Probabilistic Relational Reasoning for Differential Privacy

Gilles Barthe       Boris Köpf
Federico Olmedo       Santiago Zanella Béguelin

IMDEA Software Institute, Madrid

POPL 2012
Conflicting requirements!

Need to achieve flexible balance

User privacy

Utility of mining process
Differential Privacy
Dwork [ICALP’06]

- Fix a (symmetric) adjacency relation $\Phi$ on databases
- Fix a privacy budget $\epsilon$

A randomized algorithm $K$ is $\epsilon$-differentially private w.r.t. $\Phi$ iff, for all databases $D_1$ and $D_2$, and events $S$

$$\Phi(D_1, D_2) \iff \Pr[K(D_1) \in S] \leq \exp(\epsilon) \times \Pr[K(D_2) \in S]$$
Differential Privacy Primer

- Fundamentals
  - Laplacian mechanism
  - Composition theorems

- Expanding frontiers
  - Mechanisms: exponential, median...
  - Algorithms: streaming/graph/... algorithms
  - Definitions: approximate differential privacy, pan privacy...

Language-based tool support available

Increasingly complex, but not supported by existing tools!
Our Contribution: CERTIPriv

• Allows reasoning about approximate quantitative properties of randomized computations
• Built from first principles and fully formalized in Coq
• Machine-checked proofs of differential privacy
  • Correctness of Laplacian and Exponential mechanisms
  • State-of-art graph and streaming algorithms
• Generalizes CERTICRYPT and opens new applications to crypto
Differential privacy as quantitative 2-safety

- $K$ is $(\varepsilon, \delta)$-diff. private w.r.t. $\Phi$ iff for all $D_1$ and $D_2$ and $S$
  \[ \Phi(D_1, D_2) \implies \Pr[K(D_1) \in S] \leq \exp(\varepsilon) \times \Pr[K(D_2) \in S] + \delta \]

  (Quantitative) relational post-condition

  Relational pre-condition

- We propose a quantitative probabilistic relational Hoare Logic
  \[ c_1 \sim_{\alpha,\delta} c_2 : \Phi \implies \Psi \]
  such that $c$ is $(\varepsilon, \delta)$-diff. private w.r.t. $\Phi$ iff
  \[ c \sim_{\exp(\varepsilon),\delta} c : \Phi \implies \Xi \]
  Needs to be lifted to distributions
Characterizing differential privacy

\[ c_1 \sim_{\alpha, \delta} c_2 : \Phi \Rightarrow \Psi \text{ is valid iff for all } D_1 \text{ and } D_2 \]
\[ \Phi(D_1, D_2) \implies \text{lift}_{\alpha, \delta} \Psi ([c_1] D_1) ([c_2] D_2) \]

We define \( \alpha \)-distance such that:

- \( c \) is \((\varepsilon, \delta)\)-diff. private w.r.t. \( \Phi \) iff for all \( D_1 \) and \( D_2 \)
  \[ \Phi(D_1, D_2) \implies \Delta_\alpha([c_1] D_1, [c_2] D_2) \leq \delta \]

- Fundamental property of lifting
  \[ \Delta_\alpha(\mu_1, \mu_2) \leq \delta \iff \text{lift}_{\alpha, \delta} \equiv \mu_1 \mu_2 \]
Lifting relations to distributions

Given \( R = \{(a, x), (a, y), (c, y), (d, z)\} \), \( \alpha = 1.1 \) and \( \delta = 0.01 \)

\[
\begin{align*}
\delta_a &= \max\{0, 0.33 - \alpha (p_1 + p_2)\} \\
\delta_a + \delta_b + \delta_c + \delta_d &\leq \delta
\end{align*}
\]

Witness distribution

<table>
<thead>
<tr>
<th>( X \times Y )</th>
<th>( \mu(\cdot, \cdot) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (a, x) )</td>
<td>0.10</td>
</tr>
<tr>
<td>( (a, y) )</td>
<td>0.20</td>
</tr>
<tr>
<td>( (c, y) )</td>
<td>0.20</td>
</tr>
<tr>
<td>( (d, z) )</td>
<td>0.30</td>
</tr>
<tr>
<td>...</td>
<td>0</td>
</tr>
</tbody>
</table>
Selected rules

Sequential composition

\[ \models c_1 \sim_{\alpha, \delta} c_2 : \Psi \Rightarrow \Phi' \quad \models c_1' \sim_{\alpha', \delta'} c_2' : \Phi' \Rightarrow \Phi \]
\[ \models c_1 ; c_1' \sim_{\alpha \alpha', \delta + \delta'} c_2 ; c_2' : \Psi \Rightarrow \Phi \]

Laplacian Mechanism

\[ \models x \leftarrow \mathcal{L}_\lambda(r) \sim_{\exp(\epsilon), 0} y \leftarrow \mathcal{L}_\lambda(s) : |r\langle 1 \rangle - s\langle 2 \rangle| \leq \lambda \epsilon \Rightarrow x\langle 1 \rangle = y\langle 2 \rangle \]
Application: Vertex Cover
Gupta et al. [SODA ’10]

VertexCover\((V, E) \epsilon\)

1 \(\pi \leftarrow \text{nil}; \ n \leftarrow |V|; \ i \leftarrow 0;\)
2 \(\text{while } E \not\subseteq n \text{ do}\)
3 \(v \leftarrow \text{pick}(V, E)\epsilon, n, i);\)
4 \(\pi \leftarrow v :: \pi;\)
5 \(V \leftarrow V \setminus \{v\}; \ E \leftarrow E \setminus (\{v\} \times V);\)
6 \(\text{end} \leftarrow i + 1\)
7 \(\text{end}\)

pick\((V, E)\epsilon, \alpha, d, c, a\)\n
\(\pi = [b, g, e, h, l, k,\)
\(\text{VertexCover}(V, E, \epsilon)\).
\(\text{VertexCover}(V(V)E', \epsilon):\)
\(V\langle 1 \rangle = V\langle 2 \rangle \wedge E\langle 1 \rangle = E\langle 2 \rangle \cup \{(t, u)\} \implies \pi\langle 1 \rangle = \pi\langle 2 \rangle\)
Conclusions

- Framework for reasoning about quantitative relational properties of randomized computations
  - Laplacian and Exponential mechanisms
  - Differential privacy for streaming and graph algorithms
  - Asymmetric logic

- Further work:
  - Computational differential privacy
  - Hash functions unto elliptic curves and statistical zero-knowledge

- Challenge: logic for arbitrary quantitative relational properties
Thanks for your attention!
Define \( \alpha \)-distance as:

\[
\Delta_\alpha(d_1, d_2) = \max_A \left( \max(d_1 \ 1_A - \alpha \ (d_2 \ 1_A), d_2 \ 1_A - \alpha \ (d_1 \ 1_A)) \right)
\]

\((\alpha, \delta)\)-lifting of relations to distributions:

\[
\text{lift}_{\alpha, \delta} \ R \ (d_1 : \mathcal{D}_A) \ (d_2 : \mathcal{D}_B) = \exists (d : \mathcal{D}_{A*B}) , \\
\pi_1(d) \leq d_1 \land \Delta_\alpha(\pi_1(d), d_1) \leq \delta \land \\
\pi_2(d) \leq d_2 \land \Delta_\alpha(\pi_2(d), d_2) \leq \delta \land \text{range} \ R \ d
\]
Output perturbation makes numerical queries $\varepsilon$-diff. private

- The $\Phi$-sensitivity of a query $f : D \rightarrow \mathbb{R}$ is defined as:
  \[
  \Delta(f) = \max\{f(D_1) - f(D_2) \mid \Phi(D_1, D_2)\}
  \]

- The randomized computation
  \[
  K(D) = f(D) + \text{Lap}(\Delta(f)/\varepsilon)
  \]
  is $\varepsilon$-differentially private

Density proportional to $\exp(-\varepsilon/\Delta(f))$
Composition theorems

If $K_1$ is $(\varepsilon_1, \delta_1)$-diff. private and $K_2$ is $(\varepsilon_2, \delta_2)$-diff. private

- Sequential composition

$$K_1$$

$$K_2$$

$(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-diff. private

- Parallel composition

$$K_1$$

$$K_2$$

$(\max\{\varepsilon_1, \varepsilon_2\}, \max\{\delta_1, \delta_2\})$-diff. private

$K_1$ and $K_2$ depend on disjoint parts of the database