

*CSIT115 Data Management and Security*  
*CSIT882 Data Management Systems*

# Data Privacy

Subject Coordinators: Dr Chen Chen, Dr Thanh Le

School of Computing and Information Technology -  
University of Wollongong

# Data privacy

## Outline

### Concepts

#### Privacy in statistical databases

#### Privacy protection

#### Sample applications

# Concepts

In a social context, **trust** has several connotations; definitions of **trust** typically refer to a situation characterized by the following aspects: one party (**trustor**) is willing to rely on the actions of another party (**trustee**)

In systems, a **trusted component** has a set of properties which another component can rely on. If A **trusts** B, this means that a violation in those properties of B might **compromise** the correct operation of A

**Trustworthiness** means that something or someone is able to be relied on to do or provide what is needed: **deserving of trust**

The term **trustworthy computing** has been applied to computing systems that are inherently **secure, available** and **reliable**

**Privacy of information** allows a user to query a database while hiding identities of data items retrieved

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

3/27

# Concepts

**Private information retrieval** is like buying in a store without a seller knowing who buys what

It means, that a seller is unable to associate the purchased items with a particular person

**Private information retrieval** has many applications in medical databases, personal databases, electronic commerce databases, web searches, etc.

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

4/27

# Data privacy

## Outline

Concepts

Privacy in statistical databases

Privacy protection

Sample applications

# Privacy in statistical databases

A **simple statistical database** can be viewed as a table containing personal records, where the rows correspond to individuals and the columns correspond to different attributes, for example, a medical database may contain attributes such as **name**, **social security number**, **address**, **age**, **gender**, **ethnicity**, and **medical history** for each patient

We would like the medical researchers to have some form of access to this database to learn trends such as correlation between age and heart disease, while maintaining individual privacy

Information retrieved from **statistical databases** comes from **statistical (aggregate queries)** on a column in a relational table with an **aggregate function**

**Aggregate functions** include **COUNT**, **SUM**, **AVG**, **MAX**, **MIN**

[In HTML view press 'p' to see the lecture notes](#)

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

6/27

# Privacy in statistical databases

## Privacy problems:

- A database contains data, that is individually sensitive and because of that a direct access to data is not permitted
- Statistical queries are permitted and statistical queries access individual data items
- In such a situation, it is possible to infer information that violates privacy constraints

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

7/27

# Privacy in statistical databases

The following relational table contains information about the **names of students, sex, degree enrolled, total number of units passed, and average grade**

Relational schema

```
STUDENT(name, sex, degree, units, ave_grade)
```

A sample statistical query finds **an average grade of all students enrolled in BCompSci degree**

SELECT statement

```
SELECT AVG(ave_grade)
FROM STUDENT
WHERE degree = 'BCompSci';
```

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

8/27

# Privacy in statistical databases

Due to privacy reasons individual entries in `units` and `avg_grade` cannot be read directly

**Aggregation** refers to observation that the **sensitivity level** of an aggregated computed over a group of values is different from **sensitivity levels** of individual elements

In the majority of scenarios sensitivity level of a **result of aggregation** is lower than sensitivity level of individual elements, e.g. `AVG(salary)`

However, it is possible that a **result of aggregation** is more sensitive than individual elements, e.g. `SUM(expenses)`

**Inference problem** refers to derivation of sensitive information from non-sensitive data

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

9/27

# Privacy in statistical databases

## The types of attack

- **Direct attack** when aggregate is computed over a small sample so that information about individual data items is leaked
- **Indirect attack** which combines information related to several aggregates
- **Tracker attack**, which allows to track down information about single tuple
- **Linear system vulnerability**, which uses algebraic relations between query sets to construct equations, which can be solved to reveal information about the individual items

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

10/27

# Privacy in statistical databases

For example assume that **Carol** is a female BCompSci student

The following queries

```
SELECT COUNT(*)  
FROM STUDENTS  
WHERE sex = 'F' AND degree = 'BCompSci';
```

SELECT statement

```
SELECT AVG(avg_grade)  
FROM STUDENTS  
WHERE sex = 'F' AND degree = 'BCompSci';
```

SELECT statement

provide precise information about individual average if only one female student is included a relational table **STUDENTS**

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

11/27

# Privacy in statistical databases

A query condition that allows to track down information about a single row in a relational table is called as an **individual tracker**(**quasi identifier**) for such row

A **general tracker** is a predicate that can be used to find an answer to any **inadmissible query**

In another example assume that a condition `name= 'Carol'` uniquely identifies a tuple in `STUDENTS` table

[In HTML view press 'p' to see the lecture notes](#)

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

12/27

# Privacy in statistical databases

Then if the queries

```
SELECT SUM(units)
FROM STUDENTS
WHERE name='Carol' OR degree = 'MIS';
```

SELECT statement

```
SELECT SUM(units)
FROM STUDENTS
WHERE name = 'Carol' OR NOT (degree = 'MIS');
```

SELECT statement

```
SELECT SUM(units)
FROM STUDENTS;
```

SELECT statement

return the values 75, 77, and 136 then  $(75 + 77) - 136 = 16$  which is the total number of units passed by Carol

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

13/27

# Privacy in statistical databases

In yet another example, suppose that we compute  $\text{sum}(x_1, x_2, x_3)$  and we get a result 15

Next, suppose that we submit a query  $\text{max}(x_1, x_2, x_3)$  and the system denies the answer

The denial tells us that if the true answer to the second query were given then some value could be uniquely determined and it allows us to find the values of  $x_1, x_2, x_3$

Note that  $\text{max}(x_1, x_2, x_3) \geq 5$  because otherwise the sum could not be 15

Further, if  $\text{max}(x_1, x_2, x_3) > 5$  then the query would not have been denied since no value could be uniquely determined

So the only case when it is reasonable to deny answer is the case when  $\text{max}(x_1, x_2, x_3) = 5$  and we learn that  $x_1 = x_2 = x_3 = 5$

It is a privacy breach of all three entries.

[In HTML view](#) press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

14/27

# Data privacy

## Outline

Concepts

Privacy in statistical databases

Privacy protection

Sample applications

# Privacy protection

**Query auditing** denies one or more queries from a sequence of queries

The queries to be denied are chosen such that the sensitivity of underlying data is preserved

For a given a sequence of queries that have already been posed about the data, their corresponding answers, and for a new query, the system denies the answer if privacy can be breached or give the true answer otherwise

There are two versions of **query auditing**:

- **online** in which we do not know a sequence of queries in advance and
- **offline** in which we know entire sequence of queries in advance

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

16/27

# Privacy protection

For example a query like

```
SELECT COUNT(*)  
FROM STUDENTS  
WHERE sex = 'F' AND degree = 'BCompSci';
```

SELECT statement

is dismissed when individual tracker (quasi identifier) is used in **WHERE** clause

A problem is that almost all statistical databases have a **general tracker**, i.e. it is possible to hide an **individual tracker** in a **complex condition**

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

17/27

# Privacy protection

In the **randomization method (output perturbation)** privacy is obtained by perturbing the true answer to a database query by the addition of a small amount of random noise.

The **randomization family** includes swapping values between records, replacing the original database by a sample from the same distribution , adding noise to the values in the database, adding noise to the results of a query, and sampling the result of a query

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

18/27

# Privacy protection

For example, **randomization method** is used in a context of distorting data by probability distribution for surveys which have evasive answer bias due privacy concerns

Assume that we have a relational table R with a set of rows  $\{r_1, r_2, \dots, r_n\}$

For an attribute a in a row  $r_i$  we add a noise component which is taken from the probability distribution function  $f(y)$

The noise components are denoted by  $y_1, y_2, \dots, y_n$

A new set of distorted rows are denoted by  $\{r_1.a+y_1, r_2.a+y_2, \dots, r_n.a+y_n\}$

The added noise is so significant that it is impossible to guess the original values

[In HTML view press 'p' to see the lecture notes](#)

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

19/27

# Privacy protection

**Randomization method** is simple and it can be implemented at data collection time

Its weakness is that outlier values (values significantly different from the majority of values) are more susceptible to attack than value in dense regions

For example, we use the **randomization method** to enforce **privacy** and we assume that  $r1.a=2$ ,  $r2.a=1$ ,  $r3.a=3$ , and  $r4.a=2$

Next assume, that the following randomization record is applied  $[1, 4, 1, 2]$ , such that  $sum([1, 4, 1, 2])=8$

The values after randomization  $r1.a=3$ ,  $r2.a=5$ ,  $r3.a=4$ , and  $r4.a=4$  are revealed to a use together with 8 (summation over randomization record)

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

20/27

# Privacy protection

Then a user is able to find a correct value of  $\text{avg}(r1.a, r2.a, r3.a, r4.a)$  through computation of  $((r1.a + r2.a + r3.a + r4.a) - 8) / 4 = 2$  without knowing the original values of  $r1.a, r2.a, r3.a, r4.a$

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

21/27

# Privacy protection

The most common method of randomization is through additive perturbations

It is also possible to use **data swapping** in order to preserve privacy

Certain kinds of aggregate computations can be exactly performed without losing privacy

It cannot be implemented at data collection time

Randomization has two important weaknesses:

- Outlier records are difficult to mask
- Publicly available records can be used to identify the owners

[In HTML view press 'p' to see the lecture notes](#)

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

22/27

# Privacy protection

**Group based anonymization** constructs groups of records which are transformed in group-specific way

In many applications rows in a relational table are available by removing key identifiers

Other kinds of attributes (pseudo-identifiers) can be used to accurately identify the rows in relational table, e.g. age, zip code, sex in census rolls

**K-anonymity techniques** reduce granularity of representation of pseudo-identifiers with **generalization and suppression**

In **k-anonymity approach** each relational table must be such that every combination of pseudo-identifiers can be indistinguishably matched to at least **k** individuals

In **generalization** the values of attributes are generalized to a range in order to reduce the granularity of representation, for example a **date of birth** can be generalized to a range such as **year of birth** to reduce a risk of identification

[In HTML view press 'p' to see the lecture notes](#)

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

23/27

# Privacy protection

In **value suppression** a value of an attribute is removed completely

In HTML view press 'p' to see the lecture notes

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

24/27

# Data privacy

## Outline

Concepts

Privacy in statistical databases

Privacy protection

Sample applications

# Sample applications

The problem of **privacy preservation** has many applications in homeland security, medical databases and customer transactions databases

The **Scrub** system was designed for de-identification of clinical notes and letters which typically occurs in the forms of textual data

The system removes obvious references to patients, family members, addresses, phone numbers as well as cryptic references in a form of abbreviations that can be understood by the specialist

The system is able to remove more than 99% of identifying information from data

The **Datafly** system removes identification of subjects from medical records stored in relational tables

It includes directly identifying information like for example social security number and non-directly identifying information such as age, sex, or zip code, similar to k-anonymity approach

[In HTML view press 'p' to see the lecture notes](#)

[TOP](#)

Created by Janusz R. Getta, CSIT115 Data Management and Security, Autumn 2023

26/27

# References

C. C. Aggrawal, P. S. Yu Privacy-Preserving Data Mining Models and Algorithms, Chapter 2 General Survey of privacy-Preserving Data mining Models and Algorithms, Springer 2009