A Repeated-Game Framework for Incentives in Decentralized Infrastructure Protocols

Mustafa Qazi

Volt Capital

1 Estimating the Deterrence Ratio in Practice

Applying Theorem 1 requires an estimate of Γ, and the natural procedure—enumerate deviations, execute them with red-team nodes, measure detection rates—is biased in the unsafe direction: it lower bounds Γ with slack determined by the deviations the designer failed to imagine, while the incentive inequality requires enforcement to *cover* Γ. This section gives a calibration procedure that does not enumerate the action space. It rests on one reparametrization and two diagnostics, after which the residual estimation problem takes a different—and measurable—form in each of three canonical audit regimes. Our incentive inequality is a dynamic analogue of the classical deterrence condition of Becker [**?**], and the reparametrization places the monitoring layer in the costly-state-falsification tradition of Lacker and Weinberg [**?**] and its test-design branch [**?**]; what those literatures leave exogenous—the detection response to an adversarially chosen distortion—is precisely what the procedure below measures.

1.1 From Deviations to Costs

A deviation enters the model only through the cost it saves and the signal distribution it induces. For $x \in \mathcal{X}$ let $\mathcal{F}(x) = \bigcup_{a \in A} L(x,a)$ and, for $f \in \mathcal{F}(x)$, define the *simulation cost*

$$C(f,x) \;=\; \min\{\, c(a,x) \;:\; a \in A,\; f \in L(x,a) \,\},$$

the cheapest way of making the audit see f; the minimum exists because A is finite. Let $\mathcal{F}^{+}(x) = \{f : C(f,x) < c(a^{*},x)\}$.

Proposition 1 (Distributional Form of Γ). *Separation holds at x iff $P_f(1) > \eta(x)$ for all $f \in \mathcal{F}^{+}(x)$, and then*

$$\max_{d \in D(x)} \gamma_d(x) \;=\; \max_{f \in \mathcal{F}^{+}(x)} \frac{c(a^{*},x) - C(f,x)}{P_f(1) - \eta(x)}.$$

Proof. For $d \in D(x)$, the worst-case manipulation f_d of Definition 3 satisfies $C(f_d,x) \le c(d,x)$, so the right side weakly dominates $\gamma_d(x)$. Conversely, for $f \in \mathcal{F}^{+}(x)$ take a attaining $C(f,x)$; then $a \in D(x)$ and $\varepsilon_a(x) \le P_f(1) - \eta(x)$ since $f \in L(x,a)$, so $\gamma_a(x)$ weakly dominates the ratio at f. Both sides are maxima over the finite set $D(x)$.

The consequence for calibration: the designer never constructs A. Any lower bound on $C(\cdot, x)$ over a region of the simplex upper bounds that region's contribution to Γ, and the estimation question becomes *what it costs an arbitrary adversary to make the audit see a given distribution*—a quantity constrained by cryptography, physics, and hardware economics even when the strategy space is not.

Two diagnostics must pass before any ratio is worth computing.

Diagnostic 1: monitored sufficiency. Deviations inducing exactly the compliant outcome law (the *kernel* of the audit) are undeterrable at any stake or reward level by Proposition 1. Testing for them does not require imagining attacks; it requires only cost accounting. Decompose the compliance cost by service dimension, $c(a^*, x) = \sum_{\ell \in \mathcal{L}} c_\ell(x)$ (hardware, bandwidth, power, uptime, freshness, ...), and let $\mathcal{L}_G \subseteq \mathcal{L}$ be the dimensions on which the outcome statistic G depends.

Proposition 2 (Monitored-Sufficiency Criterion). *If $c_\ell(x) > 0$ for some $\ell \notin \mathcal{L}_G$ and expenditure on ℓ can be reduced holding the coordinates in $\mathcal{L}_G$ fixed, then a profitable kernel deviation exists and no binary public-outcome protocol implements compliance against it.*

Proof. Reducing expenditure on ℓ leaves the inputs of G, hence the outcome law, unchanged, while saving $c_\ell(x) > 0$: the induced f lies in the kernel with $C(f, x) < c(a^*, x)$. Apply Proposition 1.

The quantifier has moved from "all strategies the adversary might invent" to "all line items in the honest cost stack," which is a finite, known list—the provider bill of materials. Each unmonitored cost line is resolved in one of three ways before proceeding: enrich Z to monitor it; accept it and remove it from the compliance definition (a welfare-neutral kernel direction, e.g. replaying one's own correct cached outputs, is an efficiency gain to permit, and $\bar{c}$ falls accordingly); or acknowledge it as unenforced and disclose. Certifying the converse—that monitored coordinates pin down the *entire* cost stack—is available only where a reduction supplies it (classes (A) and, under determinism, (C) below); absent such an argument the kernel should be presumed nonempty, which errs safe.

Diagnostic 2: local slopes. Along each continuously scalable deviation axis $\{f_\theta\}_{\theta \geq 0}$, $f_0 = f_{a^*}(\cdot \mid x)$ (transmit power, storage undercommitment, quantization depth), let $s(\theta) = c(a^*, x) - C(f_\theta, x)$ and $e(\theta) = P_{f_\theta}(1) - \eta(x)$.

Lemma 1 (Local Divergence). *If $s'(0) > 0$ while $e \equiv 0$ on an initial interval (a tolerance band) or $e(\theta) = O(\theta^2)$, then $\sup_\theta s(\theta)/e(\theta) = \infty$: no finite $(\lambda S, \Phi)$ satisfies (5). If instead $e'(0) > 0$, local deviations contribute the finite ratio $s'(0)/e'(0)$.*

Proof. $s(\theta)/e(\theta) \geq (s'(0)\theta + o(\theta))/K\theta^2 \to \infty$ in the second-order case; the band case is Proposition 2 applied at f_θ; the first-order case is the quotient of expansions.

Corollary 1 (Repricing the Band). *A tolerance band of width θ_0 redefines compliance: the implemented action is the cost-minimizer inside the band, with cost $\tilde{c}(x) = c(a^*, x) - s(\theta_0)$, and all quantities downstream of $\bar{c}$ in Sections 4–5 must be evaluated at $\tilde{c}$. Deterrence analysis applies only to deviations exiting the band.*

Operationally, $(s'(0), e'(0))$ are finite differences: run the deviation at small parametric intensities and regress detection frequency on intensity. The condition $e'(0) = 0$ says the outcome statistic is first-order blind along the axis—the local analogue of a full-rank failure in [4] and of the vanishing-detection collapse in [14]—and the design response is to make detection first-order wherever savings are.

1.2 Three Bounds for Three Audit Regimes

Past the diagnostics, the residual bound on C comes from the source of the audit's detection power: a computational soundness reduction (*attestation-backed*), a physical boundary between adversary and honest resources (*physically bounded*), or output distinguishability (*statistically bounded*). Storage, wireless coverage, and compute are the canonical instances.

Attestation-backed: storage. Proof-of-storage soundness quantifies over all polynomial-time adversaries, which is exactly what removes enumeration. Adopting the notions of Fisch [**?**]—an arbitrarily tight PoS with space gap ϵ_{sp} and soundness error δ_{sound}, with sequentially hard encoding making on-demand regeneration cost ρ_{regen} per window—suppose any adversary passing a window with probability above δ_{sound} either dedicates $(1 - \epsilon_{\text{sp}})$ of the space or pays ρ_{regen}. Then for the entire non-dedication class, with σ the per-period cost of honest dedication,

$$\sup_{f \in \mathcal{D}_{\text{ns}}} \frac{c(a^*, x) - C(f, x)}{\varepsilon(f, x)} \;\leq\; \frac{\max\big(\epsilon_{\text{sp}}\sigma,\ (\sigma - \rho_{\text{regen}})^+\big)}{1 - \eta(x) - \delta_{\text{sound}}}, \tag{1}$$

by cases on whether the identified cost is paid. All inputs are protocol constants or benchmarks: ρ_{regen} is a sequential re-encoding benchmark at market hardware prices, designed to satisfy $\rho_{\text{regen}} \gg \sigma$. Note the tightness gap is a tolerance band—savings up to $\epsilon_{\text{sp}}\sigma$ are free and belong in $\tilde{c}$ per Corollary 1, which is the economic meaning of tuning ϵ_{sp} small. Two disclosed residuals: Fisch–Silas [**?**] show full rational security is likely unprovable without strong knowledge assumptions (a known kernel component), and the attestation is silent on retrieval and availability, whose cost lines fail Proposition 2 and must be resolved by enriching Z—as deployed storage networks have done with retrieval-specific programs.

Physically bounded: wireless coverage. No reduction exists here, by theorem: classical position verification is impossible against colluding adversaries [**?**], and hidden or mobile validators raise the attack's coordination cost without restoring possibility. The consequence is graph-theoretic. Let $W = (V, E)$ be the

witness graph, $B \subseteq V$ the adversary-controlled set, ∂B the cut (*attack edges*, after [?]), and E_t the randomly activated challenge edges. If, on challenges with $E_t \cap \partial B = \emptyset$, the adversary can induce the compliant signal law at hardware cost $\rho_{\mathrm{rf}}(|B|)$ (subadditive; one RF apparatus simulates many co-located identities—the content of documented attenuator-and-coaxial attacks), then conditioning on whether a challenge crosses the cut gives, for any deviation supported on B,

$$\varepsilon(f,x) \;\le\; q(B) := \Pr[E_t \cap \partial B \neq \emptyset] \;\le\; \textstyle\sum_{e\in\partial B} \Pr[e \in E_t], \qquad \Gamma \;\ge\; \max_B \frac{\Delta_B(x)}{q(B)}, \tag{2}$$

with $\Delta_B(x) = \sum_{i\in B} c(a^*, x_i) - \rho_{\mathrm{rf}}(|B|)$, and separation failing for any profitable B with $q(B) = 0$. This is the small-cut principle of graph-based Sybil defense [?,?] expressed as a bound on monitoring power: detection is purchased only on attack edges. The measurement recipe is concrete. Fix a coalition family $\mathcal{B}$ (all B up to a size cap, or geographically contiguous B from the deployment's location index); estimate $q(B)$ by Monte Carlo over the protocol's own challenge-routing distribution; price $\rho_{\mathrm{rf}}(k)$ from attack-kit bills of materials; take Δ_B from deployment cost models. The binding-coalition search is sparsest-cut-like and need only be bounded over $\mathcal{B}$, not solved exactly. Effective audit frequency against B is $q(B)$, not the nominal rate of Lemma 3, so the designer's lever is routing: choose the challenge distribution to $\max \min_{B\in\mathcal{B}} q(B)$, and evaluate program (10) at that value. Isolated profitable components ($q(B) = 0$) are undeterrable by any (λ, S, r, n) and shift enforcement to the identity margin (Section 4.4, footnote 5).

Statistically bounded: compute. Here the adversary's optimum is characterizable: substitute the cheapest computation whose output law is hardest to distinguish on the audit distribution Q—distillation under a detectability constraint. With per-query audit probability p, committed computation M^*, substitutes $\mathcal{M}$, and P_M^Q the output law of M on Q, the Neyman–Pearson bound gives $\varepsilon_M \le p \cdot d_{\mathrm{TV}}(P_{M^*}^Q, P_M^Q)$ per audited query, so with frontier $v(s) = \inf\{d_{\mathrm{TV}}(P_{M^*}^Q, P_M^Q) : c(M^*) - c(M) \ge s\}$,

$$\Gamma_{\mathrm{sub}} \;=\; \sup_{s>0} \frac{s}{p\, v(s)}, \tag{3}$$

with separation failing where $v(s) = 0$. Since on-chain audit logic is public, the bound is evaluated with the adversary knowing (Q, p) and the detector, so it survives full transparency. Two facts discipline its use. First, mixing is not an attack on Γ: routing to a substitute with probability α scales savings and total variation by exactly α (total variation is a norm), so the ratio is constant along mixture rays; the empirical collapse of test power at low substitution rates [?] reflects finite-sample suboptimality of deployed detectors, not adversarial improvement—a distinction that separates "Γ is infinite" (unfixable by statistics) from "the detector is underpowered" (fixable, with known headroom). Second, the frontier is estimated, and its two biases run in opposite directions and should be reported: sweeping only known substitution axes (quantization, pruning, student scale, speculative decoding) *over*states v, hence understates Γ_{sub}, with slack bounded by the gap between the designer's and the adversary's

distillation technology; while estimating each d_{TV} from samples via classifier two-sample tests or logprob comparisons yields a *lower* bound on the divergence (any test's power lower bounds the optimum), hence overstates Γ_{sub}, which errs safe. Empirically $v(s) \approx 0$ regions are documented at realistic audit budgets for quantized substitution of smaller models [**?**], so this vertical is where the framework most often returns "redesign the audit, not the stake." The redesign menu, ordered by what it does to the frontier: bit-deterministic committed inference with exact recomputation makes $\varepsilon = p$ for every output-altering deviation (compliant outputs become point masses), collapsing the statistical problem to replication economics at the reproducibility overhead of controlled floating-point ordering [**?**]; activation-commitment schemes [**?**] obtain most of this at low prover cost by moving the tolerance from the output level to the activation level, where the free-savings zone of Lemma 1 is far smaller; succinct proofs of inference [**?**] restore an attestation-backed reduction at prover overheads currently orders of magnitude above native cost. Each rung is priced in program (10) through $(\bar{c}, c_{\mathrm{audit}}, \varepsilon)$. Output-level tolerance without determinism defines a free quantization zone (Lemma 1); the response is Corollary 1: commit to and price the cheapest computation inside the tolerance. Caching of one's own correct outputs is welfare-neutral kernel (permit it; $\bar{c}$ falls); stale caching where freshness is a costed dimension fails Proposition 2 on the current Q and is repaired audit-side, by injecting entropy (nonces, paraphrase) so the cacheable set has Q-measure zero—Q is a design variable of the same standing as challenge routing above.

1.3 Procedure and Data Requirements

1. **Monitored sufficiency (Proposition 2).** Decompose the honest cost stack by service dimension; flag every positive cost line the outcome statistic does not depend on; resolve each (enrich Z / reprice into $\tilde{c}$ / disclose) before any ratio is computed.
2. **Local slopes (Lemma 1).** For each scalable deviation axis, estimate $(s'(0), e'(0))$ by finite-difference experiments at small intensities. First-order savings with banded or second-order detection: redefine compliance or redesign the audit; do not attempt to cover it with stake.
3. **Regime bound.** Compute (1), (2), or (3) as appropriate, from the inputs in Table 1.
4. **Design feedback.** Evaluate program (10) at the resulting bound, with the regime's audit variables—challenge routing; audit distribution Q; verification technology—added to the designer's choice set alongside (n, p, S, r).

Red-team experiments retain a subordinate role: they validate the regime assumptions—that ρ_{rf} is as priced, that the swept frontier is near the adversary's—rather than serving as the primary estimator. The structural bounds carry the quantifier over unenumerated deviations; experiments calibrate their constants.

Table 1. Calibration inputs by audit regime.

Regime	Parameters	Measurement source
Attestation (storage)	ϵ_{sp}, δ_{sound}, ρ_{regen}, σ, η	Protocol constants; sequential re-encoding benchmarks at market hardware prices; storage cost models; on-chain fault rates of reputable providers
Physical (wireless)	$q(B)$, $\rho_{\text{rf}}(k)$, Δ_B	Monte Carlo over own routing distribution and witness graph; attack-kit bills of materials and incident postmortems (denylists); deployment cost models
Statistical (compute)	$v(s)$, p, $c(M)$, η	Distillation sweep on deployed Q with classifier/logprob divergence estimates (report both bias directions); GPU spot pricing per substitute; recomputation error rates