DPella: A DSL for Privacy-Preserving Queries with Accuracy Guarantees

Differential Privacy

Addresses the paradox of learning nothing about an individual while learning useful information about a population.

Q: Average age

$D_1 \cdot \{b\}$

$D_2$

$Q(D_1) \approx Q(D_2)$

$Pr[Q(D_1) = v] \leq e^\epsilon Pr[Q(D_2) = v]$ $Q$ is $\epsilon$-differentially private

Privacy as a DSL

$Q_{\epsilon_1}(D) = Q_{\text{real}}(D) + \text{noise}$

Privacy vs Accuracy

Adding noise to the real answer ensures individual's privacy but goes in detriment of queries' accuracy.

$Pr[|Q_{\epsilon_1}(D) - Q_{\text{real}}(D)| \geq \alpha] \leq \beta$

COMBINING QUERIES

Having two DP-queries $Q_1$ and $Q_2$ over the same database $D$, how do we determine the accuracy of its addition?

$Pr[|(Q_{\epsilon_1} + Q_{\epsilon_2}) - (Q_{\epsilon_1} + Q_{\epsilon_2})| \geq \beta] \leq \beta$

Can we do better?

Accuracy as a DSL

Accuracy is implemented as a new layer in the system that builds on top of DP-computations

INDEPENDENCE

While combining DP-queries $Q_{\epsilon_1}$ and $Q_{\epsilon_2}$ we can compute better bounds over its accuracy if we can ensure that they are independent.

Example

```
> db1 = Tab r 1
> db1 = Data [("Gael", USA), ("Robin", SWE), ("Jodie", USA), ("Kim", AUS)]

> evalCountC :: Tab r s -> Budget -> IO [(Country, Double)]
> evalCountC (dat bud) = evalPINQ prog bud
> prog = countC [UK,..USA] eps dat
```

Privacy vs Accuracy

Adding noise to the real answer ensures individual's privacy but goes in detriment of queries' accuracy.

$Pr[|Q_{\epsilon_1}(D) - Q_{\text{real}}(D)| \geq \alpha] \leq \beta$

COMBINING QUERIES

Having two DP-queries $Q_1$ and $Q_2$ over the same database $D$, how do we determine the accuracy of its addition?

$Pr[|(Q_{\epsilon_1} + Q_{\epsilon_2}) - (Q_{\epsilon_1} + Q_{\epsilon_2})| \geq \beta] \leq \beta$

Can we do better?

Accuracy as a DSL

Accuracy is implemented as a new layer in the system that builds on top of DP-computations

INDEPENDENCE

While combining DP-queries $Q_{\epsilon_1}$ and $Q_{\epsilon_2}$ we can compute better bounds over its accuracy if we can ensure that they are independent.

Example

```
> db1 = Tab r 1
> db1 = Data [("Gael", USA), ("Robin", SWE), ("Jodie", USA), ("Kim", AUS)]

> evalCountC :: Tab r s -> Budget -> IO [(Country, Double)]
> evalCountC (dat bud) = evalPINQ prog bud
> prog = countC [UK,..USA] eps dat
```