Interpreting ε of Differential Privacy in Terms of Advantage in Guessing or Approximating Sensitive Attributes

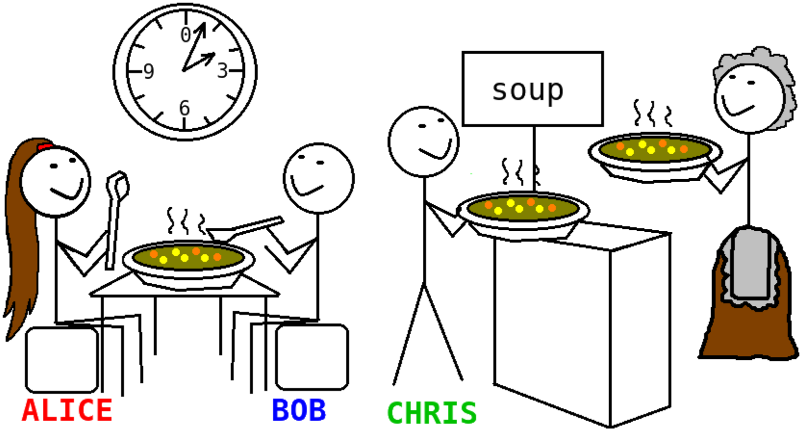
Alisa Pankova and Peeter Laud

07.08.2022

This research has been funded by the Air Force Research laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) under contract FA8750-16-C-0011. The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government. This research has also been supported by European Regional Development Fund through the Estonian Centre of Excellence in ICT Research (EXCITE), and by the European Union under the H2020 Programme Grant Agreement No. 830929 “CyberSec4Europe”.

Background

Cafeteria computes average eating time of math students.



Privacy question

- Cafeteria computes a table *t*.

student name	faculty	time (min)
Alice	math	20
Bob	math	15
Eve	computer science	25
...	...	...
...	...	...

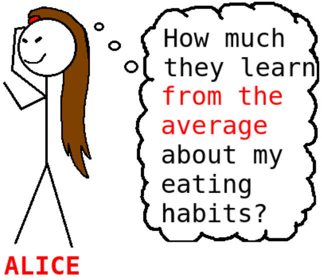
- The analyst will see only the average.

```
SELECT AVG(time) FROM t WHERE faculty = math;
```

Privacy question

- Cafeteria computes a table t.

student name	faculty	time (min)
Alice	math	20
Bob	math	15
Eve	computer science	25
...	...	...
...	...	...



- The analyst will see only the average.

```
SELECT AVG(time) FROM t WHERE faculty = math;
```

Privacy issue

Table t

student name	faculty	time (min)
Alice	math	20
Bob	math	15
Chris	math	10

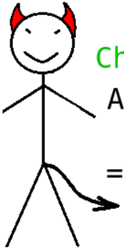
```
SELECT AVG(time)
FROM t
WHERE faculty = math;
```

Privacy issue

Table t

student name	faculty	time (min)
Alice	math	20
Bob	math	15
Chris	math	10

```
SELECT AVG(time)
FROM t
WHERE faculty = math;
```



Bob ate 15 minutes
Chris ate 10 minutes
Average is 15 minutes

=> Alice ate 20 minutes!

Privacy issue

Table t

student name	faculty	time (min)
Alice	math	20
Bob	math	15
Chris	math	10

```
SELECT AVG(time)
FROM t + noise
WHERE faculty = math;
```

ε-differential privacy for particular attributes

t

name	faculty	time (min)
Alice	math	20
Bob	math	15
Chris	math	10

t'

name	faculty	time (min)
Alice	math	25
Bob	math	15
Chris	math	10

ε-differential privacy for particular attributes

t

name	faculty	time (min)
Alice	math	20
Bob	math	15
Chris	math	10

t'

name	faculty	time (min)
Alice	math	25
Bob	math	15
Chris	math	10

Define distance $d(\cdot, \cdot)$ between two tables as the distance in some attribute of some row. We have $d(t, t') = 5$.

Let $f : X \to Y$ be a *query*.

Differential privacy: For all $Y' \subseteq Y$, for all tables $t' \in X$: $\dfrac{Pr(f(t) \in Y')}{Pr(f(t') \in Y')} \leq e^{\varepsilon \cdot d(t,t')}$

Which ε is good enough?

$$\frac{Pr(f(t) \in Y')}{Pr(f(t') \in Y')} \leq e^{\varepsilon \cdot d(t,t')} \iff Pr(f(t) \in Y') \leq e^{\varepsilon \cdot d(t,t')} \cdot Pr(f(t') \in Y') \ .$$

Which ε is good enough?

$$\frac{Pr(f(t) \in Y')}{Pr(f(t') \in Y')} \leq e^{\varepsilon \cdot d(t,t')} \iff Pr(f(t) \in Y') \leq e^{\varepsilon \cdot d(t,t')} \cdot Pr(f(t') \in Y') \ .$$

- The "goodness" of ε is linked to the distance $d(\cdot, \cdot)$.
  - $Pr(f(t) \in Y') \leq e^{\varepsilon \cdot d(t,t')} \cdot Pr(f(t') \in Y')$;

Which ε is good enough?

$$\frac{Pr(f(t) \in Y')}{Pr(f(t') \in Y')} \leq e^{\varepsilon \cdot d(t,t')} \iff Pr(f(t) \in Y') \leq e^{\varepsilon \cdot d(t,t')} \cdot Pr(f(t') \in Y') \ .$$

- The "goodness" of ε is linked to the distance $d(\cdot,\cdot)$.
  - $Pr(f(t) \in Y') \leq e^{\varepsilon \cdot d(t,t')} \cdot Pr(f(t') \in Y')$;
  - $Pr(f(t) \in Y') \leq e^{\alpha\varepsilon \cdot \frac{d(t,t')}{\alpha}} \cdot Pr(f(t') \in Y')$ for any $\alpha \in \mathbb{R}^+$.
  - Hence, there is no "universally good" ε

Which ε is good enough?

$$\frac{Pr(f(t) \in Y')}{Pr(f(t') \in Y')} \leq e^{\varepsilon \cdot d(t,t')} \iff Pr(f(t) \in Y') \leq e^{\varepsilon \cdot d(t,t')} \cdot Pr(f(t') \in Y') .$$

- The "goodness" of ε is linked to the distance $d(\cdot, \cdot)$.
  - $Pr(f(t) \in Y') \leq e^{\varepsilon \cdot d(t,t')} \cdot Pr(f(t') \in Y')$;
  - $Pr(f(t) \in Y') \leq e^{\alpha\varepsilon \cdot \frac{d(t,t')}{\alpha}} \cdot Pr(f(t') \in Y')$ for any $\alpha \in \mathbb{R}^+$.
  - Hence, there is no "universally good" ε
- **d-privacy:** treat $\varepsilon \cdot d(t, t')$ as a new distance $d'(t, t')$.

Which ε is good enough?

$$\frac{Pr(f(t) \in Y')}{Pr(f(t') \in Y')} \leq e^{\varepsilon \cdot d(t,t')} \iff Pr(f(t) \in Y') \leq e^{\varepsilon \cdot d(t,t')} \cdot Pr(f(t') \in Y') \ .$$

- The "goodness" of ε is linked to the distance $d(\cdot, \cdot)$.
  - $Pr(f(t) \in Y') \leq e^{\varepsilon \cdot d(t,t')} \cdot Pr(f(t') \in Y')$;
  - $Pr(f(t) \in Y') \leq e^{\alpha\varepsilon \cdot \frac{d(t,t')}{\alpha}} \cdot Pr(f(t') \in Y')$ for any $\alpha \in \mathbb{R}^{+}$.
  - Hence, there is no "universally good" ε
- **d-privacy:** treat $\varepsilon \cdot d(t, t')$ as a new distance $d'(t, t')$.
- How exactly should ε (or the distance d') be defined?

Guessing advantage of numerical attributes

- Attacker's question: how long has Alice been eating?

Guessing advantage of numerical attributes

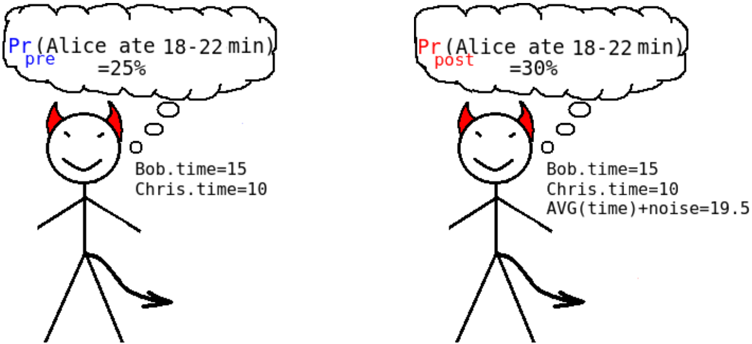
- Attacker's question: how long has Alice been eating?
  - How likely the attacker says that it was exactly 20 minutes?

Guessing advantage of numerical attributes

- Attacker's question: how long has Alice been eating?
  - How likely the attacker says that it was exactly 20 minutes?
  - What if the attacker says that it was 20.001 minutes?

Guessing advantage of numerical attributes

- Attacker's question: how long has Alice been eating?
  - How likely the attacker says that it was exactly 20 minutes?
  - What if the attacker says that it was 20.001 minutes?



- Guessing advantage: $|Pr_{post} - Pr_{pre}|$.

Defining guessing advantage

- Set X of values
- actual value $x \in X$
- Probability distribution π over it (the prior)
- Data release mechanism $\mathcal{M} : X \xrightarrow{\$} Z$

- Attacker's goal: $g : X \to \mathcal{P}(X)$
- Attacker's prior knowledge: $k \in \mathbf{Eqv}(X)$
  - Consider $X := x/k$

$$\eta := \sup_{Y \subseteq Z} \left(\Pr_{\mathbf{X} \sim \pi} [\mathbf{X} \in g(x) | \mathcal{M}(\mathbf{X}) \in Y] - \Pr_{\mathbf{X} \sim \pi} [\mathbf{X} \in g(x)] \right)$$

Prior and posterior probability of a "correct"guess

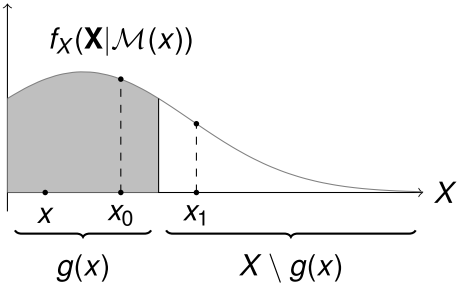
- Pr_{pre} is the *prior* probability of X that is known in advance.

- Let f_X be the probability density function of the prior distribution of X.
- Let $g(x)$ be the set of guesses considered "correct".
- Applying Bayesian inference, we get

$$
\begin{aligned}
Pr_{post}(g(x)) &= Pr_{pre}(g(x)|\mathcal{M}(x)) = \int_{g(x)} f_X(x|\mathcal{M}(x))dx \\
&= \frac{\int_{g(x)} f_X(x|\mathcal{M}(x))dx}{\int_X f_X(x|\mathcal{M}(x))dx} = \frac{1}{1 + \frac{\int_{X \setminus g(x)} f_X(x|\mathcal{M}(x))dx}{\int_{g(x)} f_X(x'|\mathcal{M}(x))dx'}}
\end{aligned}
$$

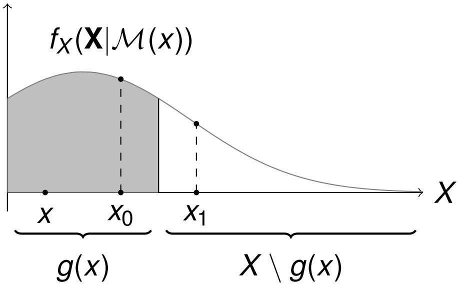
We want to bound the ratio $\frac{f_X(x|\mathcal{M}(x))}{f_X(x'|\mathcal{M}(x))}$ for $x \in X \setminus g(x)$, $x' \in g(x)$.

Intuition



- use d-privacy guarantees to ensure that the attacker would not prefer "correct" guesses in $g(x)$ to "wrong" guesses in $X \setminus g(x)$
- i.e look for a mechanism $\mathcal{M}$ such that $f_X(x_0|\mathcal{M}(x))$ is sufficiently close to $f_X(x_1|\mathcal{M}(x))$ for all $x_0 \in g(x)$, $x_1 \in X \setminus g(x)$.

Intuition



- use d-privacy guarantees to ensure that the attacker would not prefer "correct" guesses in $g(x)$ to "wrong" guesses in $X \setminus g(x)$
- i.e look for a mechanism $\mathcal{M}$ such that $f_X(x_0|\mathcal{M}(x))$ is sufficiently close to $f_X(x_1|\mathcal{M}(x))$ for all $x_0 \in g(x)$, $x_1 \in X \setminus g(x)$.

Goal of our research

- Find a d.p. mechanism that achieves a given bound on guessing advantage
- i.e. from g and η, find $\mathcal{M}$ and ε
  - Perhaps fixing d in the process

Main theorem

- Let f_Y be the probability density function of the distribution of $\mathcal{M}(x)$.
- We have

$$
\begin{aligned}
Pr_{post}(g(x)) &= \frac{1}{1 + \frac{\int_{X \setminus g(x)} f_X(x|\mathcal{M}(x))dx}{\int_{g(x)} f_X(x'|\mathcal{M}(x))dx'}} \\
&= \frac{1}{1 + \frac{\int_{X \setminus g(x)} f_Y(y|x) f_X(x)dx}{\int_{g(x)} f_Y(y|x') f_X(x')dx'}}
\end{aligned}
$$

- The ratio $\frac{f_Y(y|x)}{f_Y(y|x')}$ can be bounded using d-privacy guarantees.

Main theorem

- Let f_Y be the probability density function of the distribution of $\mathcal{M}(x)$.
- We have

$$
\begin{aligned}
Pr_{post}(g(x)) &= \frac{1}{1 + \frac{\int_{X \setminus g(x)} f_X(x|\mathcal{M}(x))dx}{\int_{g(x)} f_X(x'|\mathcal{M}(x))dx'}} \\
&= \frac{1}{1 + \frac{\int_{X \setminus g(x)} f_Y(y|x) f_X(x)dx}{\int_{g(x)} f_Y(y|x') f_X(x')dx'}} \\
&\leq \frac{1}{1 + \int_{X \setminus g(x)} \frac{f_X(x)}{\int_{g(x)} e^{\varepsilon \cdot d(x,x')} f_X(x')dx'} dx}
\end{aligned}
$$

- The ratio $\frac{f_Y(y|x)}{f_Y(y|x')}$ can be bounded using d-privacy guarantees.

Main theorem

- Let f_Y be the probability density function of the distribution of $\mathcal{M}(x)$.
- We have

$$
\begin{aligned}
Pr_{post}(g(x)) &= \frac{1}{1 + \frac{\int_{X \setminus g(x)} f_X(x|\mathcal{M}(x))dx}{\int_{g(x)} f_X(x'|\mathcal{M}(x))dx'}} \\
&= \frac{1}{1 + \frac{\int_{X \setminus g(x)} f_Y(y|x) f_X(x)dx}{\int_{g(x)} f_Y(y|x') f_X(x')dx'}} \\
&\leq \frac{1}{1 + \int_{X \setminus g(x)} \frac{f_X(x)}{\int_{g(x)} e^{\varepsilon \cdot d(x,x')} f_X(x')dx'} dx}
\end{aligned}
$$

This is precise

Cannot derive better bounds from only the d-privacy of $\mathcal{M}$

- The ratio $\frac{f_Y(y|x)}{f_Y(y|x')}$ can be bounded using d-privacy guarantees.

Simplification

$$
\begin{aligned}
Pr_{post}(g(x)) &= \dots \\
&\leq \frac{1}{1 + \int_{X \setminus g(x)} \frac{f_X(x)}{\int_{g(x)} e^{\varepsilon \cdot d(x,x')} f_X(x') dx'} dx} \\
&\leq \frac{1}{1 + e^{-\varepsilon \cdot \sup_{x,x' \in X} d(x,x')} \frac{Pr_{pre}(X \setminus g(x))}{Pr_{pre}(g(x))}} \quad .
\end{aligned}
$$

- Caveat: the quantity $R := \sup_{x,x' \in X} d(x, x')$ does not necessarily exist.

Less of a simplification

- Apply the definition of $\varepsilon \cdot d$-privacy to elements at distance $a \in \mathbb{R}^+$ from $g(x)$:
  - Let $\mathbf{B}(x, r) = \{x' \in X | d(x, x') \leq r\}$ and $\mathbf{A}(x, r) = \{x' \in X | d(x, x') = r\}$
  - Generalize to sets in $\mathbf{B}(\cdot, r)$ and $\mathbf{A}(\cdot, r)$

$$
\begin{aligned}
Pr_{post}(g(x)) &= \ldots = \frac{1}{1 + \frac{\int_{X \setminus g(x)} f_Y(y|x) f_X(x) dx}{\int_{g(x)} f_Y(y|x') f_X(x') dx'}} \\
&= \frac{1}{1 + \frac{\int_{\mathbb{R}^+} \left(\int_{X \setminus g(x) \cap \mathbf{A}(g(x), a)} f_Y(y|x) f_X(x) dx \right) da}{\int_{g(x)} f_Y(y|x') f_X(x') dx'}} \\
&\leq \frac{1}{1 + \frac{\int_{\mathbb{R}^+} e^{-\varepsilon \cdot a} Pr_{pre}(X \setminus g(x) \cap \mathbf{A}(g(x), a)) da}{Pr_{pre}(g(x))}}
\end{aligned}
$$

- Integration over $\mathbb{R}^+$ may be simpler than integration over X

Application to databases

- The attacker wants to guess certain attribute(s) of a certain victim.
  - E.g. what Alice ate and how much salt she used.
- It is easier to assume that the attacker already knows all the oher records except the victim's one:
  - The input space X has only as many dimensions as there are attributes.

Application to databases

- The attacker wants to guess certain attribute(s) of a certain victim.
  - E.g. what Alice ate and how much salt she used.
- It is easier to assume that the attacker already knows all the oher records except the victim's one:
  - The input space X has only as many dimensions as there are attributes.
  - Intuitively, for a stronger attacker, the posterior guessing probability is larger.

Application to databases

- The attacker wants to guess certain attribute(s) of a certain victim.
  - E.g. what Alice ate and how much salt she used.
- It is easier to assume that the attacker already knows all the oher records except the victim's one:
  - The input space X has only as many dimensions as there are attributes.
  - Intuitively, for a stronger attacker, the posterior guessing probability is larger.
  - However, the *advantage* can be larger for a less knowledgeable attacker.
    - The knowledge gain is 0 for someone who already knows everything.

Application to databases

- The attacker wants to guess certain attribute(s) of a certain victim.
  - E.g. what Alice ate and how much salt she used.
- It is easier to assume that the attacker already knows all the oher records except the victim's one:
  - The input space X has only as many dimensions as there are attributes.
  - Intuitively, for a stronger attacker, the posterior guessing probability is larger.
  - However, the *advantage* can be larger for a less knowledgeable attacker.
    - The knowledge gain is 0 for someone who already knows everything.
  - Generalization to weaker attackers is possible assuming that the records are independent.
    - Differential privacy (and d-privacy) mechanisms do not help much (in terms of protecting against attribute guessing) if they are not.

Guessing a single attribute

- Assume the attacker wants to guess the attribute X with precision r.
- We need to define the distance in the space X.
  - Take $d(x, x') := \frac{1}{r}|x - x'|$.
  - We have $g(x) = \{x' : d(x, x') \leq 1\}$.

Guessing a single attribute

- Assume the attacker wants to guess the attribute X with precision r.
- We need to define the distance in the space X.
  - Take $d(x, x') := \frac{1}{r}|x - x'|$.
  - We have $g(x) = \{x' : d(x, x') \leq 1\}$.
- Integration over a can be approximated with a sum over $a \in \mathbb{N}$.

$$
\begin{aligned}
Pr_{post}(g(x)) &= \ldots \leq \frac{1}{1 + \frac{\int_{a \in \mathbb{R}^+} e^{-\varepsilon a} Pr_{pre}(X \setminus g(x) \cap \mathbf{A}(g(x), a)) da}{Pr_{pre}(g(x))}} \\
&\leq \frac{1}{1 + \frac{\sum_{a=0}^{\infty} e^{-\varepsilon a} Pr_{pre}(X \setminus g(x) \cap (\mathbf{B}(g(x), a+1) \setminus \mathbf{B}(g(x), a)))}{Pr_{pre}(g(x))}}
\end{aligned}
$$

Guessing a single attribute

- Assume the attacker wants to guess the attribute X with precision r.
- We need to define the distance in the space X.
  - Take $d(x, x') := \frac{1}{r}|x - x'|$.
  - We have $g(x) = \{x' : d(x, x') \leq 1\}$.
- Integration over a can be approximated with a sum over $a \in \mathbb{N}$.

$$
\begin{aligned}
Pr_{post}(g(x)) &= \ldots \leq \frac{1}{1 + \frac{\int_{a \in \mathbb{R}^+} e^{-\varepsilon a} Pr_{pre}(X \setminus g(x) \cap \mathbf{A}(g(x), a)) da}{Pr_{pre}(g(x))}} \\
&\leq \frac{1}{1 + \frac{\sum_{a=0}^{\infty} e^{-\varepsilon a} Pr_{pre}(X \setminus g(x) \cap (\mathbf{B}(g(x), a+1) \setminus \mathbf{B}(g(x), a)))}{Pr_{pre}(g(x))}} \\
&= \frac{1}{1 + \frac{\sum_{a=1}^{\infty} e^{-\varepsilon a} Pr_{pre}(\mathbf{B}(g(x), a+1) \setminus \mathbf{B}(g(x), a))}{Pr_{pre}(\mathbf{B}(x, 1))}} \quad .
\end{aligned}
$$

Guessing AND of attributes

- Assume the attacker wants to guess:
  - The attribute X_1 with precision r_1;
  - The attribute X_2 with precision r_2;

Guessing AND of attributes

- Assume the attacker wants to guess:
  - The attribute X_1 with precision r_1;
  - The attribute X_2 with precision r_2;
- We need to define the distance in the space $X = X_1 \times X_2$.
  - Take $d(x, x') := \max(\frac{1}{r_1}|x_1 - x'_1|, \frac{1}{r_2}|x_2 - x'_2|)$.
  - We have $g(x) = \{x' : d(x, x') \leq 1\}$.

Guessing AND of attributes

- Assume the attacker wants to guess:
  - The attribute X_1 with precision r_1;
  - The attribute X_2 with precision r_2;
- We need to define the distance in the space $X = X_1 \times X_2$.
  - Take $d(x, x') := \max(\frac{1}{r_1}|x_1 - x_1'|, \frac{1}{r_2}|x_2 - x_2'|)$.
  - We have $g(x) = \{x' : d(x, x') \leq 1\}$.
- We can now treat X similarly to a single attribute, getting

$$Pr_{post}(g(x)) \leq \frac{1}{1 + \frac{\sum_{a=1}^{\infty} e^{-\varepsilon a} Pr_{pre}(\mathbf{B}(x,a+1) \setminus \mathbf{B}(x,a))}{Pr_{pre}(\mathbf{B}(x,1))}}$$

- Compute the probabilities of getting $X \in \mathbf{B}(x, a+1) \setminus \mathbf{B}(x, a)$ for different $a \in \mathbb{N}$.
  - For X_1 and X_2 it can be computed if we know the CDF of the distributions.
  - If X_1 and X_2 are independent, they can be easily combined into probabilities for X.

Where to get such mechanism $\mathcal{M}$?

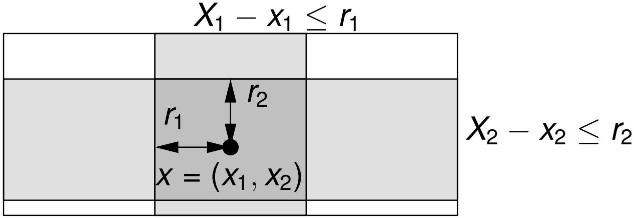
- We fixed a distance *d*
- We want a mechanism $\mathcal{M}$ that
  - is parametrized by ε
  - releases data with ε-d.p. with respect to the distance *d*
- Where to get such $\mathcal{M}$?
  - [Laud, Pankova, Pettai. A Framework of Metrics for Differential Privacy from Local Sensitivity. PET Symposium 2020] is a possible source

Guessing OR of attributes

- Assume the attacker wants to guess either:
  - The attribute X_1 with precision r_1; or
  - The attribute X_2 with precision r_2;

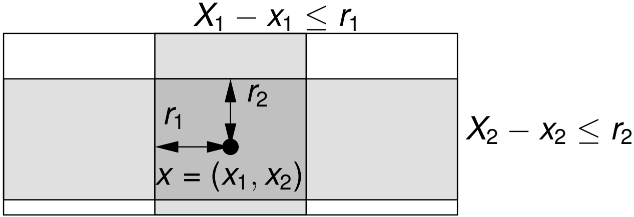
Guessing OR of attributes

- Assume the attacker wants to guess either:
  - The attribute X_1 with precision r_1; or
  - The attribute X_2 with precision r_2;
- We need to define the distance in the space $X = X_1 \times X_2$.
  - The problem is that some elements of $g(x)$ may already be at distance R.



Guessing OR of attributes

- Assume the attacker wants to guess either:
  - The attribute X_1 with precision r_1; or
  - The attribute X_2 with precision r_2;
- We need to define the distance in the space $X = X_1 \times X_2$.
  - The problem is that some elements of $g(x)$ may already be at distance R.



- We can compute a bound that depends on the single attributes.
  - Cannot get a significantly better bound. Only simplified bound for Pr_{post} is usable

$$Pr_{post}(g(x)|k(x)) \leq Pr_{post}(g_1(x)|k(x)) + Pr_{post}(g_2(x)|k(x))$$

Computing ε for a fixed guessing advantage η

- We want: $Pr_{post}(g(x)) - Pr_{pre}(g(x)) \leq \eta$.
  - For simplified bound on Pr_{post}, we can invert the formula, getting

$$\varepsilon \leq \frac{\ln\left(\frac{Pr_{pre}(X \setminus g(x))}{Pr_{pre}(g(x))} \cdot \frac{1}{(Pr_{pre}(g(x))+\eta)^{-1}-1}\right)}{\sup_{x,x' \in X} d(x,x')} \ .$$

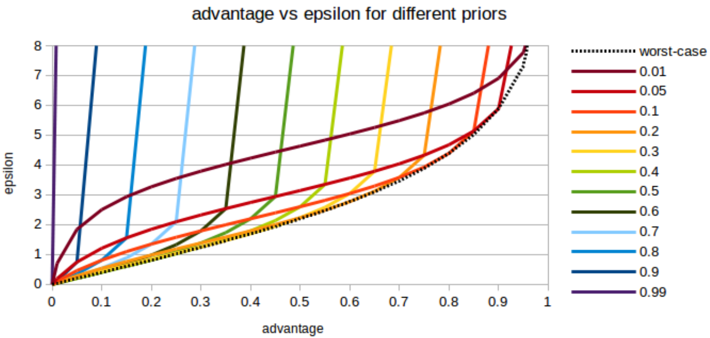
  - For precise bound on Pr_{post}, we can numerically approximate ε using e.g. window binary search over $\varepsilon > 0$.
- Analogously for $Pr_{pre}(g(x)) - Pr_{post}(g(x)) \leq \eta$.

Guessing advantage vs epsilon for different prior distributions

- For the simplified bound

$$\color{red}{Pr_{post}}(g(x)) \leq \frac{1}{1 + \color{magenta}{e^{-\varepsilon \cdot R}} \frac{\color{blue}{Pr_{pre}}(X \setminus g(x))}{\color{blue}{Pr_{pre}}(g(x))}},$$

we can plot the desired bound on advantage vs the largest suitable epsilon for different values of $\color{blue}{Pr_{pre}}(g(x))$.



Worst-case prior distribution

- Using the simplified bound

$$Pr_{post}(g(x)) \leq \frac{1}{1 + e^{-\varepsilon \cdot R \frac{Pr_{pre}(X \setminus g(x))}{Pr_{pre}(g(x))}}},$$

  we can analytically find the value p of $Pr_{pre}(g(x))$ that maximizes the guessing advantage η (if ε is given in advance) or minimizes the ε (if η is given in advance).

$$p = \frac{1-\eta}{2} \text{ for a fixed } \eta \qquad p = \frac{1}{1 + e^{R \cdot \varepsilon / 2}} \text{ for a fixed } \varepsilon.$$

- The precise bound does not provide a better bound if the prior distribution is unknown.

Conclusion — taming the ε

- Differential privacy is a nice composable notion, whose interpretation is unfortunately ambiguous without additional context.
- We can convert ε of differential privacy to more intuitive notions like guessing advantage.