1-p_{\text{rand}}^T)$-DP (Appendix A.1, Proposition 7) with one-step error $\frac{C-1}{C} p_{\text{rand}}$ (Appendix A.1, Proposition 8). This yields a sharp impossibility result.

Theorem 2 (Impossibility of fixed-privacy and non-random selection as T grows)**.** *Fix a target privacy $\delta \in [0, 1)$. Let Randomized Response (Algorithm 2) be calibrated to satisfy $(0, \delta)$-DP at a horizon T via the sufficient choice of $p_{\text{rand}} = (1-\delta)^{1/T}$ from Proposition 7. Then for every horizon T, every time $t \leq T$, and every pre-decision history $\mathcal{H}_{t-1} = h$ with a unique maximizer c_t^* we have,*

$$\Pr[a_t \neq c_t^* \mid \mathcal{H}_{t-1} = h] = \frac{C-1}{C}(1-\delta)^{1/T}.$$

Consequently, for any fixed $\delta < 1$, this error tends to $(C-1)/C$ as $T \to \infty$: the mechanism degenerates to uniform random auditing.

The proof can be found in Appendix A.1.

5. A Generic Reduction to Private Continual Counting

Randomized response adds randomness at every timestep in order to achieve privacy, causing the mechanism error to rapidly approach random selection. Using *continual counting* mechanisms, we can privatize *report counts* instead of the selection step itself, thereby achieving much better utility. These take a stream of non-negative counts and outputs noisy running totals at each step:

Definition 4 (Continual-counting mechanism)**.** *Fix a horizon $T \in \mathbb{N}$. A* continual-counting mechanism *$\mathcal{C}$ is a randomized online process which, when fed a stream $x_1, \ldots, x_T \in \mathbb{N}_0$ sequentially, outputs after each step $t \in [T]$ a noisy estimate $\hat{s}_t$ of the prefix sum $s_t := \sum_{i=1}^t x_i$. Equivalently, for each $x \in \mathbb{N}_0^T$, the mechanism induces a random output vector $\mathcal{C}(\mathbf{x}) = (\hat{s}_1, \ldots, \hat{s}_T) \in \mathbb{R}^T$.*

5.1. A Generic Private Auditing Algorithm

We now describe a generic mechanism $\mathcal{M}$ parametrized by any continual counting mechanism $\mathcal{C}$. The mechanism has two layers. First, each organization has a private running counter of pending reports. Second, the auditor selects the organization with the largest noisy pending count. The privacy proof only needs the first layer; the second layer is post-processing. The mechanism implements the Report Reset condition by restarting an organization's counter after that organization is audited. Thus, each counter tracks only reports received since the organization's most recent audit.

Algorithm 1 Generic Private Auditing Mechanism

Require: Number of organizations C; horizon T; $(0, \delta)$-DP continual-counting mechanism $\mathcal{C}$

Initialise. For each $c \in [C]$: start a fresh instance $\mathcal{C}_c$ of $\mathcal{C}$, and set $\ell_c \leftarrow 0$.
for $t = 1, 2, \ldots, T$ **do**
  for each $c \in [C]$ **do**
    Receive new report counts $r_{t,c} \in \mathbb{N}_0$
    $\ell_c \leftarrow \ell_c + 1$ (time since last audit of c)
    $\tilde{n}_{c,t} \leftarrow \mathcal{C}_c(r_{t,c}, \ell_c)$ (receive noisy count)
  end for
  Select $a_t \leftarrow \arg\max_{c \in [C]} \tilde{n}_{c,t}$ (ties broken uniformly at random).
  Restart: after auditing a_t, replace $\mathcal{C}_{a_t}$ with a fresh instance of $\mathcal{C}$ and set $\ell_{a_t} \leftarrow 0$.
end for

5.2. Privacy Guarantee

Proposition 3 (Privacy)**.** *Fix a horizon $T \in \mathbb{N}$. If $\mathcal{C}$ satisfies $(0, \delta)$-differential privacy as a continual counting mechanism on streams of length T, then Algorithm 1 $\mathcal{M}$ instantiated with $\mathcal{C}$ satisfies $(0, \delta)$-differential privacy w.r.t. adjacent report streams $\mathbf{R}_T \sim \mathbf{R}'_T$ and all measurable $A \subseteq \mathcal{A}$, i.e.*

$$\Pr[\mathcal{M}(\mathbf{R}_T) \in A] \leq \Pr[\mathcal{M}(\mathbf{R}'_T) \in A] + \delta \tag{1}$$

Proof sketch. Let $\mathbf{R}_T \sim \mathbf{R}'_T$ differ by one report for organization c^* at time t^*. Before t^*, the two executions have identical counter inputs and hence identical restart histories. The extra report can therefore enter only one active counter instance, whose two input streams are adjacent. By the assumed $(0, \delta)$-DP guarantee for the continual counter, the outputs of this affected instance differ by at most δ in total variation; all other counter randomness can be coupled identically. The audit transcript is a post-processing of these counter outputs and restart decisions, so it inherits the same $(0, \delta)$-DP guarantee. See Appendix B for the full proof. □

Remark 5.1 (Per-report privacy guarantee)**.** *Proposition 3 guarantees that the audit transcript leaks at most δ additional information about whether any* single specific report *was filed. If a whistleblower files k reports (possibly at different times or about different organizations), their overall privacy follows by basic composition (Dwork et al., 2014) and is at most $(0, k\delta)$-DP. No structural assumption about when reports are submitted is needed for the per-report guarantee.*

5.3. Gaussian Matrix-Mechanism Counter

We now construct a continual-counting mechanism calibrated for $(0, \delta)$-DP. Let $\mathbf{A}^{(T)} \in \{0, 1\}^{T \times T}$ denote the lower-triangular all-ones prefix-sum matrix, i.e. $\mathbf{A}^{(T)}_{i,j} = \mathbf{1}[i \geq j]$. Thus the exact vector of prefix sums of $\mathbf{x} \in \mathbb{N}_0^T$ is $\mathbf{A}^{(T)}\mathbf{x}$. Suppose we are given matrices $\mathbf{B}^{(T)}, \mathbf{C}^{(T)} \in \mathbb{R}^{T \times T}$ such that $\mathbf{B}^{(T)}\mathbf{C}^{(T)} = \mathbf{A}^{(T)}$. We define the associated *Gaussian matrix-mechanism counter* by

$$\mathcal{C}(\mathbf{x}) := \mathbf{B}^{(T)}\big(\mathbf{C}^{(T)}\mathbf{x} + \mathbf{z}\big), \ \mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_T). \tag{2}$$

For privacy, the key quantity is the largest column norm of the encoder:

$$M_T \ := \ \max_{j \in [T]} \|\mathbf{C}^{(T)}\mathbf{e}_j\|_2 \tag{3}$$

Proposition 4 (Base-counter Privacy)**.** *Fix $T \in \mathbb{N}$. Suppose $\mathcal{C}$ is defined by eq. (2), with $\mathbf{B}^{(T)}\mathbf{C}^{(T)} = \mathbf{A}^{(T)}$. If the noise scale is chosen as*

$$\sigma = \frac{M_T}{2\,\Phi^{-1}((1+\delta)/2)}, \tag{4}$$

where M_T is defined as in eq. (3), then $\mathcal{C}$ is $(0, \delta)$-differentially private. Throughout, we write $\kappa_\delta := \Phi^{-1}((1+\delta)/2)$, so that the noise scale can equivalently be written as $\sigma = M_T/(2\kappa_\delta)$.

The proof (Appendix C) follows by applying Lemma 14 to the encoded shift $\mathbf{C}^{(T)}\mathbf{e}_{j^*}$, whose norm is at most M_T, and post-processing through $\mathbf{B}^{(T)}$.

5.4. Concrete Instantiation via Toeplitz Factorization

To maximize utility, we use the Toeplitz factorization of Fichtenberger et al. (2023) and Dvijotham et al. (2024), which is near-optimal for continual counting.

We define the $T \times T$ lower-triangular Toeplitz matrices $\mathbf{C}^{(T)}$ (encoder) and $\mathbf{B}^{(T)}$ (decoder) by

$$\mathbf{C}^{(T)}_{i,j} = \mathbf{B}^{(T)}_{i,j} := \begin{cases} f_{i-j} & \text{if } i \geq j, \\ 0 & \text{if } i < j. \end{cases} \tag{5}$$

where the sequence $(f_k)_{k \geq 0}$ is given by

$$f_0 = 1, \qquad f_k = \Big(1 - \frac{1}{2k}\Big) f_{k-1} \quad \text{for } k \geq 1 \tag{6}$$

The convolution properties of this sequence yields $\mathbf{B}^{(T)}\mathbf{C}^{(T)} = \mathbf{A}^{(T)}$ (Lemma 1 & 2 (Fichtenberger et al., 2023)).

Lemma 5 (Sensitivity Bound for the Toeplitz Encoder)**.** *For the encoder $\mathbf{C}^{(T)}$ above,*

$$\|\mathbf{C}^{(T)}\mathbf{e}_j\|_2^2 = \sum_{k=0}^{T-j} f_k^2 \leq \sum_{k=0}^{T-1} f_k^2$$

which equals M_T^2 eq. (3), as the maximum is attained at $j = 1$. Moreover, $M_T = \mathcal{O}(\sqrt{\log T})$.

Proof. The first identity is immediate from the definition of $\mathbf{C}^{(T)}$. The $\mathcal{O}(\sqrt{\log T})$ bound follows from the analysis in Dvijotham et al. (2024) (Lemma 2.1). □

Instantiating Algorithm 1 with this Toeplitz-based counter – running one independent copy per organization and restarting after each audit – yields our final mechanism, which satisfies $(0, \delta)$-DP. This factorization was shown to be optimal among lower-triangular Toeplitz factorizations for continual counting (Dvijotham et al., 2024). We use this instantiation throughout the remainder of the paper.

6. Utility Guarantee

We now derive a utility guarantee (Definition 3) for the above mechanism. To begin, define the noise history of the mechanism.

Definition 5 (History)**.** *$\mathcal{H}_{t-1} := \sigma\big(\{\tilde{n}_{c,s} : c \in [C],\, s \leq t-1\}\big)$ is the σ-algebra generated by all noisy counter outputs up to time $t-1$. The audit decisions $a_1, \ldots, a_{t-1}$ are $\mathcal{H}_{t-1}$-measurable.*

6.1. Utility Theorem

Definition 6 (Effective gap)**.** *For any realisation of $\mathcal{H}_{t-1}$ and any challenger $c \neq c_t^*$, the* effective gap *between true leader c_t^* and challenger c at time t is*

$$\tilde{\Delta}_{c,t} \;:=\; \underbrace{(n_{c_t^*,t} - n_{c,t})}_{\text{true gap } \Delta_{c,t}} + \underbrace{(G_{c_t^*,t} - G_{c,t})}_{\text{accumulated past noise difference}} \tag{7}$$

where $\Delta_{c,t} = n_{c_t^*,t} - n_{c,t} > 0$ and $G_{c,t}$ is defined in eq. (15). Note that $\tilde{\Delta}_{c,t}$ is fully determined by $\mathcal{H}_{t-1}$ and the data.

Theorem 6 (Utility of Algorithm 1 with Toeplitz Factorization)**.** *Fix any time $t \geq 1$. For any realisation h of $\mathcal{H}_{t-1}$ such that c_t^* is the* unique *true leader ($\Delta_{c,t} > 0$ for all $c \neq c_t^*$), the conditional error satisfies*

$$\Pr\big[a_t \neq c_t^* \mid \mathcal{H}_{t-1} = h\big] \;\leq\; \sum_{c \neq c_t^*} \Phi\left(-\frac{\sqrt{2}\,\kappa_\delta\,\tilde{\Delta}_{c,t}}{M_T}\right), \tag{8}$$

where each term is exact *(equality, not merely an upper bound) for its corresponding challenger.*

For a proof, see Appendix D.

6.2. Remarks

Remark 6.1 (Optimality)**.** *The error bound eq. (8) depends on the noise calibration only through $\sigma = M_T/(2\kappa_\delta)$. Since smaller M_T means smaller σ and hence smaller error, and since Proposition 2.2 of Dvijotham et al. (2024) shows the above factorization uniquely minimises $M_T = \sqrt{\mathrm{MaxErr}(\mathbf{B}^{(T)}, \mathbf{C}^{(T)})}$ over all lower-triangular Toeplitz factorisations, the above mechanism provides the tightest error bound within this class at any fixed privacy level δ.*

Remark 6.2 (Comparison with Randomized Response)**.** *By Theorem 2, randomized response calibrated to $(0, \delta)$-DP at horizon T has error*

$$\mathrm{Error}_T^{\mathrm{RR}} \;=\; \frac{C-1}{C}\,(1-\delta)^{1/T} \;\xrightarrow{T\to\infty}\; \frac{C-1}{C}.$$

At any fixed $\delta < 1$, randomized response approaches random guessing as $T \to \infty$. This degradation is unavoidable for the randomized response approach because maintaining a fixed privacy level requires increasing the probability of random selection at every timestep.

By contrast, the error of our proposed mechanism is governed by $\Phi(-\sqrt{2}\kappa_\delta \tilde{\Delta}_{c,t}/M_T)$. Since $M_T = \mathcal{O}(\sqrt{\log T})$ (Lemma 5), while true pending-count gaps $\Delta_{c,t}$ grow proportionally to accumulated reports, the error tends to zero whenever $\Delta_{c,t}/\sqrt{\log T} \to \infty$.

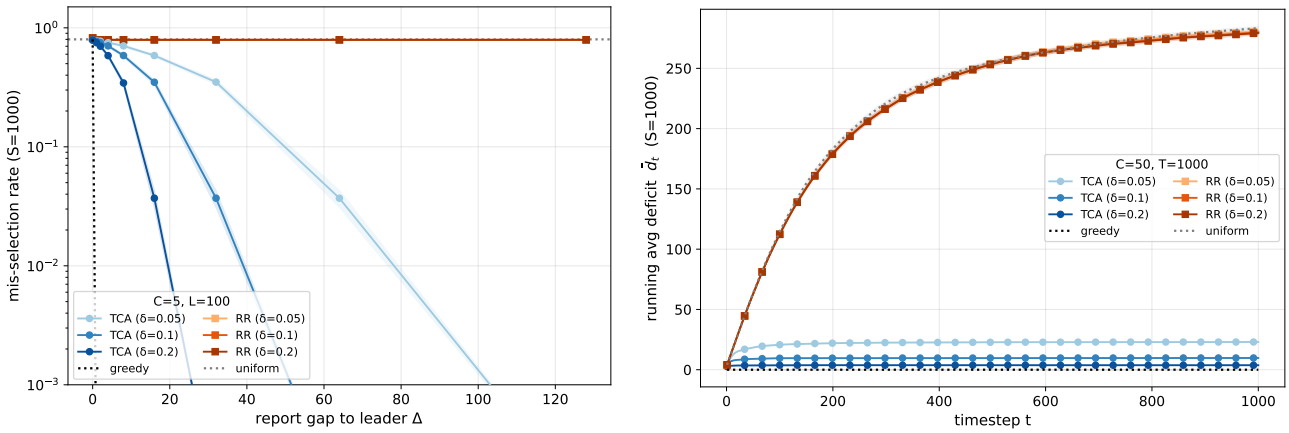
In regimes where the noisy leader–challenger gap grows faster than $\sqrt{\log T}$, the bound yields vanishing error, whereas randomized response approaches uniform random auditing.

7. Experiments

We compare *Toeplitz Continual Auditing* (TCA), with Randomized Response (Warner, 1965), uniformly random auditing, and non-private (greedy) auditing in two simulation studies. All private mechanisms are calibrated to the same per-report $(0, \delta)$-DP guarantee over the relevant horizon; full details can be found in Appendix E.

Experiment 1: Static gap sweep. Theorem 6 predicts that TCA's error should decrease as the effective gap between the true leader and its challengers grows, with noise scale governed by $M_T = \mathcal{O}(\sqrt{\log T})$. To isolate this effect, we consider a single audit decision with one leading organization c_0 having active count $n_0 = \Delta$ and all other organizations having active count $n_c = 0$. We vary Δ and estimate the mis-selection probability $\Pr[a \neq c_0]$. Figure 2a shows the expected qualitative pattern: TCA's error decreases rapidly as the gap grows, while randomized response is essentially insensitive to the gap. The advantage increases with the number of audited organizations (Appendix Figures 3–5).

Experiment 2: Dynamic auditing. The static experiment isolates a single decision, but real auditing is online: each audit resets the selected organization's active count, so different mechanisms induce different future



(a) Mis-selection rate as a function of the leading gap.

(b) Running average active-count deficit over time.

Figure 2. **Utility Simulation.** (a) Static gap sweep: TCA's mis-selection rate decreases as the leader–runner-up gap Δ grows, while randomized response remains nearly gap-insensitive. (b) Dynamic auditing: TCA incurs substantially lower running active-count deficit than randomized response and uniform auditing, approaching the non-private greedy baseline.

states. We therefore simulate a streaming report process over a fixed horizon T and measure the *active count deficit*: $d_t^{\mathcal{M}} := \max_c n_{c,t}^{\mathcal{M}} - n_{a_t,t}^{\mathcal{M}}$, computed before the reset at time t — the regret of $\mathcal{M}$'s selection against the best available active count. Figure 2b shows that TCA remains close to the greedy auditor, while randomized response accumulates much larger deficits.

8. Related Work

Whistleblower mechanisms in economics and computer science. The economics literature has long studied the design of intervention rules that protect informants from retaliation (Chassang and Miquel, 2019; Ortner and Chassang, 2018), with randomized response (Warner, 1965) supplying the canonical garbling primitive. Empirical auditing work confirms that whistleblower reports causally affect downstream audit intensity (Kuang et al., 2021). On the systems side, deployed pipelines such as SecureDrop and academic designs Dissent and Riposte (Corrigan-Gibbs et al., 2015) target sender-anonymity using mix-nets, DC-nets, or PIR. Recent cryptographic work establishes that anonymous transfer without trusted parties is fundamentally hard (Agrikola et al., 2022; Quach et al., 2023), motivating our information-theoretic, DP-based approach in which deniability is provided by the auditor's randomized policy rather than by network-level anonymity.

Differential privacy under continual observation. Our mechanism builds on the continual-release model introduced by Dwork et al. (2010) and refined by Chan et al. (2011), which gives polylogarithmic-error counters and top-k extensions. Recent matrix-factorization mechanisms achieve near-optimal constants (Fichtenberger et al., 2023; Henzinger et al., 2023) and admit efficient streaming implementations (Dvijotham et al., 2024; Andersson and Pagh, 2025). We instantiate our generic counter-and-reset scheme with a Toeplitz factorization to obtain $O(\sqrt{\log T})$ noise.

Private selection and online decision making. At each audit step we solve a private argmax over noisy counters, a problem whose canonical solutions are the exponential mechanism (McSherry and Talwar, 2007), Report-Noisy-Max (Dwork et al., 2014), and Permute-and-Flip (McKenna and Sheldon, 2020). When candidate scores are themselves DP outputs, Liu and Talwar (2019) and Cohen et al. (2024) give general selection frameworks. Closest in structure to our work are differentially private multi-armed bandit algorithms that maintain per-arm tree-counters (Tossou and Dimitrakakis, 2016). However, they differ on three concrete axes: 1. their adjacency unit is a single pull or reward (vs. a single report on our transcript); 2. their objective is cumulative regret over an exploration horizon (vs. count-leader identification at every step under a fixed $(0, \delta)$ constraint); and 3. their feedback model uses stochastic rewards after each pull to update arm estimates (vs. direct report counts with no reward channel from the audit). Bandit regret bounds therefore do not transfer to deniability guarantees, and our utility theorem (Theorem 6) has no analogue in the bandit literature.

Privacy semantics and AI governance auditing. Translating (ϵ, δ)-DP into operational deniability guarantees has recently been unified across re-identification, attribute-inference, and reconstruction risks via f-DP (Kulynych et al., 2025); this is also the framework that links DP to GDPR-style "singling out" (Cohen and Nissim, 2020). Empirical auditing of DP claims complements this view

(Steinke et al., 2023). In AI governance, calls for external scrutiny of frontier systems (Anderljung et al., 2023; Kolt et al., 2024), responsible reporting frameworks (Kolt et al., 2024), and end-to-end algorithmic-audit processes (Raji et al., 2020) all envision insider information as a key signal. Our work supplies the missing technical primitive: an auditor policy whose randomization gives whistleblowers a quantifiable deniability guarantee even after many rounds of audits—the regime in which fixed $(0, \delta)$-DP randomized response provably degenerates.

9. Limitations

There are a few limitations of the current setup. First, composed reports degrade privacy linearly under basic composition Fichtenberger et al. (2022); this could be tightened to f-DP Dvijotham et al. (2024) given a known adversary baseline, but we do not assume one. Second, the threat model assumes that the auditor is a trusted entity that will not leak the private report counts. While this is a standard threat model assumption in centralized differential privacy, it may be worth strengthening the threat model by omitting a trusted auditor. Mechanisms that satisfy local differential privacy (Kasiviswanathan et al., 2011) such as the randomized response approach described in Section 4 would likely be sufficient to guarantee privacy under this threat model. Third, we were unable to obtain any data from a real-world auditing setup. An evaluation with real-world data could expose additional practical considerations that may have been missed in this work.

10. Conclusion

We formalized an auditing setup for whistleblower anonymity and showed that the canonical primitive, randomized response, must approach uniform random selection at any fixed $(0, \delta)$ level as the horizon grows. We then proposed a generic counter-and-restart mechanism whose privacy is inherited by post-processing from any $(0, \delta)$-DP continual-counting mechanism, and instantiated it with the Toeplitz factorization of Fichtenberger et al. (2023) and Dvijotham et al. (2024), yielding per-report $(0, \delta)$-DP with noise calibrated by $M_T = O(\sqrt{\log T})$. Our utility theorem shows that the error decays whenever the noisy leader–challenger gap grows faster than $\sqrt{\log T}$. Natural extensions include accounting for whistleblowers who file multiple reports beyond basic composition and empirical evaluation on real audit pipelines. While this mechanism is intended to make reporting channels more credible by limiting what audit transcripts reveal about individual reports, the guarantee applies only under the stated threat model and should not be taken to protect against operational failures such as access-log leaks, non-anonymous reporting channels, overly narrow audit scopes, or retaliation based on side information.

Impact Statement

This work aims to improve protection for whistleblowers by giving auditors a formal way to randomize audit decisions while preserving utility. The intended positive impact is to reduce retaliation risk and make reporting channels more credible in high-stakes organizations, including AI labs and other organizations developing safety-critical technologies. More broadly, the work illustrates how differential privacy can be used to protect individuals whose actions influence downstream institutional decisions.

However, formal deniability guarantees can also be miscommunicated or over-relied upon. Our guarantee bounds what an adversary can infer from the public audit transcript under the specified threat model; it does not protect against operational failures such as access-log leaks, non-anonymous reporting channels, narrow audit scopes, compromised auditors, or retaliation based on non-technical side information. Deployments should therefore combine mechanisms of this kind with organizational, legal, and procedural protections.

References

T. Agrikola, G. Couteau, and S. Maier. Anonymous whistleblowing over authenticated channels. In E. Kiltz and V. Vaikuntanathan, editors, *Theory of Cryptography*, pages 685–714, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-22365-5.

M. Anderljung, J. Barnhart, A. Korinek, J. Leung, C. O'Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bullock, D. Cass-Beggs, B. Chang, T. Collins, T. Fist, G. Hadfield, A. Hayes, L. Ho, S. Hooker, E. Horvitz, N. Kolt, J. Schuett, Y. Shavit, D. Siddarth, R. Trager, and K. Wolf. Frontier ai regulation: Managing emerging risks to public safety, 2023. URL https://arxiv.org/abs/2307.03718.

J. D. Andersson and R. Pagh. Streaming private continual counting via binning, 2025. URL https://arxiv.org/abs/2412.07093.

S. K. Baljija and K.-s. Min. Evaluating the effectiveness of whistleblower protection: A new index. *Data & Policy*, 5:e28, 2023.

P. G. Cassematis and R. Wortley. Prediction of whistleblowing or non-reporting observation: The role of personal and situational factors. *Journal of business ethics*, 117(3): 615–634, 2013.

T.-H. H. Chan, E. Shi, and D. Song. Private and continual release of statistics. *ACM Transactions on Information and System Security (TISSEC)*, 14(3):1–24, 2011.

S. Chassang and G. P. I. Miquel. Crime, intimidation, and whistleblowing: A theory of inference from unverifiable

reports. *The Review of Economic Studies*, 86(6):2530–2553, 2019.

K. Clarke. She blew the whistle to protect seniors at the rosslyn. she paid a price. https://www.thespec.com/news/hamilton-region/she-blew-the-whistle-to-protect-seniors-at-the-rosslyn-she-paid-a-price/article_86af85f9-9067-5578-8b4a-bb186fccdd5c.html, 2020.

A. Cohen and K. Nissim. Towards formalizing the gdpr's notion of singling out. *Proceedings of the National Academy of Sciences*, 117(15):8344–8352, 2020.

E. Cohen, X. Lyu, J. Nelson, T. Sarlós, and U. Stemmer. Lower bounds for differential privacy under continual observation and online threshold queries. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1200–1222. PMLR, 2024.

H. Corrigan-Gibbs, D. Boneh, and D. Mazières. Riposte: An anonymous messaging system handling millions of users. In *Proceedings of the 2015 IEEE Symposium on Security and Privacy*, SP '15, page 321–338, USA, 2015. IEEE Computer Society. ISBN 9781467369497. doi: 10.1109/SP.2015.27. URL https://doi.org/10.1109/SP.2015.27.

K. D. Dvijotham, H. B. McMahan, K. Pillutla, T. Steinke, and A. Thakurta. Efficient and near-optimal noise generation for streaming differential privacy. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2306–2317. IEEE, 2024.

C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 715–724, 2010.

C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4):211–407, 2014.

H. Fichtenberger, M. Henzinger, and J. Upadhyay. Constant matters: Fine-grained complexity of differentially private continual observation. *arXiv preprint arXiv:2202.11205*, 2022.

H. Fichtenberger, M. Henzinger, and J. Upadhyay. Constant matters: Fine-grained error bound on differentially private continual observation. In *International Conference on Machine Learning*, pages 10072–10092. PMLR, 2023.

M. Henzinger, J. Upadhyay, and S. Upadhyay. Almost tight error bounds on differentially private continual counting. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5003–5039. SIAM, 2023.

E. . C. Initiative. Global business ethics survey. 2020. URL https://business.uccs.edu/sites/g/files/kjihxj2561/files/inline-files/ECI%202020_1-GBES-Pressure%20in%20the%20Workplace.pdf.

S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

N. Kolt, M. Anderljung, J. Barnhart, A. Brass, K. Esvelt, G. K. Hadfield, L. Heim, M. Rodriguez, J. B. Sandbrink, and T. Woodside. Responsible reporting for frontier ai development. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):768–783, Oct. 2024. doi: 10.1609/aies.v7i1.31678. URL https://ojs.aaai.org/index.php/AIES/article/view/31678.

Y. F. Kuang, G. Lee, and B. Qin. Whistleblowing allegations, audit fees, and internal control deficiencies. *Contemporary Accounting Research*, 38(1):32–62, 2021.

B. Kulynych, J. F. Gomez, G. Kaissis, J. Hayes, B. Balle, F. Calmon, and J. L. Raisaro. Unifying re-identification, attribute inference, and data reconstruction risks in differential privacy. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

M. Kutyłowski and G. Wechta. Pseudonymization and reporters' protection by design in the eu whistleblower directive. *Journal of Cybersecurity*, 11(1):tyaf028, 2025.

J. Liu and K. Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, page 298–309, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367059. doi: 10.1145/3313276.3316377. URL https://doi.org/10.1145/3313276.3316377.

R. McKenna and D. R. Sheldon. Permute-and-flip: A new mechanism for differentially private selection. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 193–203. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/01e00f2f4bfcbb7505cb641066f2859b-Paper.pdf.

F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, page 94–103, USA, 2007. IEEE Computer Society. ISBN 0769530109. doi: 10.1109/FOCS.2007.41. URL https://doi.org/10.1109/FOCS.2007.41.

J. R. Mesmer-Magnus and C. Viswesvaran. Whistleblowing in organizations: An examination of correlates of whistleblowing intentions, actions, and retaliation. *Journal of business ethics*, 62(3):277–297, 2005.

J. Ortner and S. Chassang. Making corruption harder: Asymmetric information, collusion, and crime. *Journal of Political Economy*, 126(5):2108–2133, 2018.

J. Powar and A. R. Beresford. Sok: Managing risks of linkage attacks on data privacy. *Proceedings on Privacy Enhancing Technologies*, 2023.

W. Quach, L. Tyner, and D. Wichs. Lower bounds on anonymous whistleblowing. In G. Rothblum and H. Wee, editors, *Theory of Cryptography*, pages 3–32, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-48621-0.

I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 33–44, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372873. URL https://doi.org/10.1145/3351095.3372873.

R. M. Rogers, A. Roth, J. Ullman, and S. Vadhan. Privacy odometers and filters: Pay-as-you-go composition. *Advances in Neural Information Processing Systems*, 29, 2016.

W. E. Scheuerman. Whistleblowing as civil disobedience: The case of edward snowden. *Philosophy & Social Criticism*, 40(7):609–628, 2014.

J. Sharma, S. Kanojia, and S. Sachdeva. Comparison of whistle-blower protection mechanism of select countries. *Indian Journal of Corporate Governance*, 11(1):45–68, 2018.

T. Steinke, M. Nasr, and M. Jagielski. Privacy auditing with one (1) training run. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 49268–49280. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6f6e0d6781d1cb8689192408946d73-Paper-Conference.pdf.

A. Tossou and C. Dimitrakakis. Algorithms for differentially private multi-armed bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016. doi: 10.1609/aaai.v30i1.10212. URL https://ojs.aaai.org/index.php/AAAI/article/view/10212.

N. Wallmeier and T. Promann. The hidden costs of whistleblower protection. *Review of Law & Economics*, (0), 2025.

S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American statistical association*, 60(309):63–69, 1965.

Algorithm 2 Randomized Response Auditing, $\mathcal{M}_{\mathrm{RR}}$

Require: Number of organizations C; random-selection probability $p_{\mathrm{rand}} \in (0, 1]$; horizon T

Initialize pending counts $n_c \leftarrow 0$ for all $c \in [C]$
for $t = 1, 2, \ldots, T$ **do**
  for each $c \in [C]$ **do**
    Receive new reports $r_{t,c}$
    $n_c \leftarrow n_c + r_{t,c}$
  end for
  Sample $X \sim \mathrm{Bernoulli}(p_{\mathrm{rand}})$
  if $X = 1$ **then**
    $a_t \sim \mathrm{Uniform}(\{1, \ldots, C\})$ ▷ explore
  else
    $a_t \leftarrow \arg\max_{c \in [C]} n_c$ (ties broken uniformly) ▷ exploit
  end if
  Output a_t
  Reset: $n_{a_t} \leftarrow 0$
end for

A. Baselines

A.1. Randomized Response

Proposition 7 (Privacy of Randomized Response)**.** *Fix a horizon $T \in \mathbb{N}$. Algorithm 2 satisfies $(0, \delta)$-differential privacy at horizon T for $\delta = 1 - p_{\mathrm{rand}}^T$.*

Proof. We show that, for all adjacent report streams $\mathbf{R}_T \sim \mathbf{R}'_T$ and all measurable $A \subseteq \mathcal{A}$,

$$\begin{aligned}\Pr[\mathcal{M}_{\mathrm{RR}}(\mathbf{R}_T) \in \mathbf{A}_T] \leq \Pr[\mathcal{M}_{\mathrm{RR}}(\mathbf{R}'_T) \in \mathbf{A}_T] \\ + 1 - p_{\mathrm{rand}}^T.\end{aligned} \tag{9}$$

Let $U_1, \ldots, U_T \overset{\mathrm{iid}}{\sim} \mathrm{Bernoulli}(p_{\mathrm{rand}})$ be the random coins drawn by the algorithm, independently of the report stream. Define the *full-exploration event*

$$E := \{U_t = 1 \text{ for all } t \in [T]\}, \qquad \Pr[E] = p_{\mathrm{rand}}^T.$$

Step 1: The output is data-free on E. On E, every round satisfies $U_t = 1$, so Algorithm 2 always explores and outputs $a_t \sim \mathrm{Uniform}(\{1, \ldots, C\})$ independently of the report counts. Thus the conditional output distribution is the same under $\mathbf{R}_T$ and $\mathbf{R}'_T$. Hence, for any $\mathbf{A}_T$,

$$\begin{aligned}\Pr[\mathcal{M}_{\mathrm{RR}}(\mathbf{R}_T) \in \mathbf{A}_T \mid E] = \Pr[\mathcal{M}_{\mathrm{RR}}(\mathbf{R}'_T) \in \mathbf{A}_T \mid E] \\ := P_0(\mathbf{A}_T).\end{aligned} \tag{10}$$

Step 2: Total-probability decomposition. For compactness, define

$$Q_T(\mathbf{A}_T) := \Pr[\mathcal{M}_{\mathrm{RR}}(\mathbf{R}_T) \in \mathbf{A}_T \mid \neg E],$$

and

$$Q'_T(\mathbf{A}_T) := \Pr[\mathcal{M}_{\mathrm{RR}}(\mathbf{R}'_T) \in \mathbf{A}_T \mid \neg E].$$

Conditioning on whether E holds, and using eq. (10), gives

$$\Pr[\mathcal{M}_{\mathrm{RR}}(\mathbf{R}_T) \in \mathbf{A}_T] = p_{\mathrm{rand}}^T P_0(\mathbf{A}_T) + (1 - p_{\mathrm{rand}}^T) Q_T(\mathbf{A}_T), \tag{11}$$

$$\Pr[\mathcal{M}_{\mathrm{RR}}(\mathbf{R}'_T) \in \mathbf{A}_T] = p_{\mathrm{rand}}^T P_0(\mathbf{A}_T) + (1 - p_{\mathrm{rand}}^T) Q'_T(\mathbf{A}_T). \tag{12}$$

Subtracting eq. (12) from eq. (11) yields the exact equality

$$\begin{aligned}\Pr[\mathcal{M}_{\mathrm{RR}}(\mathbf{R}_T) \in \mathbf{A}_T] - \Pr[\mathcal{M}_{\mathrm{RR}}(\mathbf{R}'_T) \in \mathbf{A}_T] \\ = (1 - p_{\mathrm{rand}}^T)\Delta_T(\mathbf{A}_T),\end{aligned} \tag{13}$$

where

$$\Delta_T(\mathbf{A}_T) := Q_T(\mathbf{A}_T) - Q'_T(\mathbf{A}_T).$$

Step 3: Bounding the conditional gap. Since $Q_T(\mathbf{A}_T)$ and $Q'_T(\mathbf{A}_T)$ are probabilities,

$$\Delta_T(\mathbf{A}_T) = Q_T(\mathbf{A}_T) - Q'_T(\mathbf{A}_T) \leq 1.$$

Conclusion. Combining eq. (13) with Step 3 gives

$$\Pr[\mathcal{M}_{\mathrm{RR}}(\mathbf{R}_T) \in \mathbf{A}_T] \leq \Pr[\mathcal{M}_{\mathrm{RR}}(\mathbf{R}'_T) \in \mathbf{A}_T] + 1 - p_{\mathrm{rand}}^T.$$

This holds for every $\mathbf{R}_T \sim \mathbf{R}'_T$ and every $\mathbf{A}_T$, so Algorithm 2 satisfies $(0,\, 1 - p_{\mathrm{rand}}^T)$-DP at horizon T. □

We further want to prove the following statement on the error bound.

Proposition 8 (One-step error of randomized response)**.** *Fix a horizon $T \in \mathbb{N}$ and privacy parameter $\delta \in [0, 1)$. A sufficient privacy calibration from Proposition 7 is*

$$p_{\mathrm{rand}} = (1 - \delta)^{1/T}.$$

Fix any time $t \leq T$ and condition on any pre-decision history $\mathcal{H}_{t-1} = h$. Let $n_{c,t}$ denote the active count of organization c at time t, before the audit decision is made. If c_t^* is the unique maximizer of the active counts, i.e.

$$c_t^* = \arg\max_{c \in [C]} n_{c,t},$$

then Algorithm 2 satisfies

$$\Pr[a_t \neq c_t^* \mid \mathcal{H}_{t-1} = h] = \frac{C-1}{C}\, p_{\mathrm{rand}}.$$

In particular, under the above sufficient privacy calibration,

$$\Pr[a_t \neq c_t^* \mid \mathcal{H}_{t-1} = h] = \frac{C-1}{C}(1 - \delta)^{1/T}.$$

Remark A.1 (Why this parameterisation?)**.** *By Proposition 7, Algorithm 2 satisfies $(0, \delta)$-DP at horizon T whenever*

$$1 - p_{\mathrm{rand}}^T \leq \delta, \qquad \textit{equivalently} \qquad p_{\mathrm{rand}} \geq (1 - \delta)^{1/T}.$$

Thus the smallest exploration probability certified by Proposition 7 is

$$p_{\mathrm{rand}} = (1 - \delta)^{1/T}.$$

Since the one-step error in Proposition 8 increases linearly with p_{rand}, this is the error-minimising choice within this sufficient privacy calibration. For fixed δ, however, $(1 - \delta)^{1/T} \to 1$ as $T \to \infty$, so the mechanism approaches uniformly random auditing over long horizons.

Proof. Fix a time $t \leq T$ and condition on a pre-decision history $\mathcal{H}_{t-1} = h$ such that the active counts $\{n_{c,t}\}_{c \in [C]}$ are fixed and have a unique maximizer c_t^*. Let $U_t \sim \mathrm{Bernoulli}(p_{\mathrm{rand}})$ be the random coin drawn by the algorithm at round t, independently of the report stream and of the past randomness.

By the law of total probability,

$$\begin{aligned}\Pr[a_t \neq c_t^* \mid \mathcal{H}_{t-1} = h] = \Pr[U_t = 1]\Pr[a_t \neq c_t^* \mid U_t = 1, \mathcal{H}_{t-1} = h] \\ + \Pr[U_t = 0]\Pr[a_t \neq c_t^* \mid U_t = 0, \mathcal{H}_{t-1} = h].\end{aligned} \tag{14}$$

If $U_t = 1$, the algorithm explores and draws $a_t \sim \mathrm{Uniform}([C])$. Since exactly one organization equals c_t^*,

$$\Pr[a_t \neq c_t^* \mid U_t = 1, \mathcal{H}_{t-1} = h] = \frac{C-1}{C}.$$

If $U_t = 0$, the algorithm exploits and chooses

$$a_t \in \arg\max_{c \in [C]} n_{c,t}.$$

By assumption, c_t^* is the unique maximizer, so

$$\Pr[a_t \neq c_t^* \mid U_t = 0, \mathcal{H}_{t-1} = h] = 0.$$

Substituting into eq. (14) gives

$$\Pr[a_t \neq c_t^* \mid \mathcal{H}_{t-1} = h] = p_{\mathrm{rand}} \frac{C-1}{C}.$$

Finally, using the sufficient privacy calibration $p_{\mathrm{rand}} = (1-\delta)^{1/T}$ gives

$$\Pr[a_t \neq c_t^* \mid \mathcal{H}_{t-1} = h] = \frac{C-1}{C}(1-\delta)^{1/T}.$$

□

A.2. Greedy Auditing

Algorithm 3 Greedy Auditing Mechanism, $\mathcal{M}_{\mathrm{greedy}}$

Require: Number of organizations C
Initialise. For each $c \in [C]$, set $n_c \leftarrow 0$.
for $t = 1, 2, 3, \ldots$ **do**
  Receive new report counts $\mathbf{r}_t \in \mathbb{N}_0^C$.
  for each $c \in [C]$ **do**
    $n_c \leftarrow n_c + r_{t,c}$ (active reports since last audit of c)
  end for
  Select $a_t \leftarrow \arg\max_{c \in [C]} n_c$ (ties broken uniformly at random).
  Reset: after auditing a_t, set $n_{a_t} \leftarrow 0$.
  Output a_t.
end for

Proposition 9 (No Privacy for Greedy Auditing)**.** *The greedy mechanism $\mathcal{M}_{\mathrm{greedy}}$ in Algorithm 3 does* not *satisfy ϵ-differential privacy for any finite ϵ.*

Proof. It suffices to construct adjacent report streams and an event that has positive probability under one stream and probability zero under the other. Let $C \geq 2$ and consider a time t^*. Let $\mathbf{R}'_T$ have one active report for organization 1 at time t^* and no active reports for organization 2; all other organizations have zero active reports. Let $\mathbf{R}_T$ be identical to $\mathbf{R}'_T$ except for one additional report for organization 2 at time t^*. Then $\mathbf{R}_T \sim \mathbf{R}'_T$ by Definition 1.

Under $\mathbf{R}'_T$, organization 1 is the unique maximizer at time t^*, so

$$\Pr[\mathcal{M}_{\mathrm{greedy}}(\mathbf{R}'_T)_{t^*} = 2] = 0.$$

Under $\mathbf{R}_T$, organizations 1 and 2 are tied for the maximum active count, so tie-breaking gives

$$\Pr[\mathcal{M}_{\mathrm{greedy}}(\mathbf{R}_T)_{t^*} = 2] > 0.$$

For the event $S := \{\mathbf{a} : a_{t^*} = 2\}$, the likelihood ratio $\Pr[\mathcal{M}_{\mathrm{greedy}}(\mathbf{R}_T) \in S] / \Pr[\mathcal{M}_{\mathrm{greedy}}(\mathbf{R}'_T) \in S]$ is therefore infinite. Hence no finite ϵ can satisfy ϵ-differential privacy. □

Proposition 10 (Utility of Greedy Auditing)**.** *Let $n_{c,t}$ denote the active count of organization c at time t, before the audit decision at time t is made. The greedy mechanism in Algorithm 3 satisfies:*

1. $a_t \in \arg\max_{c\in[C]} n_{c,t}$ *with probability* 1.
2. $\Pr[n_{a_t,t} = 0 \mid \exists c : n_{c,t} > 0] = 0$.
3. *If c^* is the unique maximizer of active counts at time t^*, then $a_{t^*} = c^*$ with probability* 1.

The proof follows by construction.

A.3. Uniformly Random Auditing

Algorithm 4 Uniformly Random Auditing, $\mathcal{M}_{\text{uni}}$

Require: organizations $[C] = \{1, \dots, C\}$
for $t = 1, 2, 3, \dots$ **do**
  Receive $r(t)$ (ignored)
  $a_t \sim \text{Uniform}([C])$
  Output a_t
end for

Proposition 11 (Perfect Privacy for Uniformly Random Auditing)**.** *The uniformly random mechanism $\mathcal{M}_{\text{uni}}$ in Algorithm 4 satisfies 0-differential privacy: for all streams $\mathbf{R}_T, \mathbf{R}'_T$ and all measurable S,*

$$\Pr[\mathcal{M}_{\text{uni}}(\mathbf{R}_T) \in S] = \Pr[\mathcal{M}_{\text{uni}}(\mathbf{R}'_T) \in S].$$

Remark A.2. *As throughout the paper, the resulting audit transcript induces active counts via condition 3: Once an organization is audited, previous reports about that organization are no longer active.*

Proof. The output distribution is independent of the report stream. At every timestep, the mechanism samples uniformly from $[C]$, regardless of the input. Therefore, for any two streams $\mathbf{R}_T, \mathbf{R}'_T$, the induced distributions over audit transcripts are identical. □

Proposition 12 (Utility of Uniformly Random Auditing)**.** *Let $n_{c,t}$ denote the active count of organization c at time t, before the audit decision at time t is made. At any time t:*

1. $\Pr[a_t = c] = 1/C$ *for all $c \in [C]$, regardless of the active counts.*
2. $\Pr[a_t \in \arg\max_c n_{c,t}] = M_t/C$, *where $M_t := |\arg\max_c n_{c,t}|$ is the number of active-count maximizers.*
3. *If t^* is a time at which organization c^* first has an active report, and t_{exit} is the first time $t \geq t^*$ at which $a_t = c^*$, then*

$$\mathbb{E}[t_{\text{exit}} - t^* + 1] = C.$$

The proof follows by construction.

B. Proof of Proposition 3 (Privacy of Algorithm 1)

Before proving Proposition 3, we quickly state an equivalent privacy characterization that will be useful for the proof.

Proposition 13 (TV characterization)**.** *A mechanism $\mathcal{M}$ satisfies $(0, \delta)$-DP if and only if for every $\mathbf{R}_T \sim \mathbf{R}'_T$,*

$$d_{\text{TV}}\big(\mathcal{M}(\mathbf{R}_T),\ \mathcal{M}(\mathbf{R}'_T)\big) \ \leq\ \delta.$$

Proof. Definition 2 requires $\Pr[\mathcal{M}(\mathbf{R}_T) \in A] - \Pr[\mathcal{M}(\mathbf{R}'_T) \in A] \leq \delta$ for every measurable A and every adjacent pair. Taking the supremum over A, and using that $\sim$ is symmetric, yields exactly $d_{\text{TV}}(\mathcal{M}(\mathbf{R}_T), \mathcal{M}(\mathbf{R}'_T)) \leq \delta$. The converse is immediate from the definition of TV distance. □

We give the full proof of Proposition 3; the sketch appears in Section 5.2.

Proof. Let $\mathbf{R}_T \sim \mathbf{R}'_T$ be any adjacent pair, differing at (t^*, c^*) with $r'_{t^*,c^*} = r_{t^*,c^*} + 1$ and $r_{t,c} = r'_{t,c}$ for all $(t, c) \neq (t^*, c^*)$. The proof has four steps.

Step 1: The transcript is a deterministic function of counter outputs.

At each step t, the decision $a_t = \arg\max_c \tilde{n}_{c,t}$ is a deterministic function of $\{\tilde{n}_{c,t}\}_{c\in[C]}$, which are outputs of the counter instances. By induction over t, the full transcript $(a_1, \ldots, a_N)$ is a deterministic function g of the collection of all outputs produced by all counter instances across all steps.

Step 2: Identify the unique counter instance affected by the extra report.

We show by induction on t that for all $t < t^*$, all counter outputs at step t are *the same random variables* under $\mathbf{R}_T$ and $\mathbf{R}'_T$.

Base case ($t = 1$, or vacuously if $t^* = 1$): at $t = 1$ all instances are freshly initialised with randomness drawn independently of all data. Since $r_{1,c} = r'_{1,c}$ for every c (the streams agree at $t = 1$ if $t^* > 1$), all counter outputs at $t = 1$ are identical under both streams.

Inductive step: assume all counter outputs at steps $1, \ldots, t-1$ are identical under $\mathbf{R}_T$ and $\mathbf{R}'_T$. Then a_{t-1} is the same under both streams. Any restart at step $t-1$ produces a fresh instance with randomness drawn independently of all data; since the restart event is identical, this instance is the same random object under both streams. At step t, each $\mathcal{C}_c$ receives the same input $r_{t,c} = r'_{t,c}$ (as $t < t^*$) and has the same internal state, so its output is identical. The induction goes through.

It follows that the audit history $\mathcal{H}_{t^*-1} = (a_1, \ldots, a_{t^*-1})$ is *identical* under $\mathbf{R}_T$ and $\mathbf{R}'_T$. In particular, the restart history of c^* up to $t^* - 1$ is the same, so the counter instance $\mathcal{C}^*_{c^*}$ that is active for c^* at time t^* — started at some step $s^* \leq t^*$ — is the same object (same randomness, started at the same step) under both streams.

For $c \neq c^*$: reports are identical under $\mathbf{R}_T$ and $\mathbf{R}'_T$ at all times, and restart decisions (driven by $\mathcal{H}$) are the same. Hence every instance of $\mathcal{C}_c$ is the same random variable under both streams.

For $c = c^*$: every instance of $\mathcal{C}_{c^*}$ *other than* $\mathcal{C}^*_{c^*}$ either ran entirely before t^* (identical inputs) or starts fresh after some audit of c^* at time $s > t^*$ (with reports identical for all $t \neq t^*$, and identical restart times conditional on $\mathcal{C}^*_{c^*}$ agreeing — see Step 4). In either case its inputs do not include the extra report.

The *only* instance whose input stream differs between $\mathbf{R}_T$ and $\mathbf{R}'_T$ is $\mathcal{C}^*_{c^*}$: it receives r_{t,c^*} under $\mathbf{R}_T$ and $r_{t,c^*} + \mathbf{1}[t = t^*]$ under $\mathbf{R}'_T$ at each step of its run, so its input streams are adjacent (differing at position $j^* = t^* - s^* + 1 \leq T$).

Step 3: Apply the DP guarantee of $\mathcal{C}$.

Since $\mathcal{C}^*_{c^*}$ processes a stream of length at most T that is adjacent between $\mathbf{R}_T$ and $\mathbf{R}'_T$, and $\mathcal{C}$ satisfies $(0, \delta)$-DP:

$$d_{\mathrm{TV}}\big(\text{outputs of } \mathcal{C}^*_{c^*} \text{ under } R,\ \text{outputs of } \mathcal{C}^*_{c^*} \text{ under } R'\big) \ \leq\ \delta.$$

Step 4: Couple all counter outputs and conclude.

Couple all counter instances identically except $\mathcal{C}^*_{c^*}$ (possible since they are the same random variables under both streams). For $\mathcal{C}^*_{c^*}$, use the *optimal coupling*, which achieves $\Pr[\mathcal{C}^*_{c^*} \text{ disagrees}] \leq \delta$.

Conditional on $\mathcal{C}^*_{c^*}$ agreeing: a_{t^*} is the same under both streams (it is a deterministic function of the noisy counts, all of which are now identical). Therefore all subsequent restarts occur at the same times, all subsequent counter inputs are identical (reports agree everywhere except at t^*, already handled), and all subsequent counter outputs can be coupled identically. Hence:

$$\Pr[\text{any counter output disagrees}] \ \leq\ \delta.$$

Since g is deterministic, agreement of all counter outputs implies agreement of transcripts. By the coupling characterisation of TV distance:

$$d_{\mathrm{TV}}\big(\mathcal{M}(R),\, \mathcal{M}(R')\big) \ \leq\ \Pr[\text{transcripts disagree}] \ \leq\ \delta.$$

Since $d_{\mathrm{TV}}(P, Q) = \sup_{\mathcal{A}} |\Pr[P \in \mathcal{A}] - \Pr[Q \in \mathcal{A}]|$, this gives $(0, \delta)$-DP for all adjacent pairs. □ □

C. Proof of Proposition 4

Before proving Proposition 4 we state a useful textbook Lemma.

Lemma 14 (TV distance for shifted Gaussians)**.** *If*

$$P = \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n), \qquad Q = \mathcal{N}(\boldsymbol{\mu} + \Delta, \sigma^2 \mathbf{I}_n),$$

then

$$d_{\mathrm{TV}}(P, Q) = 2\Phi\left(\frac{\|\Delta\|_2}{2\sigma}\right) - 1,$$

where Φ is the standard Gaussian CDF.

Proof. Let $\mathbf{x} \sim \mathbf{x}'$ be adjacent, differing only in coordinate j^*. Then $\mathbf{x}' - \mathbf{x} = \mathbf{e}_{j^*}$, so the encoded noisy vectors satisfy

$$\mathbf{C}^{(T)}\mathbf{x}' + \mathbf{z} = \mathbf{C}^{(T)}\mathbf{x} + \mathbf{z} + \mathbf{C}^{(T)}\mathbf{e}_{j^*}.$$

Thus the two encoded distributions are Gaussians with common covariance $\sigma^2 \mathbf{I}_T$ and mean shift $\Delta = \mathbf{C}^{(T)}\mathbf{e}_{j^*}$. By Lemma 14,

$$d_{\mathrm{TV}}\left(\mathbf{C}^{(T)}\mathbf{x}', \mathbf{C}^{(T)}\mathbf{x}\right) \leq 2\Phi\left(\frac{\|\mathbf{C}^{(T)}\mathbf{e}_{j^*}\|_2}{2\sigma}\right) - 1 \leq 2\Phi\left(\frac{M_T}{2\sigma}\right) - 1 = \delta.$$

Since $\mathcal{C}(\mathbf{x})$ is obtained from the encoded noisy vector by deterministic post-processing through $\mathbf{B}^{(T)}$, total variation cannot increase. Hence $\mathcal{C}$ is $(0, \delta)$-*DP*. □

D. Proof of Theorem 6 (Utility)

The proof of Theorem 6 in Section 6.1 relies on the following Lemma.

Lemma 15 (Fresh-noise decomposition)**.** *Fix any time $t \in [T]$ and any organization $c \in [C]$. Let $\ell := \ell_{c,t} \in [T]$ be the length of the active run of c at time t and define $\tau_c := t - \ell_{c,t} + 1$. For simplicity, we re-number and write $\mathbf{x}^{(c,t)} := (x_1^{(c,t)}, \ldots, x_{\ell_{c,t}}^{(c,t)})$ for the report counts in that run, and $\boldsymbol{\zeta}^{(c,t)} := (\zeta_1^{(c,t)}, \ldots, \zeta_{\ell_{c,t}}^{(c,t)})$ for the Gaussian noise vector associated with the active instance of the Gaussian matrix counter instantiated by the $\ell_{c,t} \times \ell_{c,t}$ leading principal truncations $\mathbf{B}^{(\ell_{c,t})}$ and $\mathbf{C}^{(\ell_{c,t})}$.*

Define

$$G_{c,t} := \sum_{j=1}^{\ell_{c,t}-1} f_{\ell_{c,t}-j}\zeta_j^{(c,t)}, \tag{15}$$

with $G_{c,t} := 0$ if $\ell_{c,t} = 1$. Then the noisy pending count satisfies

$$\tilde{n}_{c,t} = \underbrace{n_{c,t} + G_{c,t}}_{=:\mu_{c,t}} + \zeta_{\ell_{c,t}}^{(c,t)}, \tag{16}$$

where $n_{c,t} = \sum_{j=1}^{\ell_{c,t}} x_j^{(c,t)}$ is the count of currently active reports. Moreover:

(i) $\zeta_{\ell_{c,t}}^{(c,t)}$ is independent of the global history $\mathcal{H}_{t-1}$

(ii) for $c \neq c'$, the variables $\zeta_{\ell_{c,t}}^{(c,t)}$ and $\zeta_{\ell_{c,t}}^{(c',t)}$ are independent

Proof. Since the active run has length $\ell_{c,t}$, the current counter output is the $\ell_{c,t}$-th output of the truncated mechanism, i.e.

$$\tilde{n}_{c,t} = n_{c,t} + (\mathbf{B}^{(\ell_{c,t})}\boldsymbol{\zeta}^{(c,t)})_{\ell_{c,t}}$$

Because $\mathbf{B}^{(\ell_{c,t})}$ is lower-triangular Toeplitz with

$$\mathbf{B}_{\ell_{c,t},j}^{(\ell_{c,t})} = f_{\ell_{c,t}-j} \qquad \text{for } 1 \leq j \leq \ell_{c,t},$$

we obtain

$$
\begin{aligned}
(\mathbf{B}^{(\ell_{c,t})}\boldsymbol{\zeta}^{(c,t)})_{\ell_{c,t}} &= \sum_{j=1}^{\ell_{c,t}} f_{\ell_{c,t}-j}\zeta_j^{(c,t)} \\
&= f_0\zeta_{\ell_{c,t}}^{(c,t)} + \sum_{j=1}^{\ell_{c,t}-1} f_{\ell_{c,t}-j}\zeta_j^{(c,t)} \\
&= \zeta_{\ell_{c,t}}^{(c,t)} + G_{c,t}
\end{aligned} \tag{17}
$$

using $f_0 = 1$ and the definition of $G_{c,t}$. This proves eq. (16).

For part (i), every earlier output of the same active instance corresponds to some run-position $\ell < \ell_{c,t}$, and therefore depends only on $\zeta_1^{(c,t)}, \ldots, \zeta_l^{(c,t)}$, hence only on $\zeta_1^{(c,t)}, \ldots, \zeta_{\ell_{c,t}-1}^{(c,t)}$. Thus the coordinate $\zeta_{\ell_{c,t}}^{(c,t)}$ has not appeared in any previous output of the active instance. Since the coordinates of $\zeta_{\ell_{c,t}}^{(c,t)}$ are i.i.d. Gaussian, $\zeta_{\ell_{c,t}}^{(c,t)}$ is independent of $\zeta_1^{(c,t)}, \ldots, \zeta_{\ell_{c,t}-1}^{(c,t)}$ and therefore independent of $\mathcal{H}_{t-1}$.

For part (ii), if $c \neq c'$, then the active instances for c and c' use independently sampled Gaussian vectors. Hence the fresh coordinates $\zeta_{\ell_{c,t}}^{(c,t)}$ and $\zeta_{\ell_{c,t}}^{(c',t)}$ are independent. □

Proof of Theorem 6. Throughout, condition on $\mathcal{H}_{t-1} = h$, which fixes ℓ_c, $n_{c,t}$, G_c, $\mu_{c,t} = n_{c,t} + G_c$, and c_t^*. The only remaining randomness is $\{\zeta_{\ell_{c,t}}^{(c,t)}\}_{c\in[C]}$.

Step 1: Noisy counts in terms of fresh noise. By Lemma 15, $\tilde{n}_{c,t} = \mu_{c,t} + \zeta_{\ell_{c,t}}^{(c,t)}$ where $\{\zeta_{\ell_{c,t}}^{(c,t)}\}_{c\in[C]}$ are i.i.d. $\mathcal{N}(0,\sigma^2)$, independent of h. Hence $a_t = \arg\max_c(\mu_{c,t} + \zeta_{\ell_{c,t}}^{(c,t)})$.

Step 2: Union bound over challengers.

$$
\{a_t \neq c_t^*\} \;\subseteq\; \bigcup_{c\neq c_t^*} \{\zeta_{\ell_{c,t}}^{(c,t)} - \zeta_{\ell_{c_t^*,t}}^{(c_t^*,t)} \;\geq\; \mu_{c_t^*} - \mu_{c,t}\}.
$$

By the union bound:

$$
\Pr\big[a_t \neq c_t^* \mid h\big] \;\leq\; \sum_{c\neq c_t^*} \Pr\big[W_c \geq \tilde{\Delta}_{c,t} \mid h\big], \tag{18}
$$

where $W_c := \zeta_{\ell_{c,t}}^{(c,t)} - \zeta_{\ell_{c_t^*,t}}^{(c_t^*,t)}$.

Step 3: Exact probability for each challenger. By Lemma 15(i)–(ii), $\zeta_{\ell_{c,t}}^{(c,t)}$ and $\zeta_{\ell_{c_t^*,t}}^{(c_t^*,t)}$ are independent $\mathcal{N}(0,\sigma^2)$ variables, independent of h. Their difference satisfies $W_c \sim \mathcal{N}(0, 2\sigma^2)$, so

$$
\Pr\big[W_c \geq \tilde{\Delta}_{c,t} \mid h\big] \;=\; \Phi\left(-\frac{\tilde{\Delta}_{c,t}}{\sqrt{2}\,\sigma}\right). \tag{19}
$$

This is an *equality*, not an inequality.

Step 4: Substitute the noise calibration. With $\sigma = M_T/(2\kappa_\delta)$, we have $\sqrt{2}\,\sigma = M_T/(\sqrt{2}\,\kappa_\delta)$, so $\tilde{\Delta}_{c,t}/(\sqrt{2}\,\sigma) = \sqrt{2}\,\kappa_\delta\,\tilde{\Delta}_{c,t}/M_T$. Substituting into eq. (18) via eq. (19) gives the conditional bound eq. (8). Taking expectations and applying the law of total expectation gives eq. (**??**). □

Remark D.1 (A sufficient condition for small unconditional error)**.** *A condition on true pending counts (before noise) can be given as follows. Let*

$$
\mathsf{ME}_T := \mathrm{MaxErr}(\mathbf{B}^{(T)}, \mathbf{C}^{(T)}) = M_T^2.
$$

For any $c \neq c_t^*$, the accumulated-noise difference $G_{c_t^*,t} - G_{c,t}$ is mean-zero Gaussian with variance $V \leq 2\sigma^2 M_T^2$. Therefore,

$$\begin{aligned}\Pr\left[\tilde{\Delta}_{c,t} \leq \frac{\Delta_{c,t}}{2}\right] &= \Pr\left[G_{c_t^*,t} - G_{c,t} \leq -\frac{\Delta_{c,t}}{2}\right] \\ &\leq \exp\left(-\frac{\Delta_{c,t}^2}{16\sigma^2 M_T^2}\right) \\ &= \exp\left(-\frac{\kappa_\delta^2 \Delta_{c,t}^2}{4\mathsf{ME}_T^2}\right). \end{aligned} \tag{20}$$

Here the inequality uses the one-sided Gaussian tail bound $\Pr[Y \leq -\lambda] \leq \exp(-\lambda^2/(2V))$ for $Y \sim \mathcal{N}(0, V)$, and the final equality uses $\sigma = M_T/(2\kappa_\delta)$.

By total probability, for each challenger c,

$$\Pr[\tilde{n}_{c,t} \geq \tilde{n}_{c_t^*,t}] \leq \Phi\left(-\frac{\kappa_\delta \Delta_{c,t}}{\sqrt{2} M_T}\right) + \exp\left(-\frac{\kappa_\delta^2 \Delta_{c,t}^2}{4\mathsf{ME}_T^2}\right). \tag{21}$$

Both terms decrease rapidly as $\Delta_{c,t}$ grows, giving an unconditional error bound in terms of the true pending-count advantage alone.

Remark D.2 (Interpretation of the effective gap $\tilde{\Delta}_{c,t}$)**.** *The effective gap decomposes as $\tilde{\Delta}_{c,t} = \Delta_{c,t} + (G_{c_t^*,t} - G_{c,t})$, where $\Delta_{c,t} > 0$ is the true pending-count advantage and $G_{c_t^*,t} - G_{c,t}$ is the difference of accumulated past noise terms. The latter has mean zero and variance $\sigma^2\left(\sum_{k=1}^{\ell_{c_t^*}-1} f_k^2 + \sum_{k=1}^{\ell_{c,t}-1} f_k^2\right) \leq 2\sigma^2 M_T^2$. When $\Delta_{c,t}$ is large relative to σM_T, the accumulated noise is unlikely to reverse the ordering and the effective gap is typically close to $\Delta_{c,t}$.*

Remark D.3 (Tightness of the error bound)**.** *The only inequality in the proof is the union bound (Step 2), which is tight when one challenger dominates the sum, for example when $C = 2$ (a single challenger, in which case the bound is an equality) or when one challenger has a much smaller effective gap than all others. The per-challenger probabilities eq. (19) are equalities. The bound is therefore the tightest of this form.*

E. Experimental Details

All experiments were run on a standard laptop/CPU machine; no GPU was used. The full experiment suite completes in under 10 minutes.

E.1. Experiment 1 – Mis-selection over gap

The goal of this set of experiments is to show that TCA's mis-selection rate decays sharply with the gap Δ between the leader and the runner-up, while RR is essentially gap-independent and saturates near $(C-1)/C$. The sweep over the number of organizations C sweep demonstrates that the qualitative picture is robust across pool sizes spanning 1.5 orders of magnitude; the gap required to suppress mis-selection grows only mildly with C (roughly like $\sigma\sqrt{2\ln C}$, the scale of the maximum of $C-1$ Gaussians).

Setup. We fix a single-shot configuration: each organization has been observed for exactly L steps since its last reset, and the pending counts are $n_0 = \Delta$ for the leader (organization c_0) and $n_c = 0$ for all other organizations.

For each (L, δ, Δ) we run $S = 1000$ Monte-Carlo trials and estimate $\Pr[a \neq 0]$. For TCA we exploit that the total noise on each organization at run length L is Gaussian with variance $\sigma^2 \sum_{k=0}^{L-1} f_k^2$, drawn independently per organization. Greedy and uniform-random mis-selection probabilities are computed in closed form (0 if $\Delta > 0$, $(C-1)/C$ if $\Delta = 0$ for greedy; $(C-1)/C$ always for uniform).

Hyperparameters. Calibrated noise scales are independent of C: TCA's σ depends only on L and δ, and likewise p_{rand} for RR.

Table 1. **Experimental parameters.**

Parameter	Value
organization counts C	$\{5,\ 20,\ 50,\ 200\}$ (one figure each)
Run lengths L	$\{100,\ 1000\}$ (one panel each)
Privacy parameters δ	$\{0.05,\ 0.10,\ 0.20\}$
Gaps Δ	$\{0,\ 1,\ 2,\ 4,\ 8,\ 16,\ 32,\ 64,\ 128\}$
Monte-Carlo trials per point	$S = 1000$
RNG seed (TCA)	42
RNG seed (RR)	123

Table 2. **Calibrated noise scales for TCA and randomized response.**

L	δ	σ	p_{rand}
100	0.05	12.6862	0.999487
100	0.10	6.3306	0.998947
100	0.20	3.1400	0.997771
1000	0.05	14.4078	0.999949
1000	0.10	7.1897	0.999895
1000	0.20	3.5661	0.999777

E.2. Experiment 2 – Dynamic Utility Auditing

We evaluate the mechanisms in an online report stream with report resets. The goal of this experiment is to measure how much utility is lost over time when using a private auditing mechanism rather than the non-private greedy baseline.

Mechanisms. We compare four auditing mechanisms:

1. **Toeplitz Counter Auditing (TCA).** Each organization runs an independent Toeplitz continual-counting instance. At each timestep the mechanism audits the organization with the largest noisy active-count estimate. After an audit, the selected organization's active count and private counter instance are reset with fresh independent randomness.

2. **Randomized response auditing (RR).** With probability $p_{\mathrm{rand}} = (1-\delta)^{1/T}$, the mechanism audits a uniformly random organization. With probability $1 - p_{\mathrm{rand}}$, it audits an organization with maximal active count, with ties broken uniformly at random. After the audit, the selected organization's active count is reset. This reset step is important: RR is evaluated under the same active-report auditing semantics as TCA and Greedy.

3. **Greedy auditing.** The mechanism audits an organization with maximal active count, with ties broken uniformly at random. After the audit, the selected organization's active count is reset. This is a non-private upper baseline.

4. **Uniform random auditing.** The mechanism audits a uniformly random organization, ignoring reports. After the audit, the selected organization's active count is reset. This is a privacy-only lower baseline.

Privacy calibration. For each run, the private mechanisms are calibrated to the same transcript-level privacy guarantee $(0, \delta)$ over the same certified horizon T. TCA uses the Gaussian Toeplitz calibration

$$\sigma = \frac{M_T}{2\Phi^{-1}((1+\delta)/2)},$$

where M_T is the Toeplitz encoder sensitivity. Randomized response uses

$$p_{\mathrm{rand}} = (1-\delta)^{1/T}.$$

Thus both private mechanisms are compared under the same fixed-horizon privacy requirement.

Report stream. We generate reports from an independent Poisson model. At each timestep t, organization 0 receives reports

$$x_{t,0} \sim \text{Poisson}(\lambda_{\text{lead}}),$$

and each other organization $c \in \{1, \ldots, C-1\}$ receives reports

$$x_{t,c} \sim \text{Poisson}(\lambda_{\text{other}}).$$

All draws are independent across timesteps and organizations. Although organization 0 has the highest arrival rate in expectation, the organization with maximal active count need not be organization 0 at a given timestep, because audits reset active reports.

Mechanism-specific active counts. Each mechanism $\mathcal{M}$ maintains its own active-count vector

$$n_t^{\mathcal{M}} = (n_{1,t}^{\mathcal{M}}, \ldots, n_{C,t}^{\mathcal{M}}),$$

because different mechanisms audit different organizations and therefore induce different reset histories. At each timestep t, the new report vector x_t is added to that mechanism's active counts. The mechanism then selects an audited organization $a_t^{\mathcal{M}}$. All utility metrics are computed before applying the reset. Finally, the selected organization is reset:

$$n_{a_t^{\mathcal{M}},t}^{\mathcal{M}} \leftarrow 0.$$

For TCA, the selected organization's private counter instance is also restarted.

Metrics. Our primary metric is the *active-count deficit*

$$d_t^{\mathcal{M}} := \max_{c \in [C]} n_{c,t}^{\mathcal{M}} - n_{a_t^{\mathcal{M}},t}^{\mathcal{M}}.$$

This measures how many fewer active reports are resolved by $\mathcal{M}$'s selected audit than by the best available audit under the same reset history. We report the running average deficit

$$\bar{d}_t^{\mathcal{M}} := \frac{1}{t} \sum_{s=1}^{t} d_s^{\mathcal{M}}.$$

We also compute a normalized deficit,

$$\eta_t^{\mathcal{M}} := \frac{d_t^{\mathcal{M}}}{\max_{c \in [C]} n_{c,t}^{\mathcal{M}} + 1},$$

and its running average

$$\bar{\eta}_t^{\mathcal{M}} := \frac{1}{t} \sum_{s=1}^{t} \eta_s^{\mathcal{M}}.$$

The normalized version is useful for comparing regimes with different report volumes. Finally, we record the number of reports resolved by the selected audit,

$$u_t^{\mathcal{M}} := n_{a_t^{\mathcal{M}},t}^{\mathcal{M}},$$

and its running average

$$\bar{u}_t^{\mathcal{M}} := \frac{1}{t} \sum_{s=1}^{t} u_s^{\mathcal{M}}.$$

Lower values of $\bar{d}_t^{\mathcal{M}}$ and $\bar{\eta}_t^{\mathcal{M}}$ indicate better performance, while higher values of $\bar{u}_t^{\mathcal{M}}$ indicate that audits are resolving more active reports.

Implementation details. Unless otherwise stated, we use

$$C = 50, \qquad T = 1000, \qquad \delta = 0.10, \qquad \lambda_{\text{lead}} = 1.0, \qquad \lambda_{\text{other}} = 0.2.$$

We average over 100 independent random seeds. For each seed, the same generated report stream is fed to each mechanism, but each mechanism maintains its own active counts and reset history. We plot the mean running metric across seeds with.

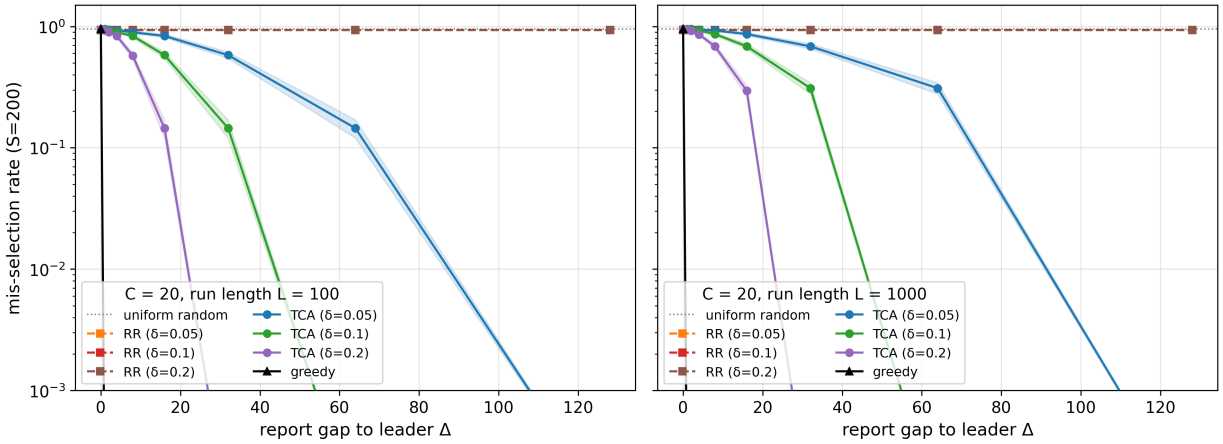


Figure 3. **Mis-selection Rate vs. Leading Gap (C=20).**

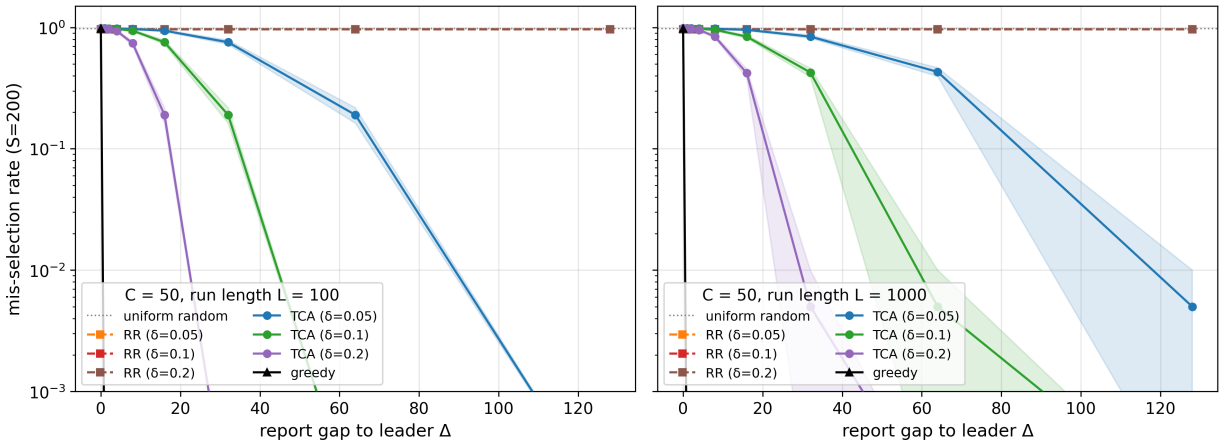


Figure 4. **Mis-selection Rate vs. Leading Gap (C=50).**

Interpretation. Greedy auditing has zero active-count deficit by construction. Uniform random auditing ignores reports and therefore provides a lower utility baseline. Randomized response is expected to behave close to uniform random auditing in long horizons, because the fixed-horizon privacy calibration makes p_{rand} close to one. TCA should accumulate substantially less active-count deficit when the active-count gaps are large relative to the calibrated counter noise.

F. Additional Results

F.1. Experiment 1 with $C \in \{20, 50, 200\}$

We vary the number of organizations in the setup of Experiment 1 and show the results in Figures 3-5. TCA curves move closer to uniform random auditing, but still have low mis-selection rates as the report gap grows.

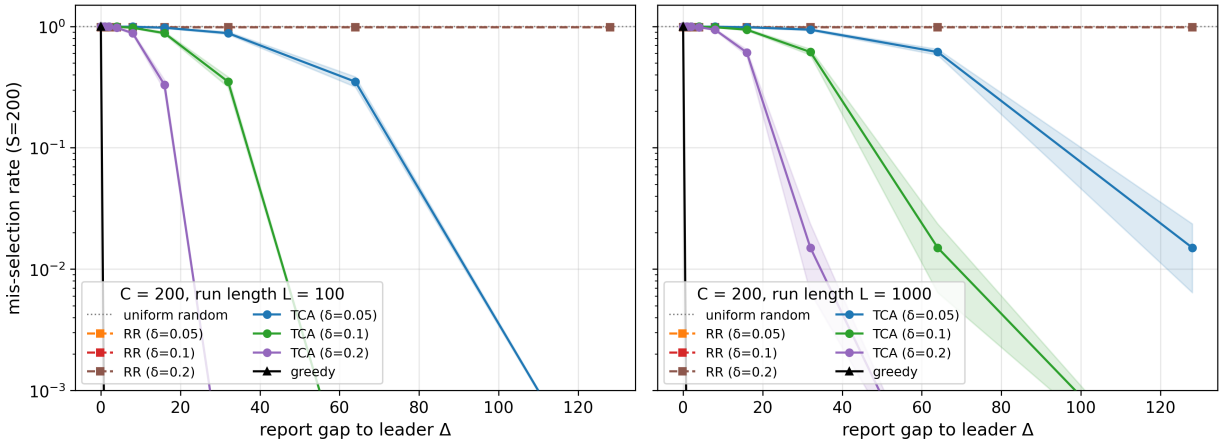


Figure 5. **Mis-selection Rate vs. Leading Gap (C=200).**