



Manipulation-Proof Oblivious Audits against Deceptive Model Providers

paper
#897



Augustin Godinot¹ Sofiane Azogagh² Julien Ferry³ Sébastien Gambs⁴
¹Max Planck Institute for Software Systems
³CIRRELT & SCALE-AI Chair, Polytechnique Montréal
²EURECOM
⁴Université du Québec à Montréal



Audits are now law...

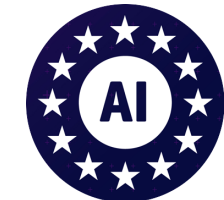


“[...] shall be subject at their own expense and at least once a year, to independent audits”

DSA Art. 37(Audits)

“[...] require access to any data and algorithms of undertakings [...]”

DMA Art. 21(RFI)



“[...] request access to the general-purpose AI model concerned through APIs [...]”

AI Act Art. 92(Evals)

... but providers see the auditor coming

Uber	showed a fake app to competitors and undercover regulators.	NYT 2017
VW	detected emission tests; up to 40× the NO _x limit on the road vs emission tests.	EPA 2015
Meta	deleted exactly the scam ads regulators searched for, without improving moderation.	Reuters 2025
	used different Llama 4 models for different benchmarks: results were “fudged a little bit” (Y. LeCun).	FT 2026

Patching is cheap

Context: demographic parity d_S audit. Audit set S , tolerance ε .
 $d_S = p_{1,S} - p_{0,S}$, audit passes iff $|d_S| \leq \varepsilon$.

Proposition 1. To pass, the provider must flip at least

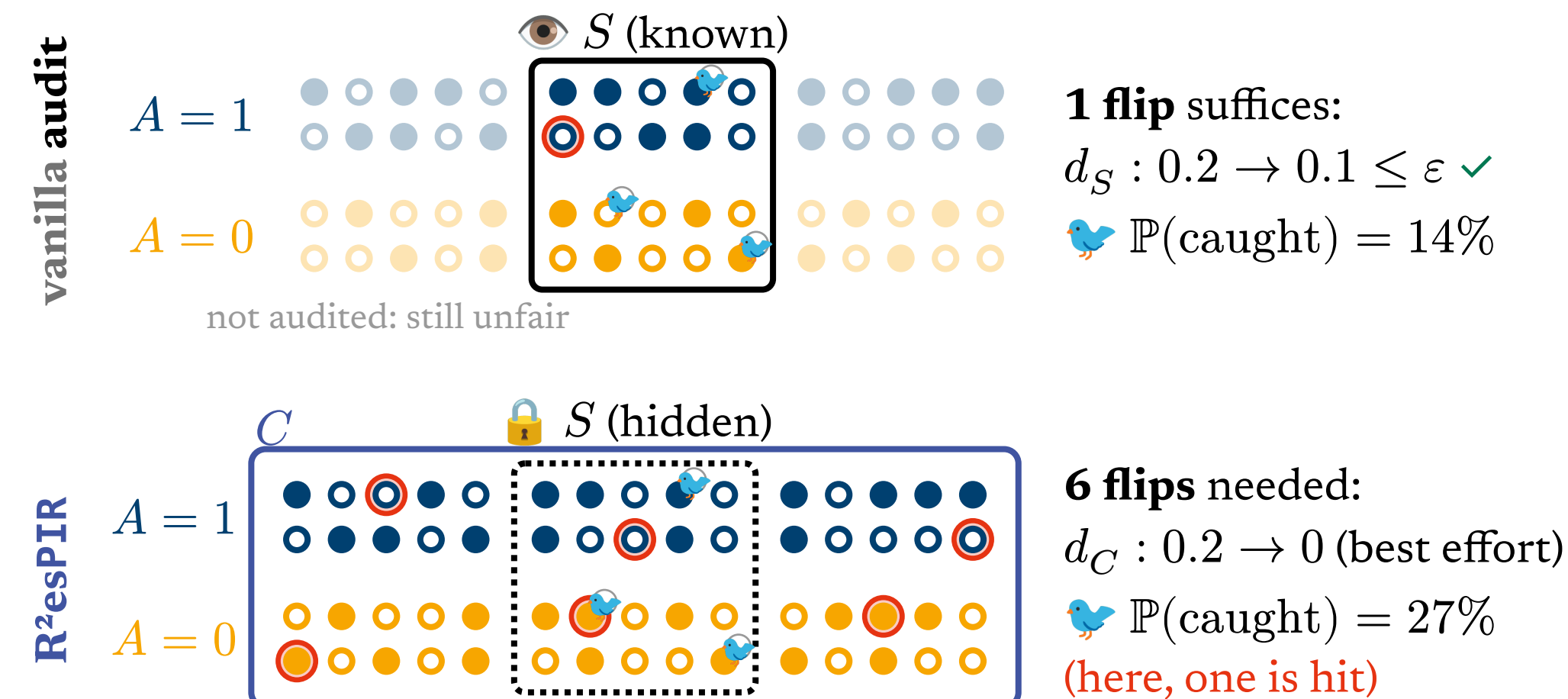
$$m_{\text{vanilla}} \geq \lceil (|d_{\text{true}}| - \varepsilon) \cdot n_{\min} \rceil$$

predictions, with $n_{\min}$ the size of the smallest group in S .

Other robust audits tools

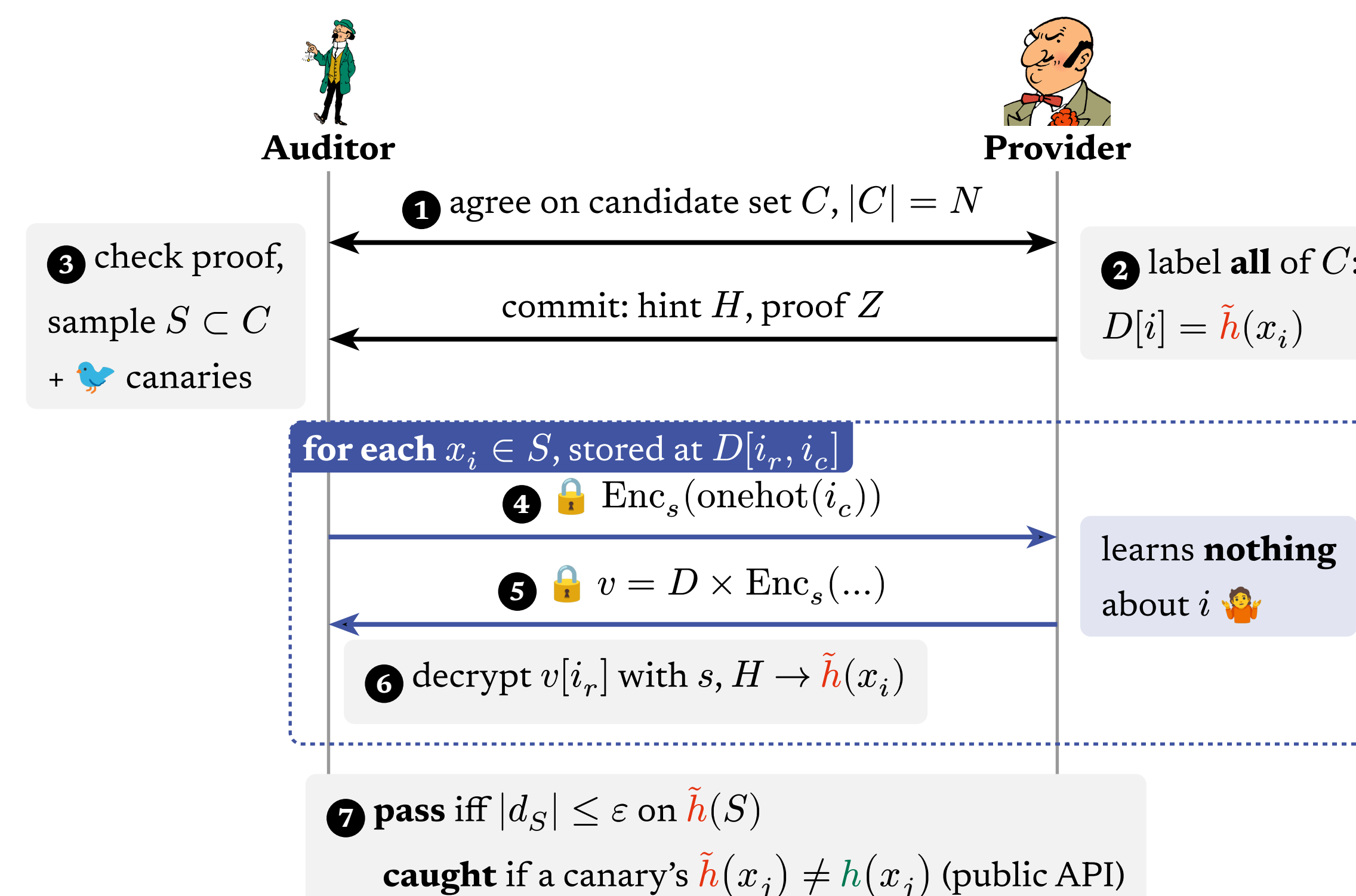
	no trust in provider	cheap for auditor	cheap for provider	any model
Other views (research APIs, canaries)	×	✓	✓	✓
External info (labels, priors)	~	×	✓	✓
Verified inference (ZKP, MPC, TEE)	✓	~	×	×
R²esPIR (canaries + PIR)	~	✓	✓	✓

● positive outcome ○ negative outcome ⊖ flipped by provider 🐦 canary



Providers patch *only* what they know you audit.

This paper: hide the audit set S in a large candidate set C with Private Information Retrieval. To pass, the provider must fix **all** of C .



R²esPIR increases detection

Theorem 1. For S uniform in C , $|d_C| \leq \varepsilon - \gamma$ ensures passing w.p. $\geq 1 - \delta$. Reaching it takes

$$m_{\text{R}^2\text{esPIR}} \geq \lceil (|d_{C,\text{true}}| - \varepsilon + \gamma) \cdot N_{\min} \rceil \quad \gamma = \sqrt{\frac{2 \ln(4/\delta)}{n_{\min}}}$$

With k canaries: $\mathbb{P}(\text{caught}) = 1 - (1 - qm/N)^k$

×4 flips, up to ×2 detection

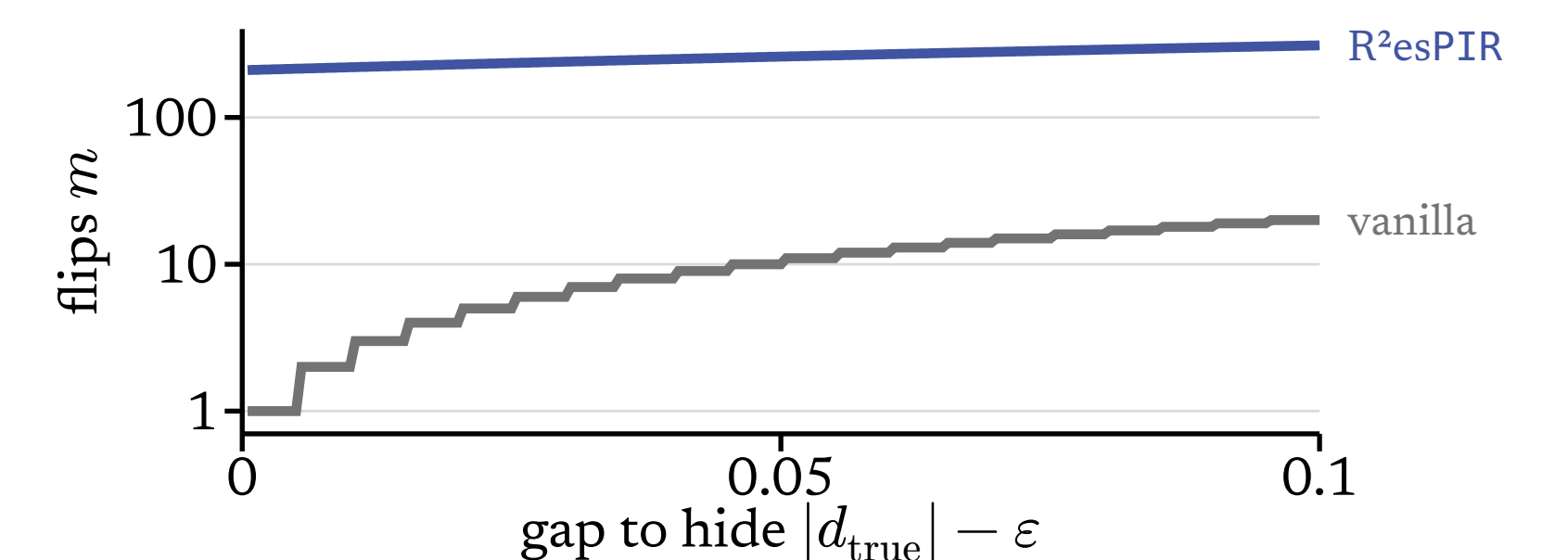


Figure 1: Flips to pass (Prop. 1) or reach $|d_C| \leq \varepsilon - \gamma$ (Thm. 1, $\varepsilon \geq \gamma \approx 0.21$), balanced $|S| = 400, |C| = 2000, \delta = 0.05$.

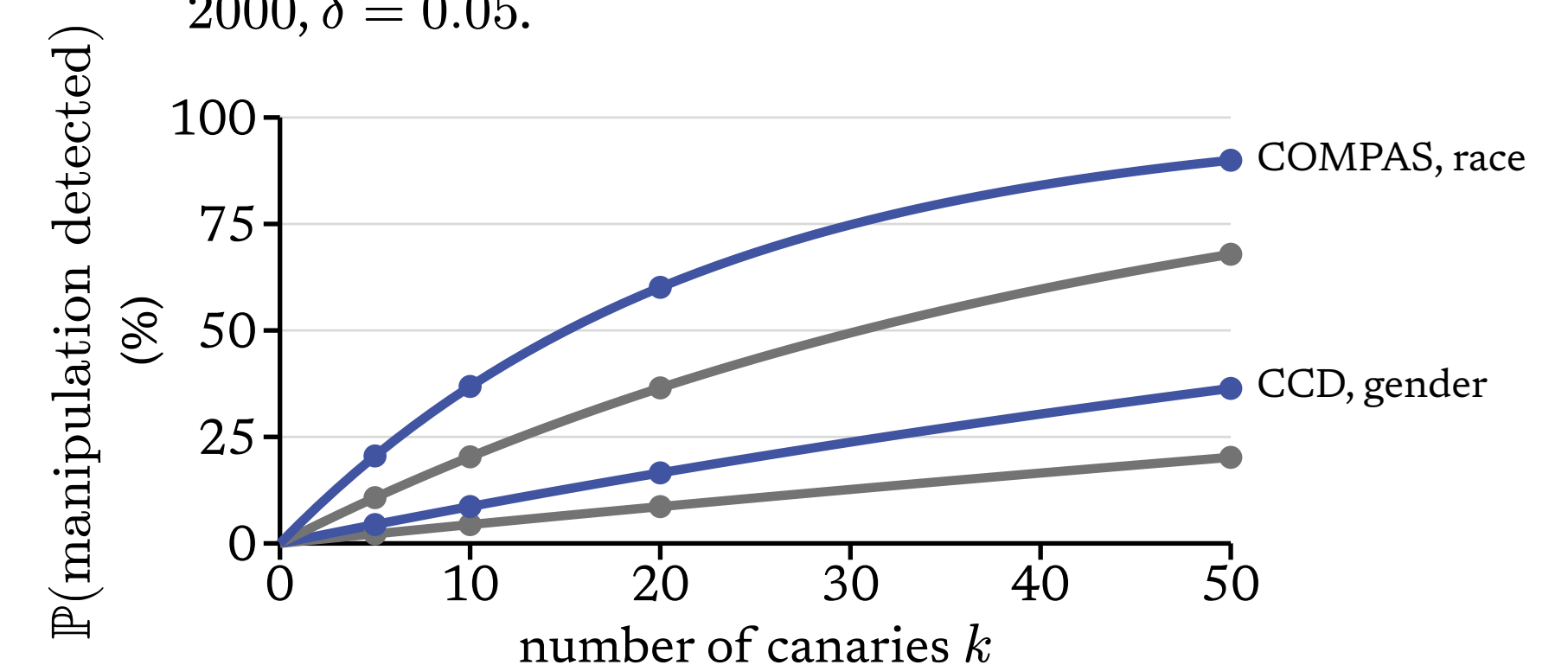


Figure 2: k canaries ($q = 0.8$) vs a provider hiding half its unfairness ($\varepsilon = |d_{C,\text{true}}|/2$), same $|S|$ for both. Here $\varepsilon < \gamma$: R²esPIR's best effort is $d_C = 0$.

PIR is efficient

labels in C	2^{20}	2^{30}	2^{33}	2^{35}	2^{36}
time / query	1 ms	28 ms	117 ms	1.6 s	3.6 s
bytes / query	8 KiB	290 KiB	855 KiB	1.7 MiB	2.4 MiB