CS 591: Formal Methods in Security and Privacy

Differential Privacy

Marco Gaboardi
gaboardi@bu.edu

Alley Stoughton
stough@bu.edu
From the previous classes
(\(\epsilon, \delta\))-Differential Privacy

**Definition**

Given \(\epsilon, \delta \geq 0\), a probabilistic query \(Q: X^n \rightarrow \mathbb{R}\) is \((\epsilon, \delta)\)-differentially private iff for all adjacent database \(b_1, b_2\) and for every \(S \subseteq \mathbb{R}\):

\[
\Pr[Q(b_1) \in S] \leq \exp(\epsilon) \Pr[Q(b_2) \in S] + \delta
\]
Releasing privately the mean of Some Data

\[ \text{Mean}(d : \text{private data}) : \text{public real} \]
\[ i:=0; \]
\[ s:=0; \]
\[ \text{while } (i<\text{size}(d)) \]
\[ \quad s:=s + d[i] \]
\[ \quad i:=i+1; \]
\[ z:=\text{Laplace}(\text{sens}/\text{eps},0) \]
\[ z:=(s/i)+z \]
\[ \text{return } z \]
Differential Privacy as a Relational Property

$c$ is differentially private if and only if for every $m_1 \sim m_2$ (extending the notion of adjacency to memories):

$\{c\}_{m_1} = \mu_1$ and $\{c\}_{m_2} = \mu_2$ implies $\Delta_\varepsilon(\mu_1, \mu_2) \leq \delta$
apRHL

Indistinguishability parameter

\[ \epsilon, \delta \vdash c_1 \sim c_2 : P \Rightarrow Q \]

Probabilistic Program

Probabilistic Program

Precondition (a logical formula)

Postcondition (a logical formula)
Validity of apRHL judgments

We say that the quadruple $\vdash_{\varepsilon,\delta} c_1 \sim c_2 : P \Rightarrow Q$ is valid if and only if for every pair of memories $m_1, m_2$ such that $P(m_1, m_2)$ we have:

\[
\{c_1\}_{m_1} = \mu_1 \quad \text{and} \quad \{c_2\}_{m_2} = \mu_2 \quad \text{implies} \quad Q_{\varepsilon,\delta^*}(\mu_1, \mu_2).
\]
Given two distributions $\mu_1 \in D(A)$, and $\mu_2 \in D(B)$, we have an $R-(\varepsilon, \delta)$-coupling between them, for $R \subseteq A \times B$ and $0 \leq \delta \leq 1$, $\varepsilon \geq 0$, if there are two joint distributions $\mu_L, \mu_R \in D(A \times B)$ such that:

1) $\pi_1(\mu_L) = \mu_1$ and $\pi_2(\mu_R) = \mu_2$,

2) the support of $\mu_L$ and $\mu_R$ is contained in $R$. That is, if $\mu_L(a,b) > 0$, then $(a,b) \in R$, and if $\mu_R(a,b) > 0$, then $(a,b) \in R$.

3) $\Delta_\varepsilon(\mu_L, \mu_R) \leq \delta$
Example of \( R-(\varepsilon, \delta) \)-Coupling

\[
\begin{align*}
\mu_1 & \quad \mu_2 \\
\begin{array}{c}
\text{OO} \quad 0.25 \\
\text{O1} \quad 0.25 \\
\text{1O} \quad 0.25 \\
\text{11} \quad 0.25 \\
\end{array} & \begin{array}{c}
\text{OO} \quad 0.20 \\
\text{O1} \quad 0.25 \\
\text{1O} \quad 0.25 \\
\text{11} \quad 0.30 \\
\end{array}
\end{align*}
\]

\[
R(a, b) = \{ a = b \}
\]

\[
\begin{array}{c|c|c|c|c}
\mu_L & \text{OO} & \text{O1} & \text{1O} & \text{11} \\
\hline
\text{OO} & 0.25 & \text{} & \text{} & \text{} \\
\text{O1} & \text{} & 0.25 & \text{} & \text{} \\
\text{1O} & \text{} & \text{} & 0.25 & \text{} \\
\text{11} & \text{} & \text{} & \text{} & 0.25 \\
\end{array}
\quad
\begin{array}{c|c|c|c|c}
\mu_R & \text{OO} & \text{O1} & \text{1O} & \text{11} \\
\hline
\text{OO} & 0.20 & \text{} & \text{} & \text{} \\
\text{O1} & \text{} & 0.25 & \text{} & \text{} \\
\text{1O} & \text{} & \text{} & 0.25 & \text{} \\
\text{11} & \text{} & \text{} & \text{} & 0.30 \\
\end{array}
\]

\[
\Delta_{0.3} (\mu_L, \mu_R) = 0
\]
Example of R-(\(\varepsilon, \delta\))-Coupling

\[ R(a,b) = \{ a \leq b \} \]

\[ \Delta_0(\mu_L, \mu_R) = 0.05 \]
Probabilistic Relational Hoare Logic

Skip

\[ \vdash_{0,0} \text{skip} \sim \text{skip} : P \Rightarrow P \]
\[\begin{align*}
\text{Probabilistic Relational Hoare Logic} \\
\text{Skip} \\
\hline \\
\vdash_{\varepsilon,0} x_1 := \$ \ Lap(\varepsilon, y_1) \\
\sim \\
\vdash_{\varepsilon,0} x_2 := \$ \ Lap(\varepsilon, y_2) \\
\vdash : |y_1 - y_2| \leq 1 \implies \quad =
\end{align*}\]
Today
Composition
Composition

A. Haeberlen

Promising approach: Differential privacy

Differential Privacy: Ensuring that the presence/absence of an individual has a negligible statistical effect on the query's result.

Trade-off between utility and privacy.

$M_1$ is $(\varepsilon_1, \delta_1)$-DP
Composition

Differential Privacy: the idea of ensuring that the presence/absence of an individual has a negligible statistical effect on the query's result.

Trade-off between utility and privacy.

\[ M_1 \text{ is } (\varepsilon_1, \delta_1)\text{-DP} \]
\[ M_2 \text{ is } (\varepsilon_2, \delta_2)\text{-DP} \]
Composition

Differential Privacy: the idea

A. Haeberlen

USENIX Security (August 12, 2011)

Private data

N(flue, >1955)?

826±10

N(brain tumor, 05-22-1955)?

3 ±700

Differential Privacy:
Ensuring that the presence/absence of an individual has a
negligible statistical effect on the query's result.

Trade-off between utility and privacy.

\[ M_1 \text{ is } (\varepsilon_1, \delta_1)-\text{DP} \]

\[ M_2 \text{ is } (\varepsilon_2, \delta_2)-\text{DP} \]

\[ \ldots \]

\[ M_n \text{ is } (\varepsilon_k, \delta_k)-\text{DP} \]
Composition

The overall process is \((\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_k, \delta_1 + \delta_2 + \ldots + \delta_k)\)-DP
Let $M_1:DB \rightarrow R_1$ be a $(\epsilon_1, \delta_1)$-differentially private program and $M_2:DB \rightarrow R_2$ be a $(\epsilon_2, \delta_2)$-differentially private program. Then, their composition $M_{1,2}:DB \rightarrow R_1 \times R_2$ defined as

$$M_{1,2}(D) = (M_1(D), M_2(D))$$

is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-differentially private.
Question: Why composition is important?
Question: Why composition is important?

Answer: Because it allows to reason about privacy as a budget!
Composition

Budget = $\varepsilon_{\text{global}}$
Differential Privacy: the idea

A. Haeberlen

Promising approach: Differential privacy

USENIX Security (August 12, 2011)

Private data

N(flu, >1955)?

826±10

N(brain tumor, 05-22-1955)?

3 ±700

Noise

Differential Privacy: Ensuring that the presence/absence of an individual has a negligible statistical effect on the query's result.

Trade-off between utility and privacy.

\[ M_1 \text{ is } \varepsilon_1\text{-DP} \]

\[ \text{Budget}=\varepsilon_{\text{global}} \]
Composition

Differential Privacy: the idea

A. Haeberlen

USENIX Security (August 12, 2011)

Private data

N(flu, >1955)?

826±10

N(brain tumor, 05-22-1955)?

3 ±700

Differential Privacy:

Ensuring that the presence/absence of an individual has a negligible statistical effect on the query’s result.

Trade-off between utility and privacy.

\[ \text{Budget} = \varepsilon_{\text{global}} - \varepsilon_1 \]

\[ M_1 \text{ is } \varepsilon_1\text{-DP} \]
Composition

Differential Privacy: the idea

A. Haeberlen

USENIX Security (August 12, 2011)

Private data

N(flue, >1955)?

826±10

N(brain tumor, 05-22-1955)?

3 ±700

Noise

Differential Privacy: Ensuring that the presence/absence of an individual has a negligible statistical effect on the query's result.

Trade-off between utility and privacy.

\[ M_1 \text{ is } \varepsilon_1\text{-DP} \]

\[ M_2 \text{ is } \varepsilon_2\text{-DP} \]

\[ \text{Budget=} \varepsilon_{\text{global}} \text{ - } \varepsilon_1 \]
Composition

Differential Privacy: the idea of ensuring that the presence/absence of an individual has a negligible statistical effect on the query's result. This involves a trade-off between utility and privacy.

\[ \text{Budget} = \varepsilon_{\text{global}} - \varepsilon_1 - \varepsilon_2 \]

- \( M_1 \) is \( \varepsilon_1\)-DP
- \( M_2 \) is \( \varepsilon_2\)-DP
Composition

Differential Privacy: the idea

A. Haeberlen

Promising approach: Differential privacy

Private data

N(\text{flu, } >1955)?

826\pm 10

N(\text{brain tumor, 05-22-1955})?

3\pm 700

Noise

Differential Privacy: Ensuring that the presence/absence of an individual has a negligible statistical effect on the query's result.

Trade-off between utility and privacy.

\text{Budget} = \varepsilon_{\text{global}} - \varepsilon_1 - \varepsilon_2 \ldots

M_1 \text{ is } \varepsilon_1\text{-DP}

M_2 \text{ is } \varepsilon_2\text{-DP}

\ldots

M_n \text{ is } \varepsilon_n\text{-DP}
A. Haeberlen

Promising approach: Differential privacy

Differential Privacy:
Ensuring that the presence/absence of an individual has a
negligible statistical effect on the query's result.

Composition

Budget = $\varepsilon_{\text{global}} - \varepsilon_1 - \varepsilon_2 \ldots - \varepsilon_n$

$M_1$ is $\varepsilon_1$-DP

$M_2$ is $\varepsilon_2$-DP

\ldots

$M_n$ is $\varepsilon_n$-DP
X = \{0,1\}^3 \text{ ordered wrt binary encoding.}

\[ q_{000}^*(D) = 0.3 + L\left(\frac{1}{\varepsilon_1}\right) \]
\[ q_{001}^*(D) = 0.4 + L\left(\frac{1}{\varepsilon_2}\right) \]
\[ q_{010}^*(D) = 0.6 + L\left(\frac{1}{\varepsilon_3}\right) \]
\[ q_{011}^*(D) = 0.6 + L\left(\frac{1}{\varepsilon_4}\right) \]
\[ q_{100}^*(D) = 0.6 + L\left(\frac{1}{\varepsilon_5}\right) \]
\[ q_{101}^*(D) = 0.9 + L\left(\frac{1}{\varepsilon_6}\right) \]
\[ q_{110}^*(D) = 1 + L\left(\frac{1}{\varepsilon_7}\right) \]
\[ q_{111}^*(D) = 1 + L\left(\frac{1}{\varepsilon_8}\right) \]
Marginals

\[ D \in X^{10} = \]

\[ q^*_1(D) = 0.4 + L(1/(10*\varepsilon_1)) \]
\[ q^*_2(D) = 0.3 + L(1/(10*\varepsilon_2)) \]
\[ q^*_3(D) = 0.4 + L(1/(10*\varepsilon_3)) \]

\[ \text{Budget} = \varepsilon_{global} - \varepsilon_1 - \varepsilon_2 - \varepsilon_3 \]
Budget=$\varepsilon_{global} - \varepsilon_1 - \varepsilon_2 - \varepsilon_3 - \varepsilon_4$
- $\varepsilon_5 - \varepsilon_6 - \varepsilon_7 - \varepsilon_8$

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>I3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>I4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>I5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I6</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>I7</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>I8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I9</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>I10</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

|   | margin | .4+$Y_1$ | .3+$Y_2$ | .4+$Y_3$ |
Releasing partial sums

```
DummySum(d : {0,1} list) : real list
i:= 0;
s:= 0;
r:= [];
while (i<size d)
  s:= s + d[i]
  z:= Lap(eps,s)
  r:= r ++ [z];
i:= i+1;
return r
```

I am using the easycrypt notation here where \( \text{Lap}(\epsilon, a) \) corresponds to adding to the value \( a \) noise from the Laplace distribution with \( b=1/\epsilon \) and mean \( \mu=0 \).
Global Sensitivity

$$GS_q = \max\left\{ |q(D) - q(D')| \mid \text{s.t. } D \sim D' \right\}$$
Probabilistic Relational Hoare Logic
Composition

\[ \vdash \epsilon_1, \delta_1 C_1 \sim C_2 : P \Rightarrow R \quad \vdash \epsilon_2, \delta_2 C_1' \sim C_2' : R \Rightarrow S \]

\[ \vdash \epsilon_1 + \epsilon_2, \delta_1 + \delta_2 C_1 ; C_1' \sim C_2 ; C_2' : P \Rightarrow S \]
Releasing partial sums

```
DummySum(d : {0,1} list) : real list
    i:=0;
    s:=0;
    r:=[];
    while (i<size d)
        z:= $Lap(\epsilon, d[i])$
        s:= s + z
        r:= r ++ [s];
        i:= i+1;
    return r
```
Parallel Composition

Let $M_1: DB \rightarrow R$ be a $(\varepsilon_1, \delta_1)$-differentially private program and $M_2: DB \rightarrow R$ be a $(\varepsilon_2, \delta_2)$-differentially private program. Suppose that we partition $D$ in a data-independent way into two datasets $D_1$ and $D_2$. Then, the composition $M_{1,2}: DB \rightarrow R$ defined as

$$MP_{1,2}(D) = (M_1(D_1), M_2(D_2))$$

is $(\max(\varepsilon_1, \varepsilon_2), \max(\delta_1, \delta_2))$-differentially private.
Probabilistic Relational Hoare Logic Composition

\[ \vdash \varepsilon_1, \delta_1 \text{c}_1 \sim \text{c}_2 : \text{P} \Rightarrow \text{R} \quad \vdash \varepsilon_2, \delta_2 \text{c}_1' \sim \text{c}_2' : \text{R} \Rightarrow \text{S} \]

\[ \vdash \varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2 \text{c}_1 ; \text{c}_1' \sim \text{c}_2 ; \text{c}_2' : \text{P} \Rightarrow \text{S} \]
apRHL

while

\[ \text{P} / \ \text{e}<1> \leq 0 \implies \neg \text{b1}<1> \]

\[ \vdash \varepsilon_k, \delta_k \ \text{c1~c2:} \ \text{P} / \ \text{b1}<1> / \ \text{b2}<2> / \ \text{k}=\text{e}<1> / \ \text{e}<1> \leq n \implies \text{P} / \ \text{b1}<1> = \text{b2}<2> / \ \text{k} < \text{e}<1> \]

\[ \vdash \Sigma \varepsilon_k, \Sigma \delta_k : \ \text{P} / \ \text{b1}<1> = \text{b2}<2> / \ \text{e}<1> \leq n \implies \text{P} / \ \neg \text{b1}<1> / \ \neg \text{b2}<2> \]
Properties of Differential Privacy
Some important properties

- Resilience to post-processing
- Group privacy
- Composition
Some important properties

- Resilience to post-processing
- Group privacy
- Composition

We will look at them in the context of $(\varepsilon,0)$-differential privacy.
Resilience to Post-processing

M is $\varepsilon$-DP
Resilience to Post-processing

M is \( \varepsilon \)-DP \( \rightarrow \) f
Resilience to Post-processing

\(f \circ M \text{ is } \epsilon\text{-DP}\)
Resilience to Post-processing

Question: Why is resilience to post-processing important?
Resilience to Post-processing

**Question:** Why is resilience to post-processing important?

**Answer:** Because it is what allows us to publicly release the result of a differentially private analysis!
Differential Privacy: the idea

A. Haeberlen

USENIX Security (August 12, 2011)

Private data

\[ N(\text{flue, } >1955) = 826 \pm 10 \]

\[ N(\text{brain tumor, } 05-22-1955) = 3 \pm 700 \]

Noise

Differential Privacy: Ensuring that the presence/absence of an individual has a negligible statistical effect on the query's result.

Trade-off between utility and privacy.

Mike is \( \varepsilon \)-DP

A conjunction query \( q \) on a dataset \( D \) gives the \( k \)-way marginal statistics at \( q \) of the dataset. Answering \( k \)-way marginals is also the base for computing contingency tables.

A generalization of counting queries are statistical queries, often called also linear queries.

Definition 1.5 (Statistical Queries).

Let \( q : X \rightarrow [0, 1] \) be a bounded function returning an element in the interval \([0, 1]\) for each on record in \( X \). A statistical query is a function \( q : X \rightarrow [0, 1] \) averaging the value of \( q \) on all the records of a dataset \( D \). In symbols:

\[
q(D) = \frac{1}{n} \sum_{i=1}^{n} q(d_i)
\]

Notice that once again we use the same symbol \( q \) for the function and the statistical query characterized by this function. Notice also that the formula defining a statistical query is the same as the one defining a counting query, what changes is just the fact that \( q \) is a predicate for a counting query and an arbitrary (bounded) function for a statistical query. As one expects from their name, statistical queries allow to define more general statistics than the ones that can be defined by using counting queries.

\[
y(y(x)) = \begin{cases} 
I & \text{if } y = x \\
0 & \text{otherwise}
\end{cases}
\]

\[
Pr[M(D) = r] \leq e^\varepsilon Pr[M(D') = r]
\]
Group Privacy

\[ \Pr[\mathcal{M}(D) \in S] \leq \exp(k\varepsilon) \Pr[\mathcal{M}(D') \in S] \]
Question: Why is group privacy important?
**Question:** Why is group privacy important?

**Answer:** Because it allows to reason about privacy at different level of granularities!
Sometimes is more convenient to think in terms of Privacy Budget: \( \text{Budget} = \varepsilon_{\text{global}} - \sum \varepsilon_{\text{local}} \)

Sometimes is more convenient to think in terms of epsilon: \( \varepsilon_{\text{global}} = \sum \varepsilon_{\text{local}} \)

Also making them uniforms is sometimes more informative.
Budget = $\varepsilon_{global} - \varepsilon_1 - \varepsilon_2 - \varepsilon_3 - \varepsilon_4$

$- \varepsilon_5 - \varepsilon_6 - \varepsilon_7 - \varepsilon_8$

$\varepsilon_{global} = \varepsilon + \varepsilon + \varepsilon + \varepsilon + \varepsilon + \varepsilon + \varepsilon + \varepsilon = 8\varepsilon$

Budget = $\varepsilon_{global} - \varepsilon_1 - \varepsilon_2 - \varepsilon_3$

$\varepsilon_{global} = \varepsilon + \varepsilon + \varepsilon = 3\varepsilon$

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>I3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>I4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>I5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I6</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>I7</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>I8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I9</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>I10</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

margin $+.4Y_1$. $+.3Y_2$. $+.4Y_3$