Optimal and Differentially Private Data Acquisition: Central and Local Mechanisms

Alireza Fallah*, Ali Makhdoumi†, Azarakhsh Malekian‡, Asuman Ozdaglar*

*: MIT  †: Duke University  ‡: University of Toronto

Data Acquisition Problem
- Advances in AI → demand for data is increasing!
- Companies collect data in various ways: offering payments, e.g., Netflix, Nielsen.
- Collecting it as a byproduct of their services, e.g., Google, Facebook.

A common concern: Privacy!
- As a user's data is harnessed, more & more information about her behavior & preferences are uncovered.

Central vs. Local: Platform's objective
- In the central setting, users trust the platform and the platform releases a private estimate.
- In the local setting, users make their data private from the beginning: user i shares $x_i$ with the platform.

Central Differential Privacy
- 

Local Differential Privacy
- 

Central & Local Differential Privacy
- In the central setting, users trust the platform and the platform releases a private estimate.
- In the local setting, users make their data private from the beginning: user i shares $x_i$ with the platform.

Private Data Acquisition Mechanism
- How $(c_i^j)_{n=1}^\infty$ are endogenized?
- A user's privacy sensitivity: per unit cost of privacy loss.

Mechanism outputs
- $c_i^j(i_i, i_j)$: Privacy loss level of user $i_i$.
- $\mathbb{E}[c_i^j(i_i, i_j) | i_i = j, i_j]$ Payment to user $i_i$.

Mechanism outputs
- User reports $x_i$ (central) or $x_i$ (local).

Platform's problem
- The cost of user $i$ with privacy sensitive $c_i$ who reports $c_i^j$:
  $\text{Cost}(c_i^j, x_i) = \mathbb{E}[\text{MSE}(c_i^j, x_i) + c_i^j x_i] = \text{MSE}(c_i^j, x_i) + c_i^j x_i$.

- The cost of a non-participating user: $\mathbb{V}$AR (her best estimate given her data alone).

Central vs. Local: Platform's objective
- We establish that the platform's optimal cost under central differential privacy setting is always weakly smaller.
- Example with two users with $c_j \sim \text{uniform} [1, 2]$.
- However, user's privacy loss can be smaller in the local setting.

Proof Ideas [for the centrally DP case]
- Lower bound: Le Cam's method
  1. Replace sup by average over arbitrary $P_1, P_2 \in \mathcal{P}$
  2. We show $\mathcal{L}(\Theta, \mathcal{P}) \geq (\frac{\mu_{1} - \mu_{2}}{8 K})^2$ where $Q_i$ is the distribution of $x_i$ when $x_i \sim P_i$.
  3. Then, we show that for any $\mathcal{R}_n = \sum_{i=1}^n (\mu_{i} - \mu_{2})^2$.
  4. Finally, choose $P_1$ & $P_2$ as two Bernoullis and optimize over their means' differences.
- Upper bound idea: $\mathbb{V}$AR grows proportional to $c_i^j$ up to some $k$ and then remains constant.