CSE544

Data Privacy

Data Security

- Dorothy Denning, 1982:

  Data Security is the science and study of methods of protecting data (...) from unauthorized disclosure and modification

- Data Security = <u>Confidentiality</u> + <u>Integrity</u>

- Quote from the paper:

  Differential privacy arose in a context in which ensuring privacy is a challenge even if all these control problems are solved: privacy-preserving statistical analysis of data.

Outline

- A famous attack
- Differential privacy (the paper)

Latanya Sweeney's Finding

- In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees
- GIC has to publish the data:

GIC(**zip, dob, sex**, diagnosis, procedure, ...)

This is private ! Right ?

Latanya Sweeney's Finding

- Sweeney paid $20 and bought the voter registration list for Cambridge Massachusetts:

VOTER(name, party, ..., **zip, dob, sex**)

GIC(**zip, dob, sex**, diagnosis, procedure, ...)

This is private ! Right ?

Latanya Sweeney's Finding

zip, dob, sex

- William Weld (former governor) lives in Cambridge, hence is in VOTER
- 6 people in VOTER share his **dob**
- only 3 of them were man (same **sex**)
- Weld was the only one in that **zip**
- Sweeney learned Weld's medical records !

Latanya Sweeney's Finding

- All systems worked as specified, yet an important data has leaked
- How do we protect against that ?

Today's Approaches

- K-anonymity
  - Useful, but not really private
- Differential privacy
  - Private, but not really useful

[Samarati&Sweeney'98, Meyerson&Williams'04]

k-Anonymity

Definition: each tuple is equal to at least k-1 others

Anonymizing: through suppression and generalization

First	Last	Age	Race	Disease
Harry	Stone	34	Afr-am	flue
John	Reyser	36	Cauc	mumps
Beatrice	Stone	47	Afr-am	mumps
John	Ramos	22	Hisp	allergy

Hard: NP-complete for supression only
Approximations exists

[Samarati&Sweeney'98, Meyerson&Williams'04]

k-Anonymity

Definition: each tuple is equal to at least k-1 others

Anonymizing: through suppression and generalization

First	Last	Age	Race	Disease
*	Stone	30-50	Afr-am	flue
John	R*	20-40	*	mumps
*	Stone	30-50	Afr-am	mumps
John	R*	20-40	*	allergy

Hard: NP-complete for supression only
Approximations exists

[Samarati&Sweeney'98, Meyerson&Williams'04]

k-Anonymity

Better: remove identifying attributes, keep only "quasi-identifiers":

Quasi identifiers (anonymized)

Sensitive attribute

Age	Race	Disease
30-50	Afr-am	flue
20-40	*	mumps
30-50	Afr-am	mumps
20-40	*	allergy

[Samarati&Sweeney'98, Meyerson&Williams'04]

k-Anonymity

BUT: Does not provide protection!

$$\underbrace{\text{Age} \quad \text{Race}}_{\text{Quasi identifiers (anonymized)}} \quad \underbrace{\text{Disease}}_{\text{Sensitive attribute}}$$

Age	**Race**	**Disease**
30-50	Afr-am	flue
20-40	*	mumps
30-50	Afr-am	mumps
20-40	*	mumps

Here we learn immediately that John Ramos, 22, has mumps (how?)

Data Privacy Ideal

Allow queries like this:

```
SELECT count(*)
FROM Patients
WHERE age > 24 and disease = ‘mumps’
```

Disallow queries like this:

```
SELECT disesase
FROM Patients
WHERE age = 22
```

“How Is Hard”

From the paper:

- What about designing a system that allows *only* count(*) queries? Will it be private?

"How Is Hard"

From the paper:

- What about designing a system that allows *only* count(*) queries? Will it be private?

- No!
  - "How many people in the database have the sickle cell trait?"
  - "How many people in the database not named 'John Ramos' have the sickle cell trait?"
- Query auditing is *not* the solution (why?)

Adding Random Noise

Answer a query like:

```
SELECT count(*)
FROM Patients
WHERE age > 24 and disease = 'mumps'
```

By adding a random noise.

This fixes the previous problem (why?).

But creates a new problem: query repeatedly, average, remove noise.

More sophisticated attach in the paper: Theorem 1, due to Dinur Nissim.

Differential Privacy

[Dwork]

DEFINITION 1. *A randomized function* $\mathcal{K}$ *gives* ε-differential privacy *if for all data-sets D and D′ differing on at most one row, and all* $S \subseteq Range(\mathcal{K})$,

$$\Pr[\mathcal{K}(D) \in S] \leq \exp(\varepsilon) \times \Pr[\mathcal{K}(D') \in S], \quad (1)$$

where the probability space in each case is over the coin flips of $\mathcal{K}$.

Differential Privacy

[Dwork]

$$\Pr[\mathcal{K}(D) \in S] \leq \exp(\varepsilon) \times \Pr[\mathcal{K}(D') \in S], \quad (1)$$

What privacy do the following values for ε ensure to an end user?

- 0
- 0.01
- 0.1
- 1
- 10

Differential Privacy

[Dwork]

$$\Pr[\mathcal{K}(D) \in S] \leq \exp(\varepsilon) \times \Pr[\mathcal{K}(D') \in S], \quad (1)$$

What privacy do the following values for ε ensure to an end user?

- 0 = total privacy: algorithm returns *same* answer on all databases
- 0.01 = the two probabilities differ by < 1%
- 0.1 = the two probabilities differ by < 10%
- 1 = the two probabilities differ by < e ≈ 2.71
- 10 = certainly not good…

Recall your math: if |ε| is small, then exp(ε) ≈ 1+ ε

[Dwork]

Achieving Differential Privacy

DEFINITION 2. *For* $f : \mathcal{D} \rightarrow \mathbf{R}^d$, *the* L_1 sensitivity *of* f *is*[7]

$$\Delta f = \max_{D,D'} \| f(D) - f(D') \|_1$$
$$= \max_{D,D'} \sum_{i=1}^{d} | f(D)_i - f(D')_i | \qquad (3)$$

for all D, D' *differing in at most one row.*

[Dwork]

Achieving Differential Privacy

Examples. What is the sensitivity of these queries?

```
SELECT count(*)
FROM Patients
WHERE disease = 'mumps'
```

```
SELECT disease, count(*)
FROM Patients
GROUP By disease
```

```
SELECT avg(age)
FROM Patients
WHERE disease = 'mumps'
```

100 queries of the form:

```
SELECT count(*)
FROM Patients
WHERE [some condition]
```

[Dwork]

Achieving Differential Privacy

Examples. What is the sensitivity of these queries?

```
SELECT count(*)
FROM Patients
WHERE disease = 'mumps'
```

Δf = 1

```
SELECT disease, count(*)
FROM Patients
GROUP By disease
```

Δf = 1

```
SELECT avg(age)
FROM Patients
WHERE disease = 'mumps'
```

Δf = can be high (say, 20 or 30)

```
100 queries of the form:
SELECT count(*)
FROM Patients
WHERE [some condition]
```

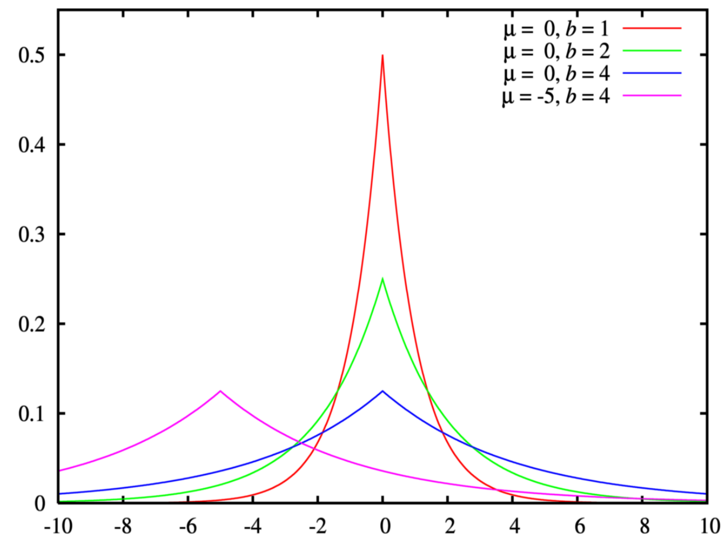
Δf = 100

Note: the number of queries dictates your ***privacy budget***

[Dwork]

Achieving Differential Privacy

Laplacian distribution



Lap(b) with mean μ=0 has the following pdf:

$$P(z \mid b) = \frac{1}{2b} \exp(-|z|/b)$$

Variance = $2b^2$

THEOREM 2. *For $f : \mathcal{D} \rightarrow \mathbf{R}^d$, the mechanism $\mathcal{K}$ that adds independently generated noise with distribution Lap $(\Delta f/\varepsilon)$ to each of the d output terms enjoys ε-differential privacy.*[7]

[Dwork]

Achieving Differential Privacy

Laplacian distribution

Suppose Δf=1 and ε=0.1

How much noise do we add?
(What is a "typical" noise value?)

Lap(b) with mean μ=0 has the following pdf:

$$P(z \mid b) = \frac{1}{2b} \exp(-|z|/b)$$

Variance = $2b^2$

THEOREM 2. *For* $f : \mathcal{D} \to \mathbf{R}^d$, *the mechanism* $\mathcal{K}$ *that adds independently generated noise with distribution Lap* $(\Delta f/\varepsilon)$ *to each of the d output terms enjoys* ε*-differential privacy.*[7]

[Dwork]

Achieving Differential Privacy

Laplacian distribution

Suppose $\Delta f=1$ and $\varepsilon=0.1$

How much noise do we add?
(What is a "typical" noise value?)

Lap(b) with mean $\mu=0$ has the following pdf:

$$P(z|b) = \frac{1}{2b}\exp(-|z|/b)$$

b = $\Delta f / \varepsilon = 10$. "Typical" noise is $b\sqrt{2} \approx 14$.
Let's compute the probability of noise > b:
$2*\int_b^\infty P(z|b)\, dz =$
$= 2*1/(2b)*\int_b^\infty \exp(-z/b)dz =$
$= \exp(-1) = 0.36$

Variance = $2b^2$

THEOREM 2. *For $f : \mathcal{D} \to \mathbf{R}^d$, the mechanism $\mathcal{K}$ that adds independently generated noise with distribution Lap $(\Delta f/\varepsilon)$ to each of the d output terms enjoys ε-differential privacy.*[7]

Is this this answer useful?

[Dwork]

Achieving Differential Privacy

Laplacian distribution

Suppose $\Delta f=1$ and $\varepsilon=0.1$

How much noise do we add?
(What is a "typical" noise value?)

Lap(b) with mean $\mu=0$ has the following pdf:

$$P(z \mid b) = \frac{1}{2b} \exp(-|z|/b)$$

$b = \Delta f / \varepsilon = 10$. "Typical" noise is $b\sqrt{2} \approx 14$.
Let's compute the probability of noise > b:
$2 * \int_b^\infty P(z|b)\, dz =$
$= 2 * 1/(2b) * \int_b^\infty \exp(-z/b) dz =$
$= \exp(-1) = 0.36$

Variance = $2b^2$

Theorem 2. *For $f : \mathcal{D} \rightarrow \mathbf{R}^d$, the mechanism $\mathcal{K}$ that adds independently generated noise with distribution Lap ($\Delta f/\varepsilon$) to each of the d output terms enjoys ε-differential privacy.*[7]

Is this this answer useful?

Yes = if the real answer is >> 10
No = if the real answer is << 10

Limitations of Differential Privacy

- *Privacy budget* ≈ the maximum number of queries that one can ask
  - Once a user exhaust her privacy budget, the system should (theoretically) refuse to answer any new query, forever! (or unitl the database gets updated significantly)
- Protects only individual users, but not general secrets
  - “Hide the fact that our hospital has significantly reduce the number of mumps cases over the last year”

Final Comments on Privacy

- In the database literature, privacy is equated with *confidentiality*
- In real life, privacy is more complex:
  - "Is the right of individuals to determine for themselves when, how and to what extent information about them is communicated to others" [Agrawal'03]

The End of CSE 544

What you achieved in 10 weeks:

1. Relational data and query model
2. Database systems
3. Database theory
4. Miscellaneous: transactions, provenance, privacy

Three homeworks, one project, nine reading assignments

- You still need to finish the project, turn in HW4