CERIAS

The Center for Education and Research in Information Assurance and Security

EUREKA: A General Framework for Black-box Differential Privacy Estimators

Yun Lu, *University of Victoria*, *yunlu@uvic.ca*
Malik Magdon-Ismail, *Rensselaer Polytechnic Institute*, *magdon@cs.rpi.edu*
*Yu Wei**, *Purdue University*, *yuwei@purdue.edu*
Vassilis Zikas, *Purdue University*, *vzikas@purdue.edu*

Overview

Question

Differential privacy (DP) is a key tool in privacy-preserving data analysis. Yet it remains challenging for non-privacy-experts to prove the DP of their algorithms. Do we have a methodology for domain experts with limited data privacy background to empirically estimate the privacy of an *arbitrary* mechanism?

Our Results

- We showed the impossibility of the above task for unrestricted input domains, and introduce a natural, application-inspired relaxation of DP which we term **relative DP**.
- We proved ***a new link*** between the problems of DP parameter-estimation and Bayes optimal classifiers in ML.
- We propose ***a general framework*** for constructing and analyzing black-box DP estimators. The instantiated estimators achieve two desirable properties:
  - *black-box*, i.e., they do not require knowledge of the underlying mechanism
  - They have a theoretically-proven accuracy

Our Result in Details

What is relative DP?

Relative DP protects individual privacy x during the query M over dataset $D \in T$.

Choose a favorite pair of neighboring database
$D_0 = D, D_1 = D \bigcup \{x\}$

Flip a coin b, sample $r \leftarrow M(D_b)$, send r

Adversary knows the mechanism M
And will choose $D_0, D_1 \in T$

Relative DP guarantees that, any adversary, with probability at least $1-\delta$, can win the above indistinguishable game with advantage at most ϵ.

Eureka Framework: construct a estimator from any binary classifier

1. Given a tested mechanism M, a binary classifier h and a database set T, and a privacy parameter ϵ
2. For any $D_0, D_1 \in T$, do the following
   1. Set r.v. $X = M(D_0), Y = M(D_1)$
   2. Set binary classification problem with likelihood $X, [Y]_\epsilon$, uniform prior and 0/1 loss function
   3. Estimate the risk $R(h)$ over this classification problem
   4. Use ***our link*** to convert the estimate in Step 3 to δ'_{D_0,D_1}
3. Compute $\delta' = \max_{D_0,D_1 \in T} \{\delta'_{D_0,D_1}\}$
4. Claim mechanism M satisfies (ϵ, δ', T) relative DP.

Eureka Moment

A link casting the DP parameter-estimation problem to a binary classification problem.

For a binary classification problem with likelihood $X, [Y]_\varepsilon$[1], uniform prior and 0/1 loss function, we have

$$\delta_{X,Y}(\varepsilon) = \max\left(1 - 2\exp(\varepsilon) R(h^*), 0\right),$$

where $R(h^*)$ is the risk of the optimal Bayes classifier for the problem, and $\delta_{X,Y}(\varepsilon)$ is defined as

$$\delta_{X,Y}(\varepsilon) = \max\left(\max_{\mathcal{S} \subseteq \mathcal{O}} \left(\Pr[X \in \mathcal{S}] - e^{\varepsilon} \Pr[Y \in \mathcal{S}]\right), 0\right).$$

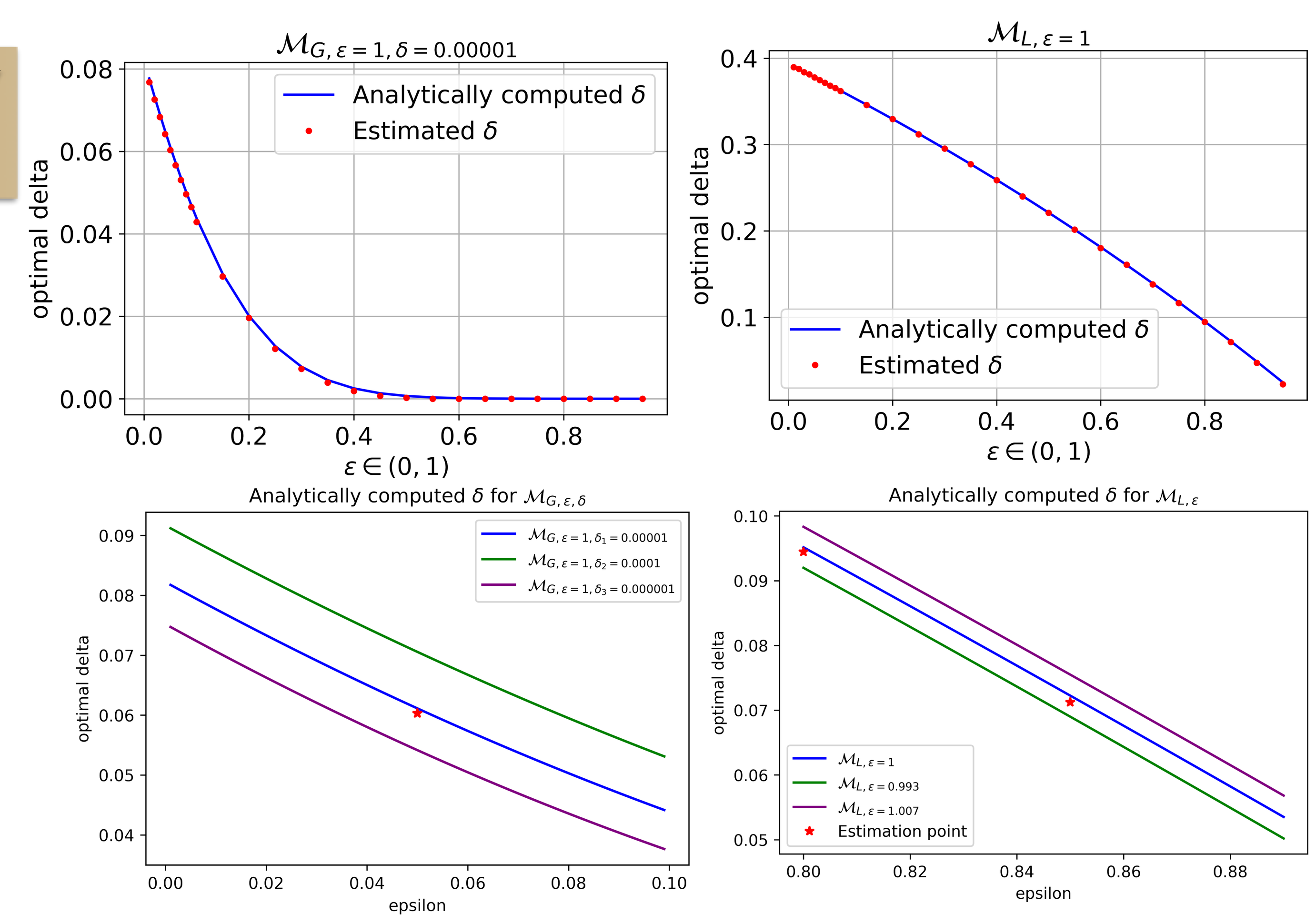
[1] $[Y]_\epsilon$ is a distribution for tossing a coin c where $\Pr[c=1] = \exp(-\varepsilon)$, outputting Y if $c = 1$ or $\perp$ (a null value) otherwise.

Evaluation on a concrete estimator

Theoretical-proven Accuracy

Consider the set of mechanisms $\mathcal{C} = \mathcal{U}^m \mapsto \mathrm{R}^d$ whose output distribution has a density. Let $T \subseteq \mathcal{U}^m$ be any set of databases with size less than t. Let h be a kNN classifier which is constructed from n samples and $k = \sqrt{n}$. Then there exists a n_0 such that for all $n > n_0$, the left-hand side algorithm is a (α, β)-Approximate Relative DP Estimator for $\mathcal{C}$, where $\alpha = 24e^{\epsilon} c_d \sqrt{\ln(8tm/\beta)/n} + 2e^{\epsilon}\sqrt{\ln(8tm/\beta)/n}$.

Estimates tightly match the analytical result



Reference

1. Yun Lu, Malik Magdon-Ismail, Yu Wei and Vassilis Zikas. EUREKA: A General Framework for Black-box Differential Privacy Estimators. S&P 2024.

PURDUE UNIVERSITY

CERIAS