Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Quantifying the Cybersecurity-Privacy Trade-Off

Ali Mohammadi Ruzbahani

ali.mohammadiruzbaha@ucalgary.ca

University of Calgary https://orcid.org/0009-0004-3287-552X

Research Article

Keywords: Cybersecurity, privacy, differential privacy, hypothesis testing, intrusion detection, security monitoring, trade-off

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11733801/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Quantifying the Cybersecurity-Privacy Trade-Off

Ali Mohammadi Ruzbahani

***Abstract*—Security monitoring needs visibility into user behavior, which is a privacy risk for the people observed. Existing trade-off formulas cover one noise mechanism each and ignore the privacy lost when an attack goes undetected. Detecting an attack and inferring a person's behavior are hypothesis tests on the same monitoring record, so its visibility μ, a Gaussian differential privacy parameter, bounds the power of both. An attack of salience κ, its telemetry footprint in privacy units, is missed with probability at least $\Phi(z_\alpha - \kappa\mu)$ at false-alarm budget α, where Φ is the normal distribution function and $z_\alpha = \Phi^{-1}(1-\alpha)$. Privacy is lost through the record and through the breaches that monitoring fails to stop. With both channels the security-privacy frontier has a closed form, and the sign of its slope, the exchange rate, separates conflict from complementarity. A pressure ratio and a critical salience classify every regime. Privacy alone calls for monitoring exactly when the ratio exceeds a threshold, which equals one for an attack the size of one privacy unit. Above it the optimal visibility is explicit; custodial trust and breach severity enter it through a logarithm. Against an attacker who chooses its own salience, the worst-case undetected impact is at least an explicit constant times $\mu^{-\gamma}$, where γ is the elasticity of impact to salience. Gaussian monitoring attains the bound, and the attacker is then caught with a probability independent of μ. Simulations of Gaussian, Laplace and logistic mechanisms agree with the closed forms to within sampling error.**

***Index Terms*—Cybersecurity, privacy, differential privacy, hypothesis testing, intrusion detection, security monitoring, trade-off.**

I. Introduction

DEFENDERS detect attacks by reading telemetry, and the people whose behavior it records bear the privacy cost. End-to-end encryption, encrypted DNS, secure aggregation and differential privacy remove visibility on purpose. Client-side scanning and traffic interception put it back. Studies document both the loss of detection under encryption and the surveillance risk of scanning [1]–[4], but they do not say how much privacy a given gain in detection costs, or whether it costs any. The General Data Protection Regulation treats security processing as a legitimate interest "to the extent strictly necessary and proportionate" [5, Recital 49] and gives no way to compute either property. Existing formulas price detection in privacy for one noise mechanism at a time and count only the privacy lost in the released data [6]–[8]. Yet a missed attack also exposes data, and risk frameworks list such incidents as a separate source of privacy risk [9], [10]. Without the second source, the reported case in which more privacy improves detection is traced to a sub-optimal detector [11].

We solve the problem by putting both sides on one scale. The defender and a privacy adversary read the same monitoring record: the defender tests attack against no attack, and the adversary tests one behavior of a person against another. The visibility μ of the record, defined through Gaussian differential privacy [12], limits how well either test can do. Detection power therefore has a price in privacy loss. The breach channel depends on the same quantity, because the probability that an attack is missed falls as visibility grows. With both channels expressed through μ, the question of how much privacy a gain in detection costs has a numerical answer at every operating point, and so does the question of how much monitoring is proportionate. The inputs are the salience κ of the attack (how far it moves telemetry, counted in privacy units), the false-alarm budget, the attack probability, and three qualitative parameters: custodial trust, data sensitivity and breach severity.

The contributions are the following.

- We define a privacy risk with two channels, the leakage of the monitoring record and the privacy lost in undetected breaches, and express both through the visibility of the record.
- We bound the probability of detecting an attack of salience κ at visibility μ and a fixed false-alarm budget by a function of the product $\kappa\mu$. The bound follows from the group-privacy property of Gaussian differential privacy [12]. Gaussian noise attains it at every salience with one detector, and among additive noise laws with an even log-concave density it is the only one that does.
- We derive the security-privacy frontier in closed form and classify the sign of the exchange rate, the slope of the frontier, into five regimes.
- We show that monitoring lowers the privacy risk itself exactly when a pressure ratio of the parameters exceeds a threshold, and we give the optimal visibility for every weight on security risk. The optima for zero weight and for the decision maker's weight bound a proportionality interval.
- We prove a power law for attackers who choose their own salience: no monitoring system of visibility μ holds the worst-case undetected impact below a constant times $\mu^{-\gamma}$, where γ measures how impact scales with salience, and Gaussian monitoring attains this value. The probability of catching such an attacker does not depend on μ.
- We give scaling rules for the choice of privacy unit, for attacks of limited duration under user-level protection, and for a benign baseline that the defender cannot predict.

II. Related work

Table I maps the evidence on how security and privacy interact to the model quantity that each body of work informs.

End-to-end encryption removes the provider's view of message content. Client-side scanning restores part of that view on the user's device, and Abelson et al. argue that it creates

The author is with MetaKet INC., Calgary, Alberta, Canada, and with the Smart Cyber-Physical Systems (SCPS) Lab, University of Calgary, Calgary, Alberta, Canada (e-mail: ali.mohammadiruzbaha@ucalgary.ca).

TABLE I
EVIDENCE AND MODEL QUANTITIES

Setting	Sources	Model quantity
Client-side scanning of encrypted messaging	[1], [2], [13]	μ, α, chosen κ
Encrypted transport and DNS	[3]	statistic q
Secure aggregation in federated learning	[4], [14]	privacy unit, κ
Differentially private sensor data	[7], [8]	detection ceiling
Fisher-information privacy	[6]	power law
Interconnected control systems	[11]	detector optimality
Machine-learning defenses	[15]	sign of χ
Private and corruption-tolerant estimation	[16]	bounded influence
Enterprise endpoint monitoring	[17]	τ
Risk and legal frameworks	[5], [9], [10]	two channels, w
Security operations	[18]	α

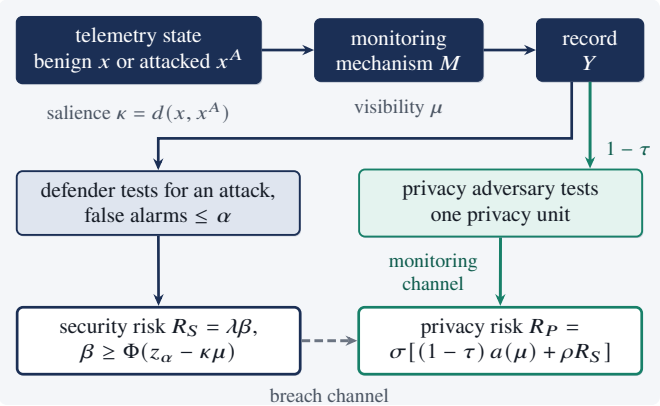


Fig. 1. Monitoring model.

a bulk surveillance capability with failure modes of its own [1]. Scheffler and Mayer systematize the designs proposed for detecting prohibited content under encryption and treat the privacy of clients and the secrecy of the server's detection rules as separate properties [2]. Jain et al. measured the detection side of one such design: small perturbations let more than 99.9% of images evade perceptual-hash matching, and thresholds large enough to make the evasion harder would probably flag more than a billion images a day [13]. Encrypted transport has the same effect on network operators. Grubbs et al. describe a decades-old conflict between user privacy and network administration and resolve one instance of it with zero-knowledge proofs that encrypted traffic complies with a policy [3]. In federated learning, secure aggregation shows the server only the sum of client updates, which makes defenses against poisoned updates harder to implement [4]; EIFFeL regains a bounded form of auditing by verifying that each update is well formed without revealing it [14].

Duddu et al. catalog cases in which a machine-learning defense against one risk raises or lowers the susceptibility to another [15]. In statistical estimation the two goals agree: Hopkins et al. give a black-box reduction that turns estimators tolerant of corrupted samples into differentially private ones [16]. Both properties cap the influence that a few data points can have on the output.

Four earlier works express a security-privacy relation in formulas. Farokhi and Mohajerin Esfahani take the inverse trace of the Fisher information as privacy and the local sensitivity of a Kullback-Leibler divergence to an injected bias as security, and prove for scalar queries that their product is constant [6]. Giraldo et al. show that an attacker can hide false data in differential-privacy noise; for a given noise density they derive the stealth-constrained attack of maximal bias and the matching saddle-point detector [7]. Ünsal and Önen give the power of the Neyman-Pearson test against one injected data point under the Gaussian mechanism [8]. Katewa et al. study interconnected dynamical systems and find attacks that the subsystems detect better when they share less information, an effect they trace to the sub-optimality of the chi-squared test [11]. The first two measure detectability by a divergence, all four concern additive noise on released data, and none counts the privacy lost in breaches. The theory below bounds detection power at a fixed false-alarm budget for every mechanism of a given visibility, and its breach term produces complementarity under an optimal detector.

On the privacy side we adopt the hypothesis-testing view. A mechanism is f-differentially private when no test separates neighboring inputs better than the trade-off function f allows, and Gaussian differential privacy (GDP) is the one-parameter family that arises in the limit of composition [12]. Audits estimate the GDP parameter of a deployed mechanism from the error rates of an attack [19], noise can be calibrated directly to a target attack risk [20], and NIST guidance interprets privacy guarantees through hypothesis tests [21].

The qualitative parameters of the model have sources in policy and in user studies. The NIST Privacy Framework separates privacy risks that arise from data processing from those that arise from cybersecurity incidents [9], and Pozen observes that cybersecurity "reduces certain sorts of privacy risks even as it generates others" [10]. These are the two channels of Section III. Contextual integrity holds that an information flow is acceptable when it matches the norms of its context [22], and Citron and Solove give a typology of privacy harms [23]. Employees whose devices run enterprise security software report that unclear collection practices erode their trust in the employer [17]. Alahmadi et al. find that analysts in security operations centers describe alarm volumes as overwhelming [18].

III. MODEL

Fig. 1 shows the model and Table II lists its parameters. A defender observes telemetry through a randomized mechanism and tests for an attack, a privacy adversary may hold the same record, and a missed attack becomes a breach.

A. Monitoring

Let $(\mathcal{X}, d)$ be a metric space. A point $x \in \mathcal{X}$ is a telemetry state: everything the monitored system would reveal about its users during one assessment period to an observer with unrestricted access. The metric is normalized by the privacy unit: two states that differ only in the behavior of one unit are

at distance at most one. The unit is an event, a user-day or a user, according to what the organization decides to protect [21].

A monitoring mechanism is a Markov kernel M from $\mathcal{X}$ to an output space $\mathcal{Y}$. The defender, and any other party that holds the monitoring record, sees $Y \sim M(x)$ and nothing else. Sampling, aggregation, truncation and noise applied before storage are all part of M.

For distributions P and Q on $\mathcal{Y}$ and a test $\psi : \mathcal{Y} \to [0,1]$, let $\alpha_\psi = \mathbb{E}_P\psi$ and $\beta_\psi = 1-\mathbb{E}_Q\psi$ be the two error probabilities. The trade-off function of the pair is $T(P,Q)(\alpha) = \inf\{\beta_\psi : \alpha_\psi \le \alpha\}$ [12, Def. 2]; it is convex and nonincreasing [12, Prop. 1]. For two unit-variance normal laws whose means differ by $\mu \ge 0$ it equals

$$G_\mu(\alpha) = \Phi\big(\Phi^{-1}(1-\alpha) - \mu\big), \tag{1}$$

where Φ is the standard normal distribution function and φ its density. For each $\alpha \in (0,1)$, $G_\mu(\alpha)$ is decreasing in μ.

Definition 1 (Visibility)**.** A mechanism M has visibility at most μ if

$$T\big(M(x), M(x')\big) \;\ge\; G_{\mu\, d(x,x')} \qquad \text{for all } x, x' \in \mathcal{X}.$$

The visibility $\mu(M)$ of M is the least such μ.

When d counts the privacy units in which two states differ, a mechanism has visibility at most μ exactly when it is μ-GDP, because a μ-GDP mechanism is $k\mu$-GDP for groups of k units [12, Th. 3]. For a general metric the definition is the Gaussian analogue of $d_\mathcal{X}$-privacy [24]. The family G_μ is the limit of the cumulative privacy loss of many small releases, which is what continuous monitoring produces [12, Th. 6]. It also excludes mechanisms that expose a unit outright with small probability, which a bound on total variation alone permits. A mechanism that stores telemetry without randomization has infinite visibility.

The reference mechanism adds Gaussian noise to a statistic. Let $q : \mathcal{X} \to \mathbb{R}$ be 1-Lipschitz and Z standard normal, and put

$$M_\mu(x) = \mu\, q(x) + Z. \tag{2}$$

Then $T(M_\mu(x), M_\mu(x')) = G_{\mu|q(x)-q(x')|} \ge G_{\mu\, d(x,x')}$, so the visibility of M_μ is at most μ, and it equals μ when q has Lipschitz constant one.

Lemma 1 (Leakage of the record)**.** *Let M have visibility at most μ. For all x, x' with $d(x,x') \le 1$,*

$$\mathrm{TV}\big(M(x), M(x')\big) \le a(\mu) := 2\Phi(\mu/2) - 1, \tag{3}$$

with equality for M_μ when $|q(x) - q(x')| = 1$.

Proof: For any P and Q, $1 - \mathrm{TV}(P,Q) = \inf_\psi(\alpha_\psi + \beta_\psi) = \min_\alpha\{\alpha + T(P,Q)(\alpha)\}$. With $P = M(x)$ and $Q = M(x')$ we have $T(P,Q) \ge G_{\mu d(x,x')} \ge G_\mu$, hence $1 - \mathrm{TV}(P,Q) \ge \min_\alpha\{\alpha + G_\mu(\alpha)\}$. Put $z = \Phi^{-1}(1-\alpha)$. The bracket equals $1 - \Phi(z) + \Phi(z-\mu)$, and its derivative in z, $\varphi(z-\mu) - \varphi(z)$, is negative for $z < \mu/2$ and positive for $z > \mu/2$. The minimum is therefore $2\Phi(-\mu/2)$, which gives (3). The trade-off function of the Gaussian pair is exactly G_μ, so equality holds for it. ■

The total variation distance in (3) is the largest advantage, true-positive rate minus false-positive rate, of any test that tries to tell from the record how one unit behaved. Advantage is one of the attack risks to which privacy noise is calibrated [20].

An attack replaces the benign state x by a state x^A. Its salience is $\kappa = d(x, x^A)$, the displacement that it causes in telemetry, measured in privacy units. Suppose, for example, that the unit is the account and that the record is built from the number of accounts with more than ten failed logins in the period, a statistic that one account changes by at most one. A password-spraying campaign that pushes k accounts over the threshold has salience k. An attack confined to the data of one unit has salience at most one, however abnormal the behavior of that unit becomes, and salience below one describes activity smaller than the variation that the privacy unit is entitled to hide. A statistic q is matched to the attack when $q(x^A) - q(x) = d(x, x^A)$, as the count is to the campaign.

A detector is a test ψ on the record. It raises a false alarm with probability $\alpha_\psi = \mathbb{E}_{M(x)}\psi$ and misses the attack with probability $\beta_\psi = 1 - \mathbb{E}_{M(x^A)}\psi$. The defender works under a false-alarm budget $\alpha \in (0, 1/2)$, an input of the model that is set by the number of alarms the analysts can review. Write $z_\alpha = \Phi^{-1}(1-\alpha) > 0$. Until Section VI the defender knows the benign value $q(x)$ of the monitored statistic, for example from a baseline learned before the period, so that the noise of M is the only source of missed detections.

B. Risk

An attack of the class under study occurs during the assessment period with probability $\lambda \in (0,1]$ and succeeds when it is missed. The security risk is

$$R_S = \lambda\, \beta_\psi. \tag{4}$$

If only a fraction of the detected attacks can be stopped, λ is multiplied by that fraction and nothing else changes.

A privacy unit can be harmed through two channels, the processing source and the incident source of privacy risk in [9], [10]. In the monitoring channel the record is used outside its defensive purpose, by function creep, insider misuse, compelled disclosure or theft of the record. Let $1 - \tau$ be the probability of such use during the period; $\tau \in [0,1)$ is the custodial trust. By Lemma 1, a party that holds the record gains an advantage of at most $a(\mu)$ about any unit. In the breach channel a missed attack exposes raw data. Let $\rho > 0$ be the breach severity: the expected harm to a unit from one missed attack, relative to a full exposure through the monitoring channel. The expectation includes the probability that the data of the unit are among those exposed. With $\sigma > 0$ the data sensitivity, that is, the harm per unit of advantage, the privacy risk of a mechanism with visibility μ is

$$R_P = \sigma\big[(1-\tau)\, a(\mu) + \rho\, R_S\big]. \tag{5}$$

The risk charges every unit the largest advantage $a(\mu)$ that the visibility of M permits, whether or not M attains it for that unit, as in privacy accounting. When $\rho \le 1$, the bracket in (5) is the union bound on the advantage of an adversary who can use either channel.

TABLE II
PARAMETERS

Symbol	Meaning	Source
Quantitative		
μ	visibility of the record	mechanism design or audit
κ	salience of the attack	displacement of a matched statistic
α	false-alarm budget	analyst capacity
λ	attack probability per period	incident history
γ	elasticity of impact to salience	attack class
δ	spread of the benign baseline	history of the statistic
Qualitative		
τ	custodial trust	governance of the record
σ	data sensitivity	data category and context
ρ	breach severity	harm assessment
w	weight of security risk	decision maker
d	privacy unit (distance one)	protection policy

Custodial trust interpolates between two readings of privacy. At $\tau = 0$ every collection counts as a loss, which is the secrecy view. As $\tau \to 1$, a flow that stays within the norms of its context costs nothing, which is the position of contextual integrity [22]. Data sensitivity σ is set by the type of harm at stake, for which Citron and Solove distinguish seven types of varying severity [23]; we give the highest values to the special categories of personal data [5, Art. 9]. The review of alarms by analysts exposes data as well. Reviews of false alarms add a constant proportional to α, and reviews of detected attacks turn ρ into the harm of a breach net of the harm of the investigation, which we take to be positive, so the form of (5) is unchanged.

IV. The security-privacy frontier

This section derives the lowest privacy risk compatible with a given security risk when the attack has a fixed salience $\kappa > 0$.

A. Detection ceiling

Proposition 1 (Detection ceiling)**.** *Let M have visibility at most μ and let the attack have salience at most κ. (i) Every detector with false-alarm probability at most α misses the attack with probability*

$$\beta_\psi \;\geq\; \Phi(z_\alpha - \kappa\mu). \tag{6}$$

(ii) The bound also holds for the average miss probability when the benign state x is drawn from an arbitrary distribution on $\mathcal{X}$, each x is paired with an attacked state at distance at most κ, and the false-alarm constraint holds on average. (iii) For the mechanism M_μ of (2) *with a statistic matched to an attack of salience κ, the detector that raises an alarm when $Y > \mu q(x) + z_\alpha$ has false-alarm probability α and miss probability $\Phi(z_\alpha - \kappa\mu)$, for every $\kappa \geq 0$.*

Proof: (i) By the definition of the trade-off function, $\beta_\psi \geq T(M(x), M(x^A))(\alpha_\psi)$. Visibility at most μ gives $T(M(x), M(x^A)) \geq G_{\mu d(x,x^A)} \geq G_{\kappa\mu}$. Since $G_{\kappa\mu}$ is nonincreasing, $\beta_\psi \geq G_{\kappa\mu}(\alpha_\psi) \geq G_{\kappa\mu}(\alpha)$, which is (6).

(ii) Let α_x and β_x be the error probabilities of ψ at the benign state x and its attacked state, and let $\mathbb{E}$ be the average over the benign state, so that $\mathbb{E}\alpha_x \leq \alpha$. By part (i), $\beta_x \geq G_{\kappa\mu}(\alpha_x)$. Jensen's inequality for the convex function $G_{\kappa\mu}$ gives $\mathbb{E}\beta_x \geq G_{\kappa\mu}(\mathbb{E}\alpha_x) \geq G_{\kappa\mu}(\alpha)$.

(iii) Under the benign state, $Y - \mu q(x)$ is standard normal and the alarm probability is $1 - \Phi(z_\alpha) = \alpha$. Under the attack, $Y - \mu q(x)$ is normal with mean $\kappa\mu$ and unit variance, and the miss probability is $\Phi(z_\alpha - \kappa\mu)$. The threshold does not involve κ. ∎

Part (i) is group privacy [12, Th. 3] applied to the attacked state: a guarantee that hides a group of κ units hides an attack of salience κ equally well. The Gaussian case with a single injected data point is in [8]. By part (ii), ignorance of the benign baseline never helps the defender, so (6) limits every monitoring system of visibility μ, and by part (iii) the limit is attained without knowledge of the salience.

B. Frontier

Call a pair (M, ψ) admissible if M has finite visibility and $\alpha_\psi \leq \alpha$, and write $R_S(M, \psi)$ and $R_P(M, \psi)$ for its risks (4) and (5).

Theorem 1 (Frontier)**.** *For $s \in (0, \lambda(1-\alpha)]$ let*

$$m(s) = \frac{z_\alpha - \Phi^{-1}(s/\lambda)}{\kappa}, \tag{7}$$

$$\mathcal{F}(s) = \sigma\big[(1-\tau)\, a\big(m(s)\big) + \rho s\big]. \tag{8}$$

(i) Every admissible pair with $R_S(M, \psi) = s$ has $R_P(M, \psi) \geq \mathcal{F}(s)$. (ii) Equality holds for the mechanism M_μ with $\mu = m(s)$, a matched statistic of Lipschitz constant one, and the threshold detector of Proposition 1. (iii) Along the frontier the exchange rate $\chi = -d\mathcal{F}/ds$ is, at $\mu = m(s)$,

$$\chi(\mu) = \sigma\Big[\frac{1-\tau}{\lambda\kappa}\, e^{\eta(\mu)} - \rho\Big], \quad \eta(\mu) = \tfrac{1}{2}(z_\alpha - \kappa\mu)^2 - \tfrac{1}{8}\mu^2. \tag{9}$$

Proof: (i) Let μ be the visibility of M. By Proposition 1, $s/\lambda = \beta_\psi \geq \Phi(z_\alpha - \kappa\mu)$, hence $\mu \geq m(s)$, and $m(s) \geq 0$ because $s \leq \lambda(1-\alpha)$. Since a is increasing, $R_P = \sigma[(1-\tau)a(\mu) + \rho s] \geq \mathcal{F}(s)$.

(ii) By Proposition 1(iii) the pair has miss probability $\Phi(z_\alpha - \kappa m(s)) = s/\lambda$, and its visibility is $m(s)$.

(iii) We have $a'(\mu) = \varphi(\mu/2)$ and $m'(s) = -1/[\kappa\lambda\,\varphi(z_\alpha - \kappa m(s))]$. Therefore $\mathcal{F}'(s) = \sigma[\rho - (1-\tau)\varphi(m/2)/(\kappa\lambda\,\varphi(z_\alpha - \kappa m))]$, and $\varphi(m/2)/\varphi(z_\alpha - \kappa m) = e^{\eta(m)}$. ∎

The frontier (8) is the trade-off formula: the least privacy risk, charged as in (5), that any monitoring system attains at security risk s. The exchange rate (9) is the privacy risk paid for one unit of security risk removed. Its first term is the price of visibility, the leakage of the record incurred per unit of security risk removed. Its second term is the breach harm that the same unit avoids. When the second term exceeds the first, $\chi < 0$, and lowering the security risk lowers the privacy risk as well.

Fig. 2 shows $\mathcal{F}$ for $\kappa \in \{0.3, 1, 3, 10\}$ with $\alpha = 0.01$, $\lambda = 0.3$, $\tau = 0.5$ and $\rho = \sigma = 1$, the values used in all figures unless stated otherwise. The horizontal axis is the miss probability R_S/λ on a normal-quantile scale, along which visibility grows linearly from right to left. All curves leave the point of no monitoring, $\mu = 0$, and tend to $R_P = \sigma(1-\tau)$ as $R_S \to 0$. A point is efficient if no point with lower security

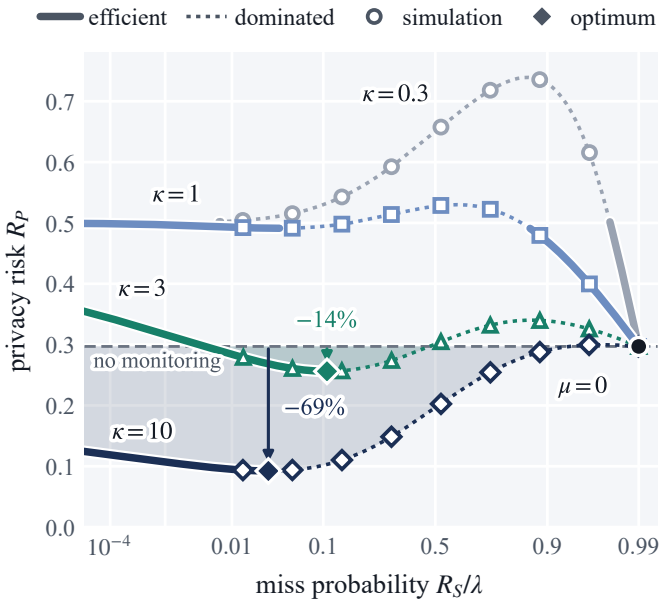


Fig. 2. Security-privacy frontiers.

risk has lower or equal privacy risk. On the dotted arcs, which are dominated, some larger visibility lowers both risks. Diamonds mark the interior privacy optima, and the shaded areas show the privacy risk that monitoring removes relative to no monitoring: 14% for $\kappa = 3$ and 69% for $\kappa = 10$.

C. Regimes

The sign of χ depends on the parameters only through κ, z_α and the pressure ratio

$$\Lambda = \frac{\lambda\rho\kappa}{1-\tau}, \tag{10}$$

the breach pressure $\lambda\rho\kappa$ divided by the custodial distrust $1-\tau$: by (9), $\chi(\mu) > 0$ exactly when $\eta(\mu) > \ln\Lambda$.

Theorem 2 (Regimes)**.** *Let $\ell = \ln\Lambda$ and $\Delta = z_\alpha^2/4 + (2\kappa^2 - \frac{1}{2})\ell$, and where the expressions are real put*

$$\mu_1 = \frac{z_\alpha^2 - 2\ell}{\kappa z_\alpha + \sqrt{\Delta}}, \qquad \mu_2 = \frac{\kappa z_\alpha + \sqrt{\Delta}}{\kappa^2 - \frac{1}{4}}. \tag{11}$$

The sign of the exchange rate on $(0,\infty)$ is the one listed in Table III. In particular χ changes sign only at μ_1 and, for $\kappa > 1/2$, at μ_2, and for $\kappa \le 1/2$ conflict does not return once complementarity has set in.

Proof: Let $p(\mu) = (\kappa^2 - \frac{1}{4})\mu^2 - 2\kappa z_\alpha\mu + z_\alpha^2 - 2\ell$, so that $2(\eta(\mu) - \ell) = p(\mu)$ and $\chi(\mu)$ has the sign of $p(\mu)$. Here $p(0) = z_\alpha^2 - 2\ell$, and $\kappa^2 z_\alpha^2 - (\kappa^2 - \frac{1}{4})(z_\alpha^2 - 2\ell) = \Delta$.

Let $\kappa > 1/2$. Then p is a convex parabola with vertex at $\kappa z_\alpha/(\kappa^2 - \frac{1}{4}) > 0$. When $\Delta \le 0$, that is, when $\ell \le -z_\alpha^2/(8\kappa^2 - 2)$, it does not change sign and $p \ge 0$, with a single zero when $\Delta = 0$; this is case (a). When $\Delta > 0$ its roots are $(\kappa z_\alpha \mp \sqrt{\Delta})/(\kappa^2 - \frac{1}{4})$. The larger one is $\mu_2 > 0$. The smaller one is positive exactly when $p(0) > 0$, that is, when $\ell < z_\alpha^2/2$, and multiplying its numerator and denominator by $\kappa z_\alpha + \sqrt{\Delta}$ gives μ_1; this is case (b). When $\ell \ge z_\alpha^2/2$ we have $\Delta \ge \kappa^2 z_\alpha^2 > 0$ and $p(0) \le 0$, so $p < 0$ on $(0, \mu_2)$ and $p > 0$ beyond; this is case (c).

Let $\kappa \le 1/2$. Then $p'(\mu) = 2(\kappa^2 - \frac{1}{4})\mu - 2\kappa z_\alpha < 0$ for $\mu \ge 0$, and $p(\mu) \to -\infty$. Hence p has exactly one positive root when $p(0) > 0$ and none otherwise. For $\ell < z_\alpha^2/2$ one checks that $\Delta > \kappa^2 z_\alpha^2$ when $\kappa < 1/2$ and $\sqrt{\Delta} = z_\alpha/2$ when $\kappa = 1/2$, and in both cases the root equals μ_1. These are cases (d) and (e). ∎

TABLE III
REGIMES OF THE EXCHANGE RATE

Case	Salience	Pressure $\ell = \ln\Lambda$	Sign of χ as μ grows
(a)	$\kappa > \frac{1}{2}$	$\ell \le -\frac{z_\alpha^2}{8\kappa^2 - 2}$	+
(b)	$\kappa > \frac{1}{2}$	$-\frac{z_\alpha^2}{8\kappa^2 - 2} < \ell < \frac{z_\alpha^2}{2}$	+, − on (μ_1, μ_2), +
(c)	$\kappa > \frac{1}{2}$	$\ell \ge \frac{z_\alpha^2}{2}$	− on $(0, \mu_2)$, then +
(d)	$\kappa \le \frac{1}{2}$	$\ell < \frac{z_\alpha^2}{2}$	+ on $(0, \mu_1)$, then −
(e)	$\kappa \le \frac{1}{2}$	$\ell \ge \frac{z_\alpha^2}{2}$	−

Fig. 3 maps the five cases; the solid curve bounds case (a), and the shading is the fraction of the privacy risk that optimal monitoring removes, $1 - \inf_\mu R_P(\mu)/R_P(0)$. Two thresholds organize the map. The first is $\Lambda = e^{z_\alpha^2/2}$: above it, the first increment of visibility already lowers the privacy risk. The second is the critical salience $\kappa = 1/2$. The privacy adversary of Lemma 1 balances its two errors, so the marginal leakage $a'(\mu) = \varphi(\mu/2)$ decays at the rate set by $\mu/2$, and the marginal gain in detection decays at the rate set by $\kappa\mu$. For $\kappa > 1/2$ the price of visibility in (9) grows without bound and conflict always returns at high visibility. For $\kappa \le 1/2$ it vanishes, and the attack cannot be detected before the record has revealed nearly all it can about each unit: at $\alpha = 0.01$ a detection probability of one half requires $\mu \ge 4.65$, where $a(\mu) \ge 0.98$.

The value $1/2$ belongs to the balanced adversary. If leakage is measured by the power of an adversary held to a false-positive rate α_p, the second risk measure of [20], then $a(\mu)$ is replaced by $\Phi(\mu - z_{\alpha_p}) - \alpha_p$, the exponent in (9) becomes $\frac{1}{2}(z_\alpha - \kappa\mu)^2 - \frac{1}{2}(\mu - z_{\alpha_p})^2$, and the critical salience is one. For $\kappa = 1$ and $\alpha_p = \alpha$ the exchange rate is then the constant $\sigma\rho(1/\Lambda - 1)$.

A decision maker who puts weight $w \ge 0$ on security risk minimizes $J_w = wR_S + R_P$. Along the frontier $J_w(\mu) = \sigma[(1-\tau)a(\mu) + (\rho + w/\sigma)\lambda\Phi(z_\alpha - \kappa\mu)]$, which is the privacy risk with ρ replaced by $\rho + w/\sigma$. Theorem 2 therefore applies to J_w with

$$\Lambda_w = \frac{\lambda\kappa(\rho + w/\sigma)}{1-\tau} \tag{12}$$

in place of Λ; write Δ_w for the corresponding value of Δ. Because σ multiplies both channels, it does not affect the privacy-optimal visibility and enters the weighted optimum only through w/σ.

Proposition 2 (Monitoring threshold)**.** *Let $D(\mu) = \Phi(z_\alpha) - \Phi(z_\alpha - \kappa\mu)$ be the gain in detection probability over no*

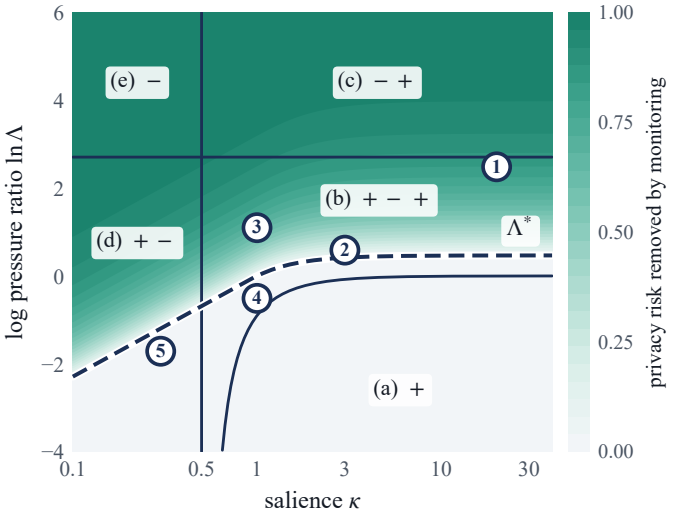


Fig. 3. Regime map at $\alpha = 0.01$ with the scenarios of Table V.

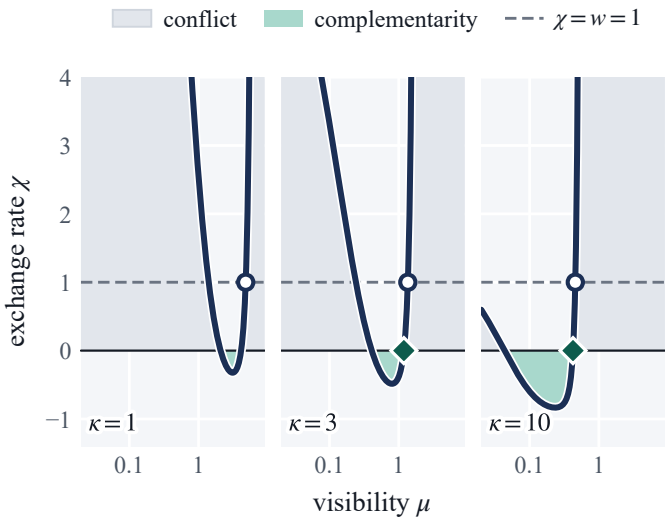


Fig. 4. Exchange rate against visibility.

monitoring, and

$$\Lambda^*(\kappa) = \kappa \inf_{\mu>0} \frac{a(\mu)}{D(\mu)}. \tag{13}$$

(i) Some visibility gives a lower weighted risk than no monitoring, that is, $\inf_{\mu>0} J_w(\mu) < J_w(0)$*, if and only if* $\Lambda_w > \Lambda^*(\kappa)$*. (ii)* Λ^* *is nondecreasing in* κ*, with* $\Lambda^*(\kappa) = \kappa/(1-\alpha)$ *for* $\kappa \le 1/2$*,* $\Lambda^*(1) = 1$ *for every* α*, and* $\Lambda^*(\kappa) \le (2\pi)^{-1/2} \min_{t>0} t/[\Phi(z_\alpha) - \Phi(z_\alpha - t)]$ *for all* κ*.*

Proof: (i) The miss probability without monitoring is $1 - \alpha = \Phi(z_\alpha)$, so $J_w(\mu) - J_w(0) = \sigma[(1-\tau)a(\mu) - (\rho + w/\sigma)\lambda D(\mu)]$, which is negative exactly when $\Lambda_w > \kappa a(\mu)/D(\mu)$.

(ii) Write $t = \kappa\mu$, so that $\kappa a(\mu)/D(\mu) = \kappa a(t/\kappa)/[\Phi(z_\alpha) - \Phi(z_\alpha - t)]$. Because a is concave with $a(0) = 0$, $\kappa a(t/\kappa)$ is nondecreasing in κ and at most $t\,a'(0) = t/\sqrt{2\pi}$, which gives the monotonicity and the upper bound. For $\kappa = 1$, $D(\mu)$ is the standard normal probability of an interval of length μ, which is at most $a(\mu)$, with equality when the interval is centered at zero, that is, when $\mu = 2z_\alpha$; hence the infimum in (13) is one. For $\kappa \le 1/2$ we have $D(\mu) \le \Phi(z_\alpha) - \Phi(z_\alpha - u)$ with $u = \mu/2$. The ratio $\Phi(z_\alpha - u)/\Phi(-u)$ is increasing in $u \ge 0$ because φ/Φ is decreasing, and it equals $2\Phi(z_\alpha)$ at $u = 0$, so $\Phi(z_\alpha - u) > 2\Phi(z_\alpha)\Phi(-u)$ for $u > 0$. This is the inequality $D(\mu) < (1-\alpha)a(\mu)$, and $a(\mu)/D(\mu) \to 1/(1-\alpha)$ as $\mu \to \infty$. ∎

At $\alpha = 0.01$ the upper bound in part (ii) is 1.60, and $\Lambda^*(3) = 1.52$ and $\Lambda^*(10) = 1.59$ are close to it. The dashed curve in Fig. 3 is $\ln \Lambda^*$, and the shading is positive exactly above it. For $\kappa = 1$ and $w = 0$ the condition $\Lambda_w > \Lambda^*(\kappa)$ reads $\lambda\rho > 1 - \tau$: the breach pressure exceeds the custodial distrust.

Corollary 1 (Optimal visibility)**.** *Let* $\kappa > 1/2$*. If* $\Lambda_w > \Lambda^*(\kappa)$*, the weighted risk* J_w *has the unique minimizer*

$$\mu^\circ(w) = \frac{\kappa z_\alpha + \sqrt{z_\alpha^2/4 + (2\kappa^2 - \tfrac{1}{2}) \ln \Lambda_w}}{\kappa^2 - \tfrac{1}{4}}, \tag{14}$$

and $d\mu^\circ / d \ln \Lambda_w = \Delta_w^{-1/2}$*. If* $\Lambda_w < \Lambda^*(\kappa)$*, the unique minimizer is* $\mu^\circ(w) = 0$*. Let* $\kappa \le 1/2$*. Then* $\inf J_w$ *is* $J_w(0)$ *if* $\Lambda_w \le \Lambda^*(\kappa)$*, and otherwise it is* $\sigma(1-\tau)$*, approached as* $\mu \to \infty$*.*

Proof: Let $\kappa > 1/2$. By Theorem 2 the only interior local minimum of J_w is at μ_2, computed with Λ_w, in cases (b) and (c), and J_w increases beyond it. In case (a) J_w is increasing. By Proposition 2, J_w takes values below $J_w(0)$ exactly when $\Lambda_w > \Lambda^*(\kappa)$, which excludes case (a) and makes μ_2 the minimizer; when $\Lambda_w < \Lambda^*(\kappa)$, $J_w(\mu) > J_w(0)$ for all $\mu > 0$. Differentiating (14) gives $(2\kappa^2 - \frac{1}{2})/[2\sqrt{\Delta_w}(\kappa^2 - \frac{1}{4})] = \Delta_w^{-1/2}$. For $\kappa \le 1/2$, J_w has no interior local minimum and tends to $\sigma(1-\tau)$, and the claim follows from Proposition 2. ∎

When $\Lambda > \Lambda^*(\kappa)$, a defender who monitors less than the privacy-optimal visibility $\mu^\circ(0)$ loses on both risks, so privacy protection itself sets a lower limit on monitoring. The weighted optimum $\mu^\circ(w)$ is an upper limit: above it the exchange rate exceeds the weight given to security. The interval $[\mu^\circ(0), \mu^\circ(w)]$ is what necessity and proportionality [5, Recital 49] amount to in the model, and we call it the proportionality interval.

The qualitative parameters act in two ways. Whether monitoring pays at all depends on the side of $\Lambda^*(\kappa)$ on which Λ_w falls, and at the crossing the optimum jumps between 0 and μ_2 when $\kappa > 1/2$, and between 0 and unbounded visibility otherwise. How much monitoring is optimal, once it pays, depends on them only through $\ln \Lambda_w$: a factor of two in λ, in $\rho + w/\sigma$ or in $1 - \tau$ moves $\mu^\circ(w)$ by about $(\ln 2)/\sqrt{\Delta_w}$, which is below $1.4/z_\alpha$ for $\kappa \ge 1$. In Fig. 4 the diamonds mark $\mu^\circ(0)$ for $\kappa \in \{3, 10\}$ and the circles mark $\mu^\circ(1)$, where the line $\chi = w$ meets the rising branch of each curve.

V. Strategic attackers

An attacker who knows how the defender monitors will choose how visible to be. Fix a benign state x and an attack pattern, a family $\{x_\kappa : 0 \le \kappa \le \bar{\kappa}\} \subset \mathcal{X}$ with $x_0 = x$ and $d(x_\kappa, x_{\kappa'}) = |\kappa - \kappa'|$. The attacker chooses the salience κ and

obtains impact κ^γ if the attack is missed, where $\gamma > 0$ is the elasticity of impact with respect to salience; $\gamma = 1$ describes impact proportional to salience, as in exfiltration at a constant rate. The defender commits to an admissible pair (M, ψ) with visibility at most μ. Let $\beta_{M,\psi}(\kappa)$ be the miss probability at x_κ and

$$V(M,\psi) = \sup_{0\le\kappa\le\bar\kappa} \kappa^\gamma \beta_{M,\psi}(\kappa) \tag{15}$$

the worst-case expected undetected impact.

Theorem 3 (Power law)**.** *Let $t_{\alpha,\gamma}$ be the unique positive solution of $t\,\varphi(z_\alpha - t) = \gamma\,\Phi(z_\alpha - t)$, let*

$$K_\gamma(\alpha) = t_{\alpha,\gamma}^{\gamma}\,\Phi(z_\alpha - t_{\alpha,\gamma}) = \max_{t\ge 0} t^\gamma \Phi(z_\alpha - t), \tag{16}$$

and suppose that $\mu \ge t_{\alpha,\gamma}/\bar\kappa$. (i) Every admissible pair with visibility at most μ has $V(M,\psi) \ge K_\gamma(\alpha)\,\mu^{-\gamma}$. (ii) Equality holds for M_μ with the statistic $q(x') = \inf_\kappa\{\kappa + d(x', x_\kappa)\}$ and the threshold detector ψ^* of Proposition 1. The attacker's best response is $\kappa^* = t_{\alpha,\gamma}/\mu$, and $\big((M_\mu,\psi^*),\kappa^*\big)$ is a saddle point of the zero-sum game with payoff $\kappa^\gamma\beta_{M,\psi}(\kappa)$. (iii) At the saddle point the attack is detected with probability $1 - \Phi(z_\alpha - t_{\alpha,\gamma})$, which does not depend on μ.

Proof: Let $r(u) = \Phi(u)/\varphi(u)$. Then $r'(u) = 1 + u\,r(u) = [\varphi(u) + u\Phi(u)]/\varphi(u)$, and $\varphi(u) + u\Phi(u) = \int_{-\infty}^{u}\Phi(v)\,dv > 0$, so r is increasing, with $r(u) \to 0$ as $u \to -\infty$. The function $g(t) = t^\gamma\Phi(z_\alpha - t)$ has $g'(t) = t^{\gamma-1}\varphi(z_\alpha - t)[\gamma\,r(z_\alpha - t) - t]$. The bracket is strictly decreasing in t, positive at $t = 0$ and negative for large t. Hence g has a unique stationary point $t_{\alpha,\gamma}$, which is its maximum on $[0,\infty)$, and $\max g = K_\gamma(\alpha)$.

(i) For every $\kappa \le \bar\kappa$, Proposition 1 gives $\kappa^\gamma\beta_{M,\psi}(\kappa) \ge \kappa^\gamma\Phi(z_\alpha - \kappa\mu) = \mu^{-\gamma}g(\kappa\mu)$. The choice $\kappa = t_{\alpha,\gamma}/\mu \le \bar\kappa$ yields the claim.

(ii) The function q is 1-Lipschitz as an infimum of 1-Lipschitz functions, and $q(x_\kappa) = \inf_{\kappa'}\{\kappa' + |\kappa - \kappa'|\} = \kappa$, so q is matched to every attack of the pattern. By Proposition 1(iii), $\beta_{M_\mu,\psi^*}(\kappa) = \Phi(z_\alpha - \kappa\mu)$ for all κ, so $\kappa^\gamma\beta_{M_\mu,\psi^*}(\kappa) = \mu^{-\gamma}g(\kappa\mu) \le K_\gamma(\alpha)\mu^{-\gamma}$ with equality at κ^*. Together with part (i) at $\kappa = \kappa^*$, the payoff at $\big((M_\mu,\psi^*),\kappa^*\big)$ is the largest over κ and the smallest over admissible pairs.

(iii) At κ^* the detection probability is $1 - \Phi(z_\alpha - \kappa^*\mu) = 1 - \Phi(z_\alpha - t_{\alpha,\gamma})$. ∎

That more privacy leaves more room for an undetected attack is known for Fisher-information privacy [6], for the maximal stealthy bias under a given noise density [7] and for the largest undetected shift of one data point under the Gaussian mechanism [8]. Theorem 3 holds for every mechanism of visibility μ and every elasticity, with an explicit constant and a saddle point: a 1% increase in visibility lowers the worst-case undetected impact by γ%. Part (iii) shows that added visibility does not raise the detection probability, because an optimizing attacker lowers its salience in proportion and is caught as often as before. With r as in the proof, the number $u = z_\alpha - t_{\alpha,\gamma}$ solves $z_\alpha = u + \gamma\,r(u)$, and $r(0) = \sqrt{\pi/2}$, so the detection probability at the saddle point is below one half exactly when $z_\alpha > \gamma\sqrt{\pi/2}$; for $\gamma = 1$ this is every budget $\alpha < 0.105$. As $\gamma \to 0$, $t_{\alpha,\gamma} \to 0$ and the detection probability falls to α: an attacker whose impact does not depend on its salience is caught no more often than a false alarm is raised. Table IV gives the constants and Fig. 6b the detection probability.

TABLE IV
ATTACKER CONSTANTS

α	$t_{\alpha,1}$	$K_{1/2}(\alpha)$	$K_1(\alpha)$	$K_2(\alpha)$	Detection, $\gamma = 1$
10^{-1}	1.267	0.615	0.641	0.979	0.494
5×10^{-2}	1.461	0.742	0.837	1.446	0.427
10^{-2}	1.874	0.975	1.264	2.711	0.325
10^{-3}	2.400	1.219	1.812	4.840	0.245
10^{-4}	2.873	1.405	2.302	7.228	0.199

In this model the security risk is the expected undetected impact, $R_S = \lambda K_\gamma(\alpha)\mu^{-\gamma}$, and ρ now measures the harm to a privacy unit per unit of impact. By (5) the privacy risk at the saddle point is

$$\Pi(\mu) = \sigma\big[(1-\tau)\,a(\mu) + \rho\lambda K_\gamma(\alpha)\,\mu^{-\gamma}\big], \qquad \mu \ge t_{\alpha,\gamma}/\bar\kappa. \tag{17}$$

For $\mu \le t_{\alpha,\gamma}/\bar\kappa$ the attacker uses the largest salience $\bar\kappa$, and the privacy risk is

$$\bar\Pi(\mu) = \sigma\big[(1-\tau)\,a(\mu) + \rho\lambda\bar\kappa^\gamma\,\Phi(z_\alpha - \bar\kappa\mu)\big], \tag{18}$$

which is (5) at salience $\bar\kappa$ with ρ replaced by $\rho\bar\kappa^\gamma$. Let W_0 and W_{-1} be the two real branches of the Lambert function, the inverse of $v \mapsto ve^v$.

Theorem 4 (Optimal visibility under strategic attack)**.** *Let*

$$\Gamma = \frac{\sqrt{2\pi}\,\gamma\rho\lambda K_\gamma(\alpha)}{1-\tau}, \qquad \xi = \frac{\Gamma^{2/(\gamma+1)}}{4(\gamma+1)}, \tag{19}$$

and treat Π in (17) as a function on $(0,\infty)$. (i) If $\xi \ge 1/e$, then Π is nonincreasing. (ii) If $\xi < 1/e$, then Π decreases on $(0,\mu^*)$, increases on $(\mu^*,\mu^\sharp)$ and decreases on $(\mu^\sharp,\infty)$, where

$$\mu^* = 2\sqrt{-(\gamma+1)W_0(-\xi)}, \quad \mu^\sharp = 2\sqrt{-(\gamma+1)W_{-1}(-\xi)}. \tag{20}$$

(iii) Let $\xi < 1/e$ and $\mu^* \ge t_{\alpha,\gamma}/\bar\kappa$. Then $\Pi(\mu^*) = \sigma(1-\tau)[a(\mu^*) + \mu^*\varphi(\mu^*/2)/\gamma]$, and μ^* minimizes the privacy risk over all $\mu \ge 0$, which is $\bar\Pi$ on $[0, t_{\alpha,\gamma}/\bar\kappa]$ and Π beyond, if and only if

$$\Pi(\mu^*) \le \sigma\min\big\{1-\tau,\ \rho\lambda\bar\kappa^\gamma(1-\alpha)\big\}. \tag{21}$$

(iv) As $\Gamma \to 0$, $\mu^* = \Gamma^{1/(\gamma+1)}\big(1 + \xi/2 + O(\xi^2)\big)$.

Proof: We have $\Pi'(\mu) = \sigma(1-\tau)(2\pi)^{-1/2}\mu^{-\gamma-1}[H(\mu) - \Gamma]$ with $H(\mu) = \mu^{\gamma+1}e^{-\mu^2/8}$. Since $H'(\mu) = \mu^\gamma e^{-\mu^2/8}[\gamma + 1 - \mu^2/4]$, H increases from 0 to its maximum at $\mu = 2\sqrt{\gamma+1}$ and then decreases to 0. Raising $H(\mu) = \Gamma$ to the power $2/(\gamma+1)$ and writing $v = \mu^2/[4(\gamma+1)]$ gives $ve^{-v} = \xi$, that is, $-v = W(-\xi)$. If $\xi > 1/e$ there is no solution and $\Pi' < 0$; if $\xi = 1/e$ the only solution is a double root. This proves (i). If $\xi < 1/e$ the two solutions are $v = -W_0(-\xi) \in (0,1)$ and $v = -W_{-1}(-\xi) > 1$, which are (20), and $H > \Gamma$ between them. This proves (ii).

(iii) The stationarity condition $H(\mu^*) = \Gamma$ reads $\rho\lambda K_\gamma(\alpha)(\mu^*)^{-\gamma} = (1-\tau)\mu^*\varphi(\mu^*/2)/\gamma$, which gives the value of $\Pi(\mu^*)$. On $[t_{\alpha,\gamma}/\bar\kappa,\infty)$ the infimum of Π is, by part (ii), the

smaller of $\Pi(\mu^*)$ and the limit $\sigma(1-\tau)$, which is not attained. On $[0, t_{\alpha,\gamma}/\bar{\kappa}]$ the risk is $\bar{\Pi}$, and $\bar{\Pi}(0) = \sigma\rho\lambda\bar{\kappa}^{\gamma}(1-\alpha)$. At $t_{\alpha,\gamma}/\bar{\kappa}$ the two branches have the same value and, by the equation that defines $t_{\alpha,\gamma}$, the same derivative, which is nonpositive because $t_{\alpha,\gamma}/\bar{\kappa} \le \mu^*$. The derivative of $\bar{\Pi}$ has the sign of an exchange rate of Section IV and follows one of the patterns of Table III, so on $(0, t_{\alpha,\gamma}/\bar{\kappa})$ it changes sign at most once, from positive to negative. Hence the minimum of $\bar{\Pi}$ on $[0, t_{\alpha,\gamma}/\bar{\kappa}]$ is $\bar{\Pi}(0)$ or $\Pi(t_{\alpha,\gamma}/\bar{\kappa}) \ge \Pi(\mu^*)$.

(iv) From $W_0(-\xi) = -\xi - \xi^2 + O(\xi^3)$ we get $(\mu^*)^2 = 4(\gamma+1)\xi(1+\xi+O(\xi^2)) = \Gamma^{2/(\gamma+1)}(1+\xi+O(\xi^2))$. ∎

The two bounds in (21) are the alternatives to μ^*. The first is the risk at unbounded visibility, and it is met when μ^* is at most 0.90, 1.50 and 2.38 for $\gamma = 1/2$, 1 and 2. The second is the risk without monitoring, when the attacker uses its largest salience. Part (iv) is a root law for the optimum. For $\gamma = 1$ the privacy-optimal visibility is $(2\pi)^{1/4}\sqrt{\rho\lambda K_1(\alpha)/(1-\tau)}$ to first order, so a fourfold increase of the attack probability doubles it, and so does a fourfold decrease of the custodial distrust. At the optimum, and to the same order, the monitoring channel carries γ times the risk carried by the breach channel, because the stationarity condition equates $(1-\tau)\mu^*\varphi(\mu^*/2)$ with γ times the breach term. For example, $\alpha = 0.01$, $\gamma = 1$, $\lambda = 0.3$, $\tau = 0.5$ and $\rho = 0.01$ give $\Gamma = 0.019$ and $\mu^* = 0.138$; the attacker responds with salience 13.6, and each channel carries a risk of $0.0275\,\sigma$. The sum is below the risk without monitoring, $0.00297\,\bar{\kappa}\,\sigma$, when $\bar{\kappa} \ge 18.6$. A weight w on security risk replaces ρ by $\rho + w/\sigma$ in (19) and (21).

VI. Extensions

This section changes the privacy unit, the noise law and the defender's knowledge of the benign baseline, one at a time.

Proposition 3 (Privacy unit and dwell time)**.** *Visibility and salience depend on the privacy unit as follows. (i) Replace d by d/c with $c > 0$. The visibility of every mechanism is multiplied by c and the salience of every attack is divided by c. The detection ceiling* (6) *is unchanged and the leakage $a(\mu)$ becomes $a(c\mu)$. (ii) Let monitoring run for N rounds, with state $x = (x_1, \dots, x_N)$ and independent mechanisms M_n of visibility at most μ_0 for metrics d_n. The joint mechanism has visibility at most μ_0 for $d(x,x') = (\sum_n d_n(x_n, x_n')^2)^{1/2}$. If the privacy unit is a user who may change every round by one event unit, the visibility is at most $\mu = \mu_0\sqrt{N}$, and an attack of salience κ_0 in each of N_A rounds has salience*

$$\kappa = \kappa_0\sqrt{N_A/N}. \tag{22}$$

Proof: (i) The identity $G_{\mu d} = G_{(c\mu)(d/c)}$ gives the first two claims, and $\kappa\mu$ is invariant.

(ii) The trade-off function of two product distributions is the tensor product of the trade-off functions of the factors, the tensor product is monotone, and $G_{\nu_1} \otimes G_{\nu_2} = G_{(\nu_1^2+\nu_2^2)^{1/2}}$ [12, Sec. 3.1]. Hence the joint trade-off function between x and x' is at least $\bigotimes_n G_{\mu_0 d_n(x_n,x_n')} = G_{\mu_0 d(x,x')}$. A user who changes every round by one unit moves the state by at most $\sqrt{N}$, so the user-level metric is $d/\sqrt{N}$ and part (i) applies with $c = \sqrt{N}$. The attack moves the state by $\kappa_0\sqrt{N_A}$. ∎

Part (i) turns the choice of privacy unit into a rescaling inside the model: a unit that protects more is equivalent to a less salient attack. Part (ii) applies the product rule behind the composition of GDP [12, Cor. 2] to an attack of limited duration. With (22) and Theorem 2, an attack is above the critical salience under user-level protection only if it is active in more than a fraction $1/(4\kappa_0^2)$ of the rounds. The same rule prices retention: a record kept for N periods has visibility up to $\sqrt{N}$ times that of one period.

Proposition 4 (Additive noise)**.** *Let $\mathcal{X} = \mathbb{R}$ with $d(x,x') = |x-x'|$, let ζ have distribution function F with a positive, even, log-concave density f, and let $M_b^F(x) = x + b\zeta$ with $b > 0$. Put*

$$L_F = \sup_{u\in\mathbb{R}} \frac{f(u)}{\varphi\big(\Phi^{-1}(F(u))\big)} \in (0,\infty]. \tag{23}$$

(i) The visibility of M_b^F is L_F/b. (ii) The best detector with false-alarm probability α misses an attack of salience κ with probability $F(F^{-1}(1-\alpha) - \kappa/b)$. (iii) When L_F is finite, this miss probability equals the ceiling $\Phi(z_\alpha - \kappa L_F/b)$ for all $\kappa \ge 0$ at some budget $\alpha \in (0, 1/2)$ if and only if ζ is Gaussian. (iv) For the standard Laplace law, with density $e^{-|u|}/2$, $L_F = \sqrt{\pi/2}$.

Proof: A log-concave density has monotone likelihood ratio in its location parameter [25, Example 8.2.1], so the most powerful level-α test between F and its shift by $\theta \ge 0$ rejects above $F^{-1}(1-\alpha)$ [25, Th. 3.4.1] and has miss probability $F(F^{-1}(1-\alpha) - \theta)$. With $\theta = \kappa/b$ this is part (ii).

(i) Let $h = \Phi^{-1} \circ F$, an increasing bijection of $\mathbb{R}$ with $h' = f/(\varphi \circ h)$. For a constant μ', the inequality $F(F^{-1}(1-\alpha)-\theta) \ge \Phi(\Phi^{-1}(1-\alpha) - \mu'\theta)$ for all α is, after the substitution $u = F^{-1}(1-\alpha)$, the inequality $h(u) - h(u-\theta) \le \mu'\theta$ for all u. It holds for all u and all $\theta \ge 0$ exactly when $\mu' \ge \sup h' = L_F$. States at distance θ in the metric d correspond to the shift θ/b, so the visibility is L_F/b.

(iii) Let $u_0 = F^{-1}(1-\alpha) > 0$. Equality for all κ means $h(u_0) - h(u_0 - \theta) = L_F\theta$ for all $\theta \ge 0$, so h is affine with slope L_F on $(-\infty, u_0]$. Since f is even, h is odd, hence affine with the same slope on $[-u_0, \infty)$. The two half-lines overlap, so $h(u) = L_F u$ and $F(u) = \Phi(L_F u)$. The converse is Proposition 1(iii).

(iv) For $u \ge 0$ the Laplace law has $f(u) = 1 - F(u) = e^{-u}/2$, so $h'(u) = [1 - \Phi(v)]/\varphi(v)$ with $v = h(u) \ge 0$. This Mills ratio is decreasing in v and equals $\sqrt{\pi/2}$ at $v = 0$. The case $u < 0$ follows by symmetry. ∎

The visibility of an additive-noise mechanism is thus the Lipschitz constant of the map that transports its noise law to the standard normal. Fig. 5 compares three noise laws at $\kappa = 3$; for the logistic law, numerical evaluation of (23) gives $L_F = 0.627$. By part (iii) no other even log-concave noise law attains the ceiling at every salience, and in the figure the two alternatives lie above the frontier at every noise scale. The gap is large for the Laplace mechanism of pure differential privacy, whose tails are heavier than Gaussian tails and raise the alarm threshold. At a miss probability of 0.02 it carries 0.16 more privacy risk than the frontier, and against a strategic attacker with $\gamma = 1$ at $\alpha = 0.01$ it leaves an undetected impact of

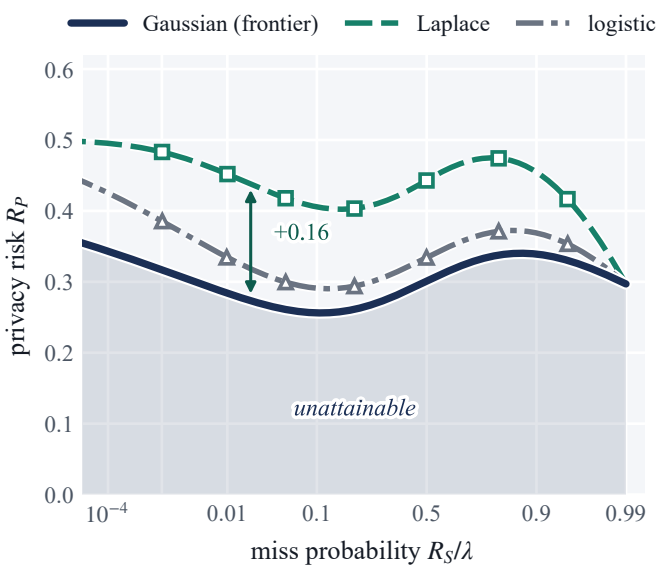


Fig. 5. Additive-noise mechanisms against the frontier ($\kappa = 3$).

$3.03/\mu$, the upper edge of the band in Fig. 6a, against $1.26/\mu$ for Gaussian noise. Both comparisons are at equal visibility.

Proposition 5 (Baseline uncertainty)**.** *Let the benign value of the statistic in* (2) *be* $q(x) = \bar{q} + \delta B$*, where* $\bar{q}$ *is known to the defender,* B *is standard normal and unknown to the defender, and* $\delta \geq 0$ *is the spread of the baseline in privacy units. Put*

$$\mu_{\text{eff}} = \frac{\mu}{\sqrt{1+\delta^2\mu^2}}, \qquad \textit{so that} \qquad \mu_{\text{eff}}^{-2} = \mu^{-2} + \delta^2. \tag{24}$$

(i) For M_μ *and an attack that adds* κ *to the statistic, the most powerful detector at false-alarm budget* α *misses with probability* $\Phi(z_\alpha - \kappa\mu_{\text{eff}})$*. (ii) For* M_μ*, therefore,* $R_S \geq \lambda\Phi(z_\alpha - \kappa/\delta)$ *at every visibility. Against the strategic attacker of Theorem 3 its worst-case undetected impact is* $K_\gamma(\alpha)\,\mu_{\text{eff}}^{-\gamma} \geq K_\gamma(\alpha)\,\delta^\gamma$ *when* $\mu_{\text{eff}} \geq t_{\alpha,\gamma}/\bar{\kappa}$*, with the same detection probability at the saddle point. (iii) The leakage bound* (3) *is unchanged.*

Proof: (i) Without the attack the record is normal with mean $\mu\bar{q}$ and variance $1+\delta^2\mu^2$, and the attack shifts its mean by $\kappa\mu$. The likelihood ratio is increasing in Y, the level-α threshold is $\mu\bar{q} + z_\alpha\sqrt{1+\delta^2\mu^2}$, and the miss probability is $\Phi(z_\alpha - \kappa\mu/\sqrt{1+\delta^2\mu^2})$. (ii) Since $\mu_{\text{eff}} < 1/\delta$, the first claim follows from (i). The attacker's payoff is $\kappa^\gamma\Phi(z_\alpha - \kappa\mu_{\text{eff}}) = \mu_{\text{eff}}^{-\gamma}\, g(\kappa\mu_{\text{eff}})$ with g as in the proof of Theorem 3, whose maximum $K_\gamma(\alpha)$ is attained at $\kappa\mu_{\text{eff}} = t_{\alpha,\gamma}$. (iii) Definition 1 compares two given states, so the adversary of Lemma 1 knows the state apart from one unit. ∎

The noise of the mechanism and the variability of the baseline add in variance, so the defender detects as if the visibility were $\mu_{\text{eff}} < 1/\delta$, while the privacy adversary faces μ as before. Even as $\mu \to \infty$ an attack is missed with probability $\Phi(z_\alpha - \kappa/\delta)$, so what counts is salience relative to the spread of the baseline. The frontier of the reference mechanism becomes $R_P = \sigma[(1-\tau)a(\mu) + \rho\lambda\Phi(z_\alpha - \kappa\mu_{\text{eff}})]$, whose minimum we compute numerically in Section VII.

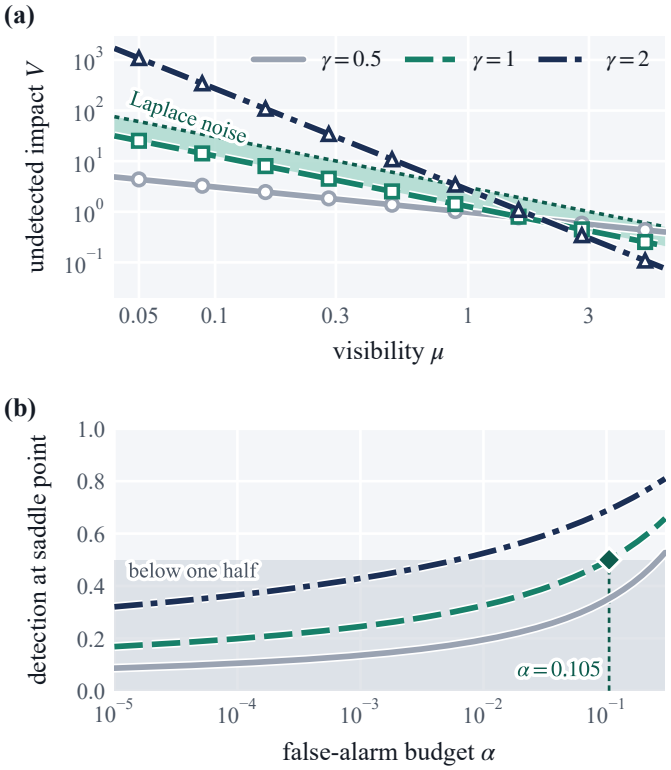


Fig. 6. Strategic attacker: (a) undetected impact, (b) detection probability at the saddle point.

VII. Numerical results

The formulas were checked by simulating the mechanisms and the detectors. Each marker in Figs. 2, 5, 6 and 7 is an average over 2×10^6 simulated periods (2×10^5 in Fig. 7).

For the frontier we drew the record of M_μ under the benign and the attacked state and applied the threshold detector. A privacy adversary ran the likelihood-ratio test between two states one unit apart on records drawn in the same way. The simulated risks differ from (4) and (5) by at most 4×10^{-4} (Fig. 2). The Laplace and logistic mechanisms were simulated with their visibility taken from Proposition 4 (Fig. 5), and the Gaussian mechanism with a random baseline of spread $\delta \in \{0.25, 0.5, 1\}$ reproduces the miss probability of Proposition 5 to within 6×10^{-4}.

For the strategic model we computed the attacker's best response on a grid of saliences and simulated the detector at that response. The product $V\mu^\gamma$ stays within 0.2% of $K_\gamma(\alpha)$ over two decades of visibility and three elasticities (Fig. 6a).

For Proposition 3 we simulated $N = 100$ rounds of Gaussian monitoring with user-level visibility μ and an attack of per-round salience $\kappa_0 = 3$ that is active in N_A rounds and is detected by a matched filter over those rounds. The detection probability follows $1 - \Phi(z_\alpha - \kappa_0\mu\sqrt{N_A/N})$ to within 0.003 (Fig. 7), and the advantage of a user-level adversary matches $a(\mu)$ to within 0.002. The strip in the figure marks dwell fractions below $1/(4\kappa_0^2)$.

Table V applies the theory to five scenarios with $\alpha = 0.01$ and $\lambda = 0.3$. In each, a missed attack ends in the exposure of the data of every unit, so $\rho = 1$, and $\sigma = 1$; the

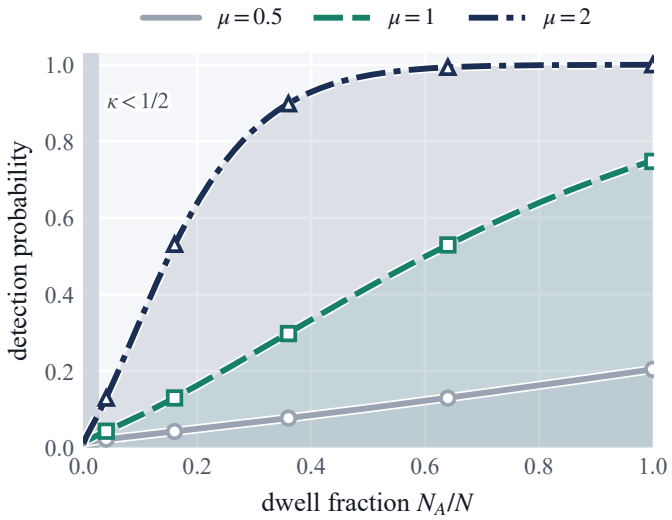


Fig. 7. Detection against dwell fraction ($\kappa_0 = 3$, user-level μ).

TABLE V
SCENARIOS

						at $\mu^\circ(0)$			
No.	κ	τ	Λ	Λ^*	$\mu^\circ(0)$	Det.	R_P	$\mu^\circ(1)$	$\mu^\circ(100)$
1	20	0.5	12	1.59	0.23	0.987	0.049	0.24	0.30
2	3	0.5	1.8	1.52	1.19	0.892	0.256	1.35	1.90
3	1	0.9	3	1.00	5.41	0.999	0.100	5.78	7.30
4	1	0.5	0.6	1.00	0	0.010	0.297	4.80	6.76
5	0.3	0.5	0.18	0.30	0	0.010	0.297	∞	∞

privacy risk without monitoring is then 0.297. Scenario 1 is a credential campaign over twenty accounts and scenario 2 a compromise of three accounts. Scenarios 3 and 4 are a single compromised account under a trusted and a distrusted custodian, and scenario 5 is an intruder whose salience is below the critical value.

For the campaign the privacy optimum is light monitoring ($\mu = 0.23$, an advantage of 0.09 about any unit) that catches 98.7% of attacks and cuts the privacy risk by a factor of six. A weight of 100 on security moves the optimum only to 0.30, so here the dispute over weights concerns a narrow interval. For a single compromised account the answer is all or nothing, and custodial trust decides it through the threshold $\Lambda^*(1) = 1$. Under a trusted custodian, near-total visibility of each unit is privacy-optimal. Under a distrusted one the privacy optimum is no monitoring, any weight $w \geq 1$ on security moves the optimum to $\mu \geq 4.8$, and the proportionality interval $[0, 4.8]$ is wide. Below the critical salience the optimum is either no monitoring or unbounded visibility, for every weight.

An unpredictable baseline changes these answers for the less salient attacks (Proposition 5). With spread $\delta = 0.5$ the privacy optima of scenarios 1 and 2 are 0.23 and 1.21, with detection probabilities 0.987 and 0.782, and scenario 3 moves to unbounded visibility, where the attack is caught with probability 0.37. With $\delta = 1$ the optimum of scenario 1 is still 0.23, and in scenarios 2 and 3 the privacy optimum becomes no monitoring.

VIII. Implications

Detection of prohibited content on the client is a contest with an adaptive sender under a tight false-alarm budget. A perturbation that defeats a perceptual hash leaves the content intact, so the sender gains the same at every salience. This is the limit $\gamma \to 0$ of Theorem 3, where detection at the saddle point is as rare as a false alarm, consistent with the evasion rates measured in [13].

A verified bound on the norm of each client update [14] caps the salience of one client at one privacy unit. Poisoning by k colluding clients then has salience at most k. By Proposition 1 the visibility needed to detect it falls as $1/k$, and the privacy cost falls with it. A single malicious client is detected only at a visibility that exposes every client, and for that case bounding influence is the remaining defense.

Zero-knowledge middleboxes [3] let the record depend on the traffic through one predicate only. In the model this is the choice of the statistic q: behavior that leaves q unchanged is hidden at any visibility, and a statistic matched to the policy violation gives the violation its full salience.

Under user-level protection over a long window, (22) shrinks the salience of a short intrusion: an attack that is active in 3% of the rounds keeps 17% of its per-round salience. User-level protection is the default that NIST guidance recommends [21]. An event-level unit removes the factor, and the guarantee then covers one event only.

To apply the theory to a deployed system, an assessor fixes the privacy unit and the monitored statistic, estimates κ for each attack class as the displacement of a matched statistic in privacy units and δ as the spread of its benign value, reads μ from the mechanism or audits it [19], and takes α and λ from operations. The qualitative parameters τ, σ, ρ and w are given numerical values. The assessor then computes Λ_w and compares it with $\Lambda^*(\kappa)$, and where monitoring pays computes the exchange rate (9) at the current visibility and the proportionality interval. A negative exchange rate means that the system monitors too little for privacy itself.

IX. Conclusion

Detection and privacy loss draw on one resource, the visibility of the monitoring record, and the frontier prices each in units of the other. The pressure ratio grows in proportion to salience while its threshold stays bounded, so sufficiently salient attacks make the privacy-optimal visibility positive, and the weight given to security then moves it only through a logarithm. For an attack of salience one the answer turns on custodial trust and on how predictable the benign baseline is. Gaussian noise is the only additive design with an even log-concave density that reaches the detection ceiling at every salience.

References

[1] H. Abelson *et al.*, "Bugs in our pockets: The risks of client-side scanning," *Journal of Cybersecurity*, vol. 10, no. 1, p. tyad020, 2024.
[2] S. Scheffler and J. Mayer, "SoK: Content moderation for end-to-end encryption," *Proceedings on Privacy Enhancing Technologies*, vol. 2023, no. 2, pp. 403–429, 2023.

[3] P. Grubbs, A. Arun, Y. Zhang, J. Bonneau, and M. Walfish, "Zero-knowledge middleboxes," in *Proc. 31st USENIX Security Symposium*, Boston, MA, USA, Aug. 2022, pp. 4255–4272.

[4] P. Kairouz *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[5] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," Official Journal of the European Union, L 119, pp. 1–88, May 2016.

[6] F. Farokhi and P. Mohajerin Esfahani, "Security versus privacy," in *Proc. IEEE Conference on Decision and Control (CDC)*, Miami Beach, FL, USA, Dec. 2018, pp. 7101–7106.

[7] J. Giraldo, A. A. Cardenas, M. Kantarcioglu, and J. Katz, "Adversarial classification under differential privacy," in *Proc. Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, Feb. 2020.

[8] A. Ünsal and M. Önen, "Calibrating the attack to sensitivity in differentially private mechanisms," *Journal of Cybersecurity and Privacy*, vol. 2, no. 4, pp. 830–852, 2022.

[9] National Institute of Standards and Technology, "NIST privacy framework: A tool for improving privacy through enterprise risk management, version 1.0," National Institute of Standards and Technology, Gaithersburg, MD, USA, NIST Cybersecurity White Paper CSWP 6, Jan. 2020.

[10] D. E. Pozen, "Privacy-privacy tradeoffs," *University of Chicago Law Review*, vol. 83, pp. 221–247, 2016.

[11] V. Katewa, R. Anguluri, and F. Pasqualetti, "On a security vs privacy trade-off in interconnected dynamical systems," *Automatica*, vol. 125, p. 109426, 2021.

[12] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 84, no. 1, pp. 3–37, 2022.

[13] S. Jain, A.-M. Creţu, and Y.-A. de Montjoye, "Adversarial detection avoidance attacks: Evaluating the robustness of perceptual hashing-based client-side scanning," in *Proc. 31st USENIX Security Symposium*, Boston, MA, USA, Aug. 2022, pp. 2317–2334.

[14] A. Roy Chowdhury, C. Guo, S. Jha, and L. van der Maaten, "EIFFeL: Ensuring integrity for federated learning," in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Los Angeles, CA, USA, Nov. 2022, pp. 2535–2549.

[15] V. Duddu, S. Szyller, and N. Asokan, "SoK: Unintended interactions among machine learning defenses and risks," in *Proc. IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, May 2024, pp. 2996–3014.

[16] S. B. Hopkins, G. Kamath, M. Majid, and S. Narayanan, "Robustness implies privacy in statistical estimation," in *Proc. 55th Annual ACM Symposium on Theory of Computing (STOC)*, Orlando, FL, USA, Jun. 2023, pp. 497–506.

[17] J. Stegman, P. J. Trottier, C. Hillier, H. Khan, and M. Mannan, "'My privacy for their security': Employees' privacy perspectives and expectations when using enterprise security software," in *Proc. 32nd USENIX Security Symposium*, Anaheim, CA, USA, Aug. 2023, pp. 3583–3600.

[18] B. A. Alahmadi, L. Axon, and I. Martinovic, "99% false positives: A qualitative study of SOC analysts' perspectives on security alarms," in *Proc. 31st USENIX Security Symposium*, Boston, MA, USA, Aug. 2022, pp. 2783–2800.

[19] M. Nasr *et al.*, "Tight auditing of differentially private machine learning," in *Proc. 32nd USENIX Security Symposium*, Anaheim, CA, USA, Aug. 2023, pp. 1631–1648.

[20] B. Kulynych, J. F. Gomez, G. Kaissis, F. du Pin Calmon, and C. Troncoso, "Attack-aware noise calibration for differential privacy," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[21] J. P. Near, D. Darais, N. Lefkovitz, and G. S. Howarth, "Guidelines for evaluating differential privacy guarantees," National Institute of Standards and Technology, Gaithersburg, MD, USA, NIST Special Publication 800-226, Mar. 2025.

[22] H. Nissenbaum, "Privacy as contextual integrity," *Washington Law Review*, vol. 79, no. 1, pp. 119–158, 2004.

[23] D. K. Citron and D. J. Solove, "Privacy harms," *Boston University Law Review*, vol. 102, pp. 793–863, 2022.

[24] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, "Broadening the scope of differential privacy using metrics," in *Privacy Enhancing Technologies (PETS 2013)*, ser. Lecture Notes in Computer Science, vol. 7981. Berlin, Germany: Springer, 2013, pp. 82–102.

[25] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 4th ed. Cham, Switzerland: Springer, 2022.