Robust Local Differential Privacy

Milan Lopuhaä-Zwakenberg, TUE
Jasper Goseling, UT

SITB 2021

Partially sensitive data

- users have data $X = (S, U) \in$ finite alphabet $\mathcal{X}$
- S is considered sensitive, U is not
- They may be correlated!

Analyst Goal: Create sanitization protocol $\mathcal{Q}$ s.t. $Y := \mathcal{Q}(X)$ has much info about X but does not tell too much about S.

Utility: Maximise $\mathrm{I}(X;Y)$.

(Local Differential) Privacy: Choose $\varepsilon > 0$. Ensure

$$\mathbb{P}(Y = y | S = s) \leq e^{\varepsilon} \mathbb{P}(Y = y | S = s'). \qquad (1)$$

Robustness

Let $\mathcal{P}_{\mathcal{X}} = \{\text{prob.distr. on } \mathcal{X}\}$.

- Problem: Analyst has no access to true distribution P^*.
- Assumption: There is a publicly available dataset $\vec{x}$ drawn from P^*.
- From $\vec{x}$ analyst creates *estimate* $\hat{P} \in \mathcal{P}_{\mathcal{X}}$ and *confidence set* $\mathcal{F} \subset \mathcal{P}_{\mathcal{X}}$ via χ^2-test.

Robust privacy: Satisfy (1) for all $P \in \mathcal{F}$.

(Robust utility: maximize $\min_{P \in \mathcal{F}} \mathrm{I}_P(X;Y)$: Work in progress!)

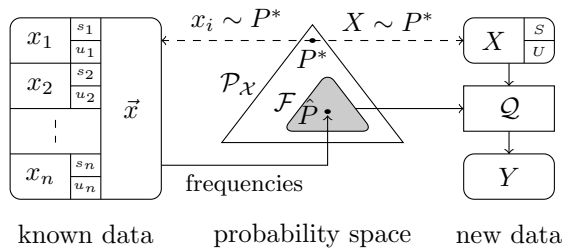

Figure 1: Overview.

Polyhedral optimisation

- Take polyhedron $\mathcal{D} \supset \mathcal{F}$.
- Polyhedral duality: $\{\mathcal{Q} : \mathcal{Q} \text{ satisfies (1) } \forall P \in \mathcal{D}\}$ is itself a polyhedron.
- $\mathrm{I}(X;Y)$ is convex on this
- Maximum is found by vertex enumeration.
- Smaller $\mathcal{D} \to$ better optimum at computation cost.

Low-complexity mechanisms

- Vertex enumeration is computationally infeasible for large $\mathcal{X}$
- Instead we need explicit methods
- **Idea**: Obfuscate S and U separately. Two ways to do this:
  - *Independent reporting* (IR): obfuscate S and U independently.
  - *Conditional reporting* (CR): obfuscate S; if obfuscated version is 'correct' then slightly obfuscate U; otherwise send random U.
- Low privacy regime: optimal obfuscation methods for S and U known.
- Only 1-dim. optimisation is needed to divide 'privacy budget' between S and U.

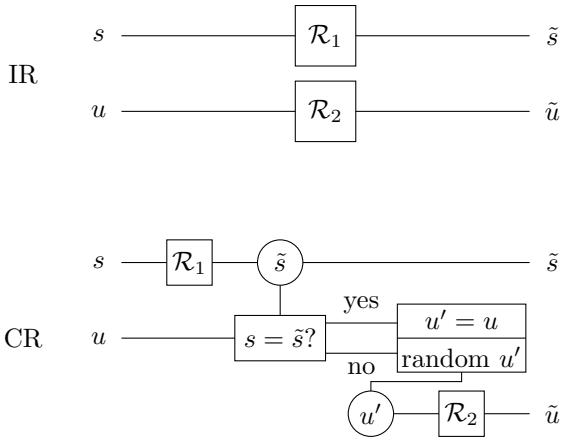

Figure 1: IR and CR

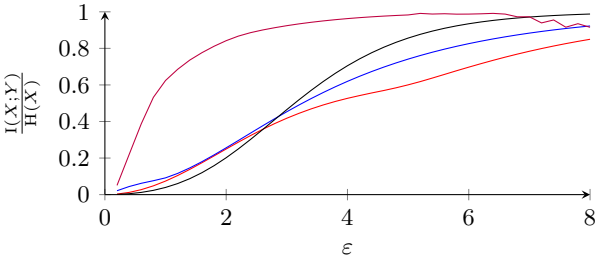

Figure 2: Experiments on synthetic data ($|\mathcal{X}| = 9$). (—— Polyh.Opt., —— RR, —— IR, —— CR)

Experiments

- Synthetic data: average utility over 200 P^*
- Comparison to standard LDP protocol *randomized response* (RR) which obfuscates all of X
- Conclusion: PolyOpt clearly best, IR/CR still better than RR for low ε (high privacy).
- Real-world data: similar results.

Learn more?

`https://arxiv.org/abs/2101.09139`