Data Security and Privacy for Outsourced Data in the Cloud

Introductory Lecture to the Lab Session

MDD '22 Summer School

Tristan Allard
Univ. Rennes, CNRS, Irisa
tristan.allard@irisa.fr

23rd June 2022

Progress of the Talk

Menu of the Lab Session

Based on this morning's lecture, implement a privacy-preserving DBMS!

Menu of the Lab Session

Based on this morning's lecture, implement a privacy-preserving DBMS!

...

Menu of the Lab Session

Based on this morning's lecture, implement a privacy-preserving DBMS!

...

Given an implementation of a privacy-preserving index for range queries [3]:

- Implement the functions related to privacy: encryption and perturbation.
- Implement a query processing strategy.
- Measure performances and analyze them.

Menu of the Lab Session

Based on this morning's lecture, implement a privacy-preserving DBMS!

...

Given an implementation of a privacy-preserving index for range queries [3]:

- Implement the functions related to privacy: encryption and perturbation.
- Implement a query processing strategy.
- Measure performances and analyze them.

Why? A simple illustration of the privacy-quality-performance tradeoff.

Target technique: PinedRQ [3]

B+-Trees

- Well-known efficient data access structures
- A hierarchy of ranges
- But cleartext data and unprotected structure!

PinedRQ Adaptations

- Perturb the index and encrypt the data (*and the query is in the clear*)[1].
- Requires to adapt the index structure:
  - A level of the tree = histogram.
  - A node = a bin.
  - Bins are pertubed so are the number of pointers.
  - Records are encrypted.
- Satisfies a probabilistic computational variant of *differential privacy* (see below) against a honest-but-curious *cloud*.

[1] What about access pattern leakages?!

PinedRQ by Example

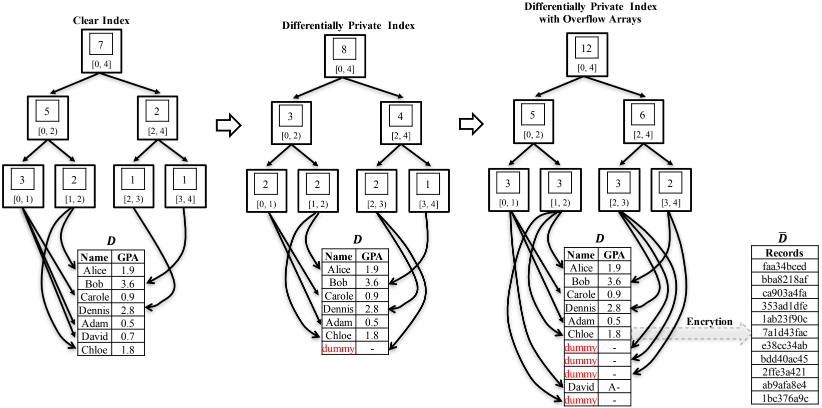

Figure: Sample index at three steps of its construction. Where (1) there is no consistency constraints between nodes, (2) negative noises lead to removing records → store them in *overflow arrays* (size=1 here)

Before starting programming

Lets just have a look to differentially private perturbations.

Progress of the Talk

Differential Privacy and Privacy-Preserving Data Publishing

Privacy-Preserving Data Publishing (PPDP) :

- Publish *personal data* for analysis purposes (accurate aggregate queries)...
- ...while preserving individuals' *privacy* (uncertain point queries)
- Also called *sanitization*

Differential privacy is one way to perform sanitization.

Components of a Sanitization Solution

Three components:

1. **Privacy model**: What does it mean for the data released to be privacy-preserving?
   (e.g., ϵ-differential privacy)
2. **Privacy mechanism**: How to produce the privacy-preserving data to be released?
   (e.g., answer to count queries only and add Laplace random variables (aka perturbations) to counts)
3. **Utility metric**: How much useful is the released data?
   (e.g., variance of the perturbations)

Differential Privacy Paradigm

- Global trends are not private and must be learnt : there must be a knowledge gain !
- Privacy is about each individual value, i.e., **each individual contribution** to the global trend is private.

Differential Privacy Paradigm

A function `f` satisfies differential privacy iif: the possible impact of any individual on its result (its possible outputs) is limited.

Differential Privacy Paradigm

- Global trends are not private and must be learnt : there must be a knowledge gain !
- Privacy is about each individual value, i.e., **each individual contribution** to the global trend is private.

Differential Privacy Paradigm

A function `f` satisfies differential privacy iif: the possible impact of any individual on its result (its possible outputs) is limited.

Intuitions - Mechanism

- Differential privacy originally considers **aggregate queries** (counts, sums)...
- For ex : $q =$ `SELECT COUNT(*) FROM PATIENTS WHERE DIAGNOSIS LIKE 'FLU'`
- How to hide the impact of any single individual participation to the aggregate result ?
  - Add random noise to the true result ! Answer $q(\mathcal{D}) + \texttt{noise}$
  - Such that the noise is **proportional to the participation of one individual**.
  - For ex : `noise` above should be proportionnal to the impact of one individual on q, *i.e.*,, proportionnal to 1 !
  - What if q had been a sum of salaries ?

Initial Model

ϵ-differential privacy (from [1])

A **random function** f satisfies ϵ-differential privacy iff: **For all** $\mathcal{D}$ and $\mathcal{D}'$ **differing in at most one record**, and for any possible output $\mathcal{S}$ of f, then it is true that:

$\mathtt{Pr}[\mathtt{f}(\mathcal{D}) = \mathcal{S}] \leq e^{\epsilon} \times \mathtt{Pr}[\mathtt{f}(\mathcal{D}') = \mathcal{S}]$

Initial Model

ϵ-differential privacy (from [1])

A **random function** $\texttt{f}$ satisfies ϵ-differential privacy iff: **For all** $\mathcal{D}$ and $\mathcal{D}'$ **differing in at most one record**, and for any possible output $\mathcal{S}$ of $\texttt{f}$, then it is true that:
$\texttt{Pr}[\texttt{f}(\mathcal{D}) = \mathcal{S}] \leq e^{\epsilon} \times \texttt{Pr}[\texttt{f}(\mathcal{D}') = \mathcal{S}]$

- $\texttt{f}$: here, an agregate query perturbed by adding random noise to its output
- "For all $\mathcal{D}$ and $\mathcal{D}'$": all possible datasets
- "$\mathcal{D}$ and $\mathcal{D}'$ differing in at most one record": here, $\mathcal{D}$ is $\mathcal{D}'$ with one tuple more or one tuple less (variant: one tuple with different values). Called *neighboring datasets*
- ϵ : the privacy parameter, public, common values: 0.01, 0.1, $\ln 2$, $\ln 3$
- $e^{\epsilon} \times \texttt{Pr}[\dots]$: if one side is zero, the other must be zero too

Query Sensitivity

Different individuals, different impacts. . .

Query Sensitivity

Different individuals, different impacts...

- Presence/absence of an individual on the result of a `COUNT`: at worst +/- 1
- Presence/absence of an individual on the result of a `SUM`: $\mathtt{max}(|domain_{min}|, |domain_{max}|)$

Quantification of the worst-case impact of any possible individual on the output of a query `g`: called *query sensitivity*, and denoted $S_{\mathtt{g}}$.

Query Sensitivity

Different individuals, different impacts...

- Presence/absence of an individual on the result of a `COUNT`: at worst +/- 1
- Presence/absence of an individual on the result of a `SUM`: $\mathtt{max}(|domain_{min}|, |domain_{max}|)$

Quantification of the worst-case impact of any possible individual on the output of a query g: called *query sensitivity*, and denoted S_{g}.

In general: $S_{\mathrm{g}} = \mathtt{max}_{\mathcal{D},\mathcal{D}'} ||\mathrm{g}(\mathcal{D}) - \mathrm{g}(\mathcal{D}')||_1$ where $\mathcal{D}$ and $\mathcal{D}'$ are two neighboring datasets.

Laplace Mechanism for Real-Valued Interactive Queries

A - "Excellent, but how to achieve differential privacy ?"
B - "Just add random noise to each query output, he said !"
A - "But from which distribution ? Uniform ? Gaussian ? Gamma ? Poisson ? . . . ? Any ?"

Laplace Mechanism for Real-Valued Interactive Queries

Given g and ϵ, adding a random variable sampled from a Laplace distribution with mean 0 and scale factor $S_{\mathtt{g}}/\epsilon$ satisfies ϵ-differential privacy [2].

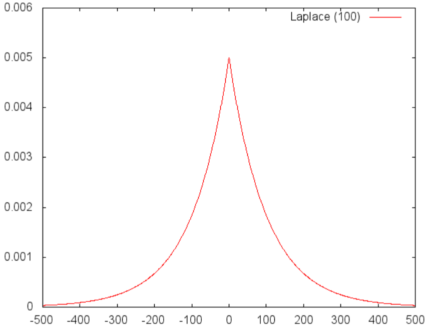

Figure: Laplace (0, 1/0.01)

Laplace probability distribution function : $\mathtt{Pr}_{\mathtt{Lap}(0,b)}(x) = \frac{e^{-|x|/b}}{2b}$

Differential Privacy Properties

- **Self-composability** : composing the outputs of two independant releases sanitized by differentially-private function(s) satisfies differential privacy :
  - Where $\epsilon_{final} = \sum \epsilon_i$ If input datasets are **not** disjoint
  - Or $\epsilon_{final} = \max \epsilon_i$ otherwise
- **No breach from post-processing** :
  - (*Laplace mechanism is independent from data*)
  - Any function applied to a differentially-private input produces a differentially-private output

Inherent Limits

- Noise distribution centered on 0 . . .
  $\Rightarrow$ Sum of noises converges to 0 . . .
  $\Rightarrow$ No unlimited number of queries !
- Composability properties $\Rightarrow$ the privacy parameter ϵ can be seen as a **budget** that must be distributed over the queries to execute ($\epsilon_{final} = \sum \epsilon_i$)

Progress of the Talk

Ready to go?

Go to `https://gitlab.inria.fr/tallard/mdd2022-public`**,**
follow the instructions, and start playing!

Progress of the Talk

[1] C. Dwork.
Differential privacy.
In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, pages 1–12, Berlin, Heidelberg, 2006. Springer-Verlag.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith.
Calibrating noise to sensitivity in private data analysis.
In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[3] C. Sahin, T. Allard, R. Akbarinia, A. E. Abbadi, and E. Pacitti.
A differentially private index for range query processing in clouds.
2018 IEEE 34th International Conference on Data Engineering (ICDE), pages 857–868, 2018.