



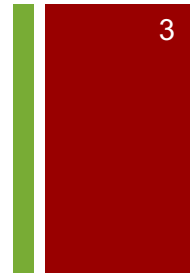
Tutorial 8: Data Privacy

+ Q1

- What is CIA Triad of information security?
- What is data privacy according to The General Data Protection Regulation (GDPR)?

+ Q1 CIA Triad of information security

- Confidentiality: Ensures that data or an information system is accessed by only an authorized person.
- Integrity: Integrity assures that the data or information system can be trusted. Ensures that it is edited by only authorized persons and remains in its original state when at rest.
- Availability: Data and information systems are available when required.



+ Q1 Data Privacy

4

- Issues related to **appropriate use of information**
- Goes beyond security
 - Not only your data, but also about you as a person
 - Personal
 - Name, identity card number, passport number, social security number, birthday, diagnostic health information, GPS position, IP address, behavioral profile, ethnic origin, religious beliefs, location derived from telecommunication systems...
- How your data could be treated and who could access to it
 - Privacy policy (like when installing an app)
 - How your data is going to be collected, saved, or transferred, and even if it will be transferred to third parties
- From an end user perspective
 - Preventing storage of personal information
 - Ensuring appropriate use of personal information
- Data Publishing
 - K- Anonymity, L-Diversity, T-closeness, Differential Privacy



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA


+ Q2 Quasi-identifiers

5

- What is a quasi-identifier?
 - A piece of information that alone is not an identifier, but when combined with other quasi-identifiers it can create a unique identifier.
 - Gender
 - Birth date
 - Street Name
- Can a dataset have no quasi-identifier, or multiple quasi-identifiers?
 - Yes, in fact, the issue that arises is datasets having too many quasi-identifiers that make it very easy to determine the entity that a record belongs to

+ Q2 Quasi-identifiers

- What is the role a quasi-identifier plays in the k-anonymity approach?
 - k-anonymity aims to conceal these quasi-identifiers in groups to prevent statistical attacks
 - If we hide the quasi-identifier of Age, it will no longer be clear what Bob suffers (Bob is 53 years old)

Age	Medial Issue		Age	Medial Issue
32	Broken arm		[30,60]	Broken arm
48	Headaches		[30,60]	Headaches
53	Broken toe		[30,60]	Broken toe

- By hiding the age in a group of size k (in this example k=3), no longer can we tell which row refers to Bob (more detailed examples coming)

+ Q3 Data Publishing

7

- Considering the statistical attack example in Lecture Notes
- How can you protect data privacy (i.e., not revealing individual's salary) using k-anonymity, l-diversity and differential privacy?
- Please also discuss their potential problems

+ An Example of Statistical Attack

8

- Privacy rules
 - Cannot query about individual's salary
- Attack queries:

```
select count(*)  
from staff  
where title = "Professor"
```

```
select sum(salary)  
from staff  
where title = "Professor"
```


+ Q3

9

- Easiest approach

```
select count(*)  
from staff  
where title = "Professor" and Name = "Andy"
```

- If the result is 1

```
Select sum(salary)  
from staff  
where title = "Professor" and Name = "Andy"
```

ID	Name	Title	Age	Salary
1	Andy	Professor	43	18.4K
2	Bob	Professor	35	16.3K
3	Claire	Professor	47	18.4K
4	Doug	Professor	34	16.3K
5	Emma	Lecturer	33	13.5K
6	Fabio	Lecturer	31	11.7K
7	George	Lecturer	30	11.6K

+ Q3

- Remove ID and Name

```
select count(*)  
from staff  
where title = "Professor" and Age = 43
```

- If the result is 1

```
Select sum(salary)  
from staff  
where title = "Professor" and Age = 43
```

ID	Name	Title	Age	Salary
1	Andy	Professor	43	18.4K
2	Bob	Professor	35	16.3K
3	Claire	Professor	47	18.4K
4	Doug	Professor	34	16.3K
5	Emma	Lecturer	33	13.5K
6	Fabio	Lecturer	31	11.7K
7	George	Lecturer	30	11.6K

+ Q3 K-Anonymity

11

- Each combination of quasi-identifiers (QI) is hidden in a group of size at least k
 - Title and Age
 - Requires that each (Title, Age) combination can be matched to at least k salaries

Select sum(*)
from staff
where title = "Professor" and Age = 43

select count(salary)
from staff
where title = "Professor" and Age = 43

ID	Name	Title	Age	Salary
1	Andy	Professor	43	18.4K
2	Bob	Professor	35	16.3K
3	Claire	Professor	47	18.4K
4	Doug	Professor	34	16.3K
5	Emma	Lecturer	33	13.5K
6	Fabio	Lecturer	31	11.7K
7	George	Lecturer	30	11.6K



Title	Age	Salary
Professor	[40,50]	18.4K
Professor	[30,40]	16.3K
Professor	[40,50]	18.4K
Professor	[30,40]	16.3K
Lecturer	[30,40]	13.5K
Lecturer	[30,40]	11.7K
Lecturer	[30,40]	11.6K



+ Q3 K-Anonymity

select count(*)
from staff
where title = "Professor" and Age IN [40,50]

Select sum(salary)
from staff
where title = "Professor" and Age IN [40,50]

Select max(salary)
from staff
where title = "Professor" and Age IN [40,50]

Select min(salary)
from staff
where title = "Professor" and Age IN [40,50]

	Title	Age	Salary
2-Anonymity	Professor	[40,50]	18.4K
	Professor	[40,50]	18.4K
2-Anonymity	Professor	[30,40]	16.3K
	Professor	[30,40]	16.3K
3-Anonymity	Lecturer	[30,40]	13.5K
	Lecturer	[30,40]	11.7K
	Lecturer	[30,40]	11.6K

+ Q3 K-Anonymity

13

```
select count(*)  
from staff  
where title = "Professor" and Age IN [40,50]
```

```
Select sum(salary)  
from staff  
where title = "Professor" and Age IN [40,50]
```

```
Select max(salary)  
from staff  
where title = "Professor" and Age IN [40,50]
```

```
Select min(salary)  
from staff  
where title = "Professor" and Age IN [40,50]
```

Min = Max = Sum / 2 !

- Hiding in a group of k is not sufficient
- The group should have a diverse set of sensitive values

	Title	Age	Salary
2-Anonymity	Professor	[40,50]	18.4K
	Professor	[40,50]	18.4K
2-Anonymity	Professor	[30,40]	16.3K
	Professor	[30,40]	16.3K
3-Anonymity	Lecturer	[30,40]	13.5K
	Lecturer	[30,40]	11.7K
	Lecturer	[30,40]	11.6K

+ Q3 L-Diversity

14

select count(*)
from staff
where title = "Lecturer" and Age IN [30,40]

Select sum(salary)
from staff
where title = "Lecturer" and Age IN [30,40]

Select max(salary)
from staff
where title = "Lecturer" and Age IN [30,40]

Select min(salary)
from staff
where title = "Lecturer" and Age IN [30,40]

Select med(salary)
from staff
where title = "Lecturer" and Age IN [30,40]

4-Anonymity

2-Diversity

3-Anonymity

3-Diversity

Title	Age	Salary
Professor	[30,50]	18.4K
Professor	[30,50]	18.4K
Professor	[30,50]	16.3K
Professor	[30,50]	16.3K
Lecturer	[30,40]	13.5K
Lecturer	[30,40]	11.7K
Lecturer	[30,40]	11.6K

+ Q3 L-Diversity

15

select count(*)
from staff
where title = "Lecturer" and Age IN [30,40]

Select sum(salary)
from staff
where title = "Lecturer" and Age IN [30,40]

Select max(salary)
from staff
where title = "Lecturer" and Age IN [30,40]

Select min(salary)
from staff
where title = "Lecturer" and Age IN [30,40]

Select med(salary)
from staff
where title = "Lecturer" and Age IN [30,40]

But if the attacker knows
George's salary is low...

4-Anonymity

2-Diversity

3-Anonymity

3-Diversity

Title	Age	Salary
Professor	[30,50]	18.4K
Professor	[30,50]	18.4K
Professor	[30,50]	16.3K
Professor	[30,50]	16.3K
Lecturer	[30,40]	13.5K
Lecturer	[30,40]	11.7K
Lecturer	[30,40]	11.6K

+ Q3 L-Diversity

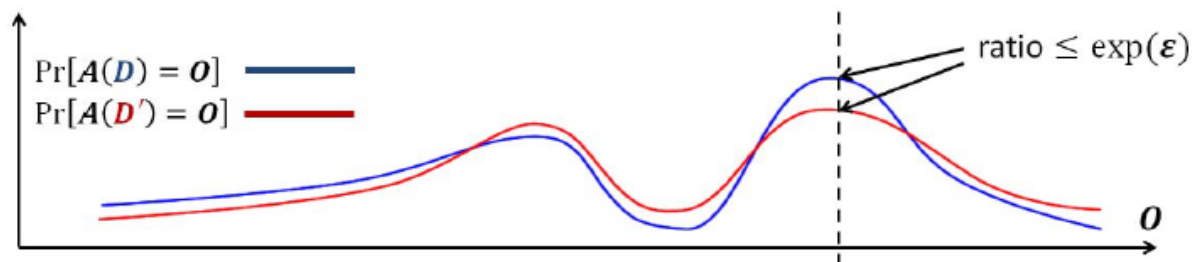
- What if there are more attributes?
- What if the attacker has more background knowledge?
- What if it is harder to generalize data?
 - Hard to win this war...

	Title	Age	Salary
4-Anonymity	Professor	[30,50]	18.4K
	Professor	[30,50]	18.4K
2-Diversity	Professor	[30,50]	16.3K
	Professor	[30,50]	16.3K
3-Anonymity 3-Diversity	Lecturer	[30,40]	13.5K
	Lecturer	[30,40]	11.7K
	Lecturer	[30,40]	11.6K

+ Q5 Differential Privacy

17

■ Illustration of ϵ -differential privacy



where D and D' are neighboring databases that differ by **at most one** tuple

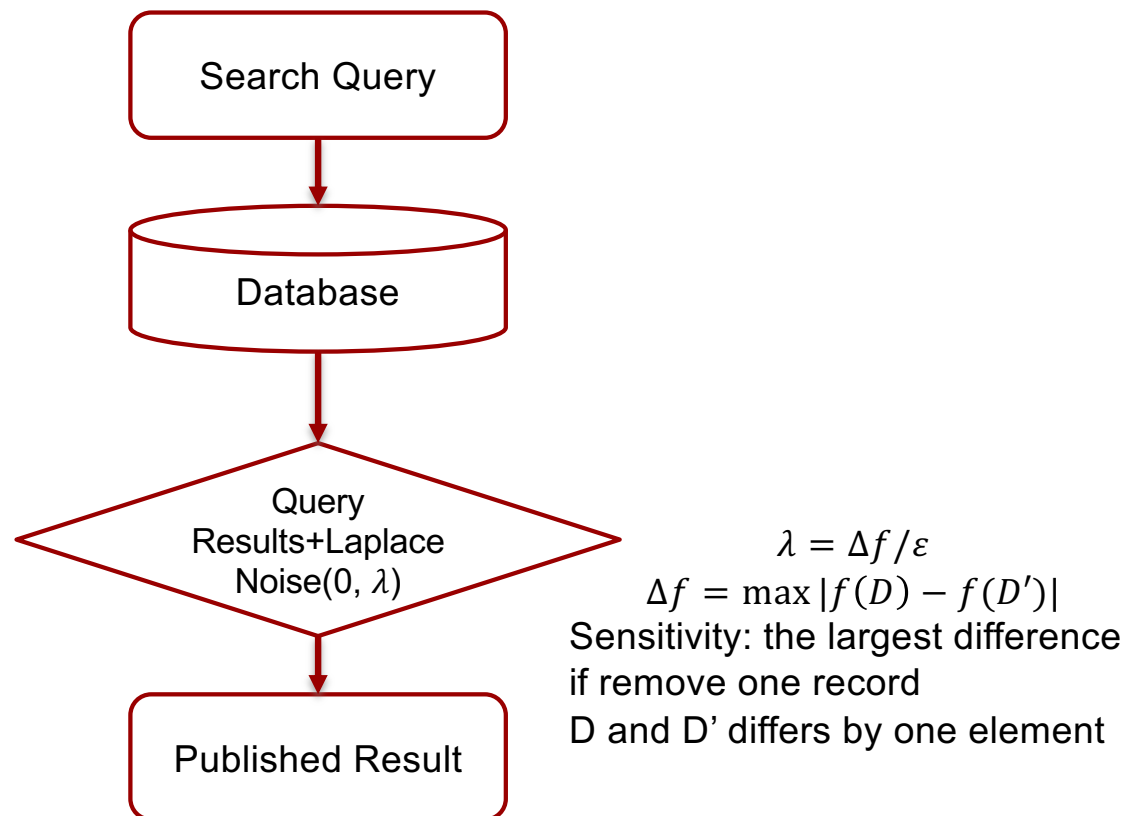
$$\exp(-\epsilon) \leq \frac{\Pr[A(D) = O]}{\Pr[A(D') = O]} \leq \exp(\epsilon)$$



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

+ Q3 Differential Privacy

18



The noise only depends on Δf and ϵ
It has nothing to do with database !

+ Q3 Differential Privacy

■ Add random noise using Laplace Distribution

■ Count

- Sensitivity $\Delta f = 1$,
- Query result: $3 + \eta$
 - η is drawn from $\text{Lap}(1/\epsilon)$
 - Sensitivity of Count is $\Delta f = 1$
 - Mean = 0
 - Variance = $2/\epsilon^2$

```
select count(*)
from staff
where title = "Lecturer" and Age IN [30,40]
```

variance: $2\lambda^2$

■ Sum

- Sensitivity $\Delta f = \text{Max}(\text{Salary}) = 13.5$
- Query result: $\text{sum} + \eta$
- η is drawn from $\text{Lap}(13.5/\epsilon)$
 - Mean = 0
 - Variance = $2 \times 13.5^2 / \epsilon^2$

Title	Age	Salary
Professor	[30,50]	18.4K
Professor	[30,50]	18.4K
Professor	[30,50]	16.3K
Professor	[30,50]	16.3K
Lecturer	[30,40]	13.5k
Lecturer	[30,40]	11.7K
Lecturer	[30,40]	11.6K



+ Q3 Differential Privacy

20

- Still open problems out there...
 - Might be too strong
 - It requires that changing one tuple should not bring much change to the published result
 - How to choose an appropriate ϵ ?
 - How to quantify the cost of privacy and the gain of utility in releasing data?