Computer Security (EDA263 / DIT 641)

Lecture 12:
Database Security

Erland Jonsson
Department of Computer Science and Engineering
Chalmers University of Technology
Sweden

Outline

- Introduction to databases
- Database security requirements
- Sensitive data
- Inference
  - basics
  - in statistical databases (SDBs)

What is a database?

- **Database** = collection of data + set of rules that specify certain relationships among the data.
- Data is stored in one or more files
- The database file consists of **records**, which in turn consists of **fields or elements.**
- The logical structure of the database is called a **schema.**
- A **subschema** is that part of the database, to which a particular user may have access.
- Data can be organised in tables. All columns are given names, which are the **attributes** of the database.
- A **relation** is a set of columns

What is a database? (2)

- **Database management system (DBMS)** (databashanterare) is a program with which the user interacts with the data base
- **Database administrator** is a person that defines the rules that organise the data and who should have access to which parts of the data. (expresses an access policy)
- Several databases could be joined ("samköra")
- Users interact with the database through commands to the DBMS. A command is called a **query.**
- Security requirements (in general):
  - Confidentiality, Integrity, Availability (!)

What makes database security a problem?

- the sensitivity for the"same type" of elements may differ
- differentiated sensitivity may be necessary (>2)
- **inference** (Sw. slutledning), i.e. "unwanted" conclusions can be drawn
  - the sensitivity of a combination of data differs from the sensitivity of the data elements
  - data are semantically related

Database security requirements

- Physical database integrity - power failures etc
- Logical database integrity - the structure is preserved
- Element integrity - data must be accurate
- Auditability - possibility to track changes
- Access control
- User authentication
- Availability
- Confidentiality - protection of sensitive data

Integrity of the database

- **Overall Goal:** data must always be correct
- Mechanisms for the whole database:
- DBMS must regularly **back up** all files
- DBMS must maintain a **transaction log**

Reliability and integrity mechanisms

- **record locking** (write):
  - we want atomic and serialisable operations:
  - *atomic*: (cp "read-modify-write" for instructions)
  means that operations can not be interrupted
  => either OK and data correctly updated or
  NOT OK and data unchanged
  - *serialisable*:
  the result of a number of transactions that are started at the same time must be the same as if they were made in strict order

Sensitive data

There are several reasons why data are sensitive:

- **inherently sensitive** (location of missiles)
- **from a sensitive source** (an informer's identity may be compromised)
- **declared sensitive** (military classification, anonymous donour)
- **part of a sensitive record**/attribute
- sensitive **in relation to previously disclosed information** (longitude plus latitude)

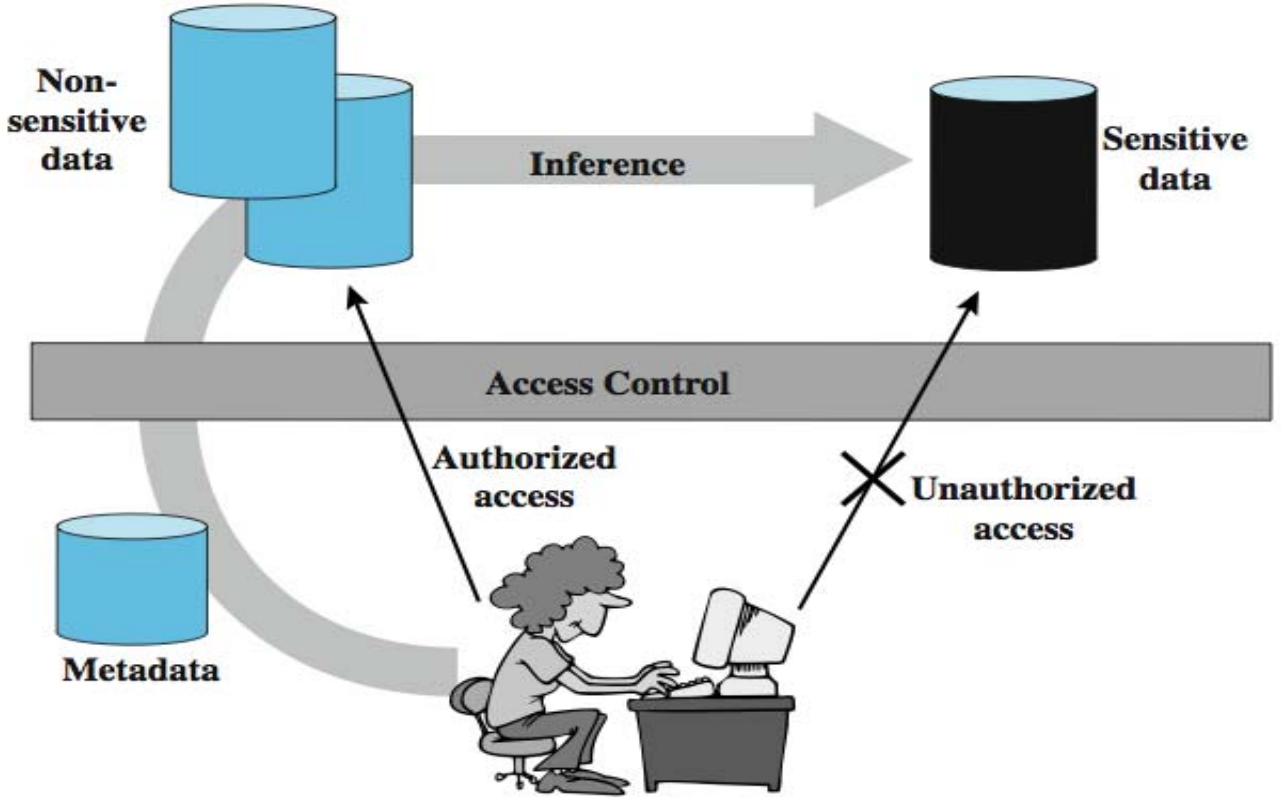
Sensitive data – types of disclosures

There are various ***forms of disclosure*** for sensitive data:

- **exact data**
- **bounds**
  - e.g. giving a lower and an upper bound for the data item
- **negative result** revealing that the data item does not have a specific value can be compromising, in particular that the value not 0.
- the **existence** of a data may be sensitive, e.g. a criminal record
- **probable values:** it might be possible to determine the probability that an element has a certain value

Inference principle

INFERENCE
means deriving sensitive data from non-sensitive data



Inference Example

Name	Position	Salary ($)	Department	Dept. Manager
Andy	senior	43,000	strip	Cathy
Calvin	junior	35,000	strip	Cathy
Cathy	senior	48,000	strip	Cathy
Dennis	junior	38,000	panel	Herman
Herman	senior	55,000	panel	Herman
Ziggy	senior	67,000	panel	Herman

(a) Employee table

Position	Salary ($)
senior	43, 000
junior	35,000
senior	48,000

Name	Department
Andy	strip
Calvin	strip
Cathy	strip

(b) Two views

Name	Position	Salary ($)	Department
Andy	senior	43,000	strip
Calvin	junior	35,000	strip
Cathy	senior	48,000	strip

(c) Table derived from combining query answers

Inference Countermeasures

- inference detection at database design
  - alter database structure or access controls
- inference detection at query time
  - by monitoring and altering or rejecting queries
- needs some inference detection algorithm
  - a difficult problem
  - cf. employee-salary example

Statistical Databases

- provides data of a statistical nature
  - e.g. counts, averages
- two types:
  - pure statistical database
  - ordinary database with statistical access
    - some users have normal access, others statistical
- the access control objective is to allow statistical use without revealing individual entries

Statistical Database Security

- use a characteristic formula C
  - a logical formula over the values of attributes
  - e.g. `(Sex=Male) AND ((Major=CS) OR (Major=EE))`
- the query set X(C) is the set of records matching C
- a statistical query is a query that produces a value calculated over a query set

Statistical Database Example

(a) Database with statistical access with $N = 13$ students

Name	Sex	Major	Class	SAT	GP
Allen	Female	CS	1980	600	3.4
Baker	Female	EE	1980	520	2.5
Cook	Male	EE	1978	630	3.5
Davis	Female	CS	1978	800	4.0
Evans	Male	Bio	1979	500	2.2
Frank	Male	EE	1981	580	3.0
Good	Male	CS	1978	700	3.8
Hall	Female	Psy	1979	580	2.8
Iles	Male	CS	1981	600	3.2
Jones	Female	Bio	1979	750	3.8
Kline	Female	Psy	1981	500	2.5
Lane	Male	EE	1978	600	3.0
Moore	Male	CS	1979	650	3.5

(b) Attribute values and counts

Attribute A_j	Possible Values	$\|A_j\|$
Sex	Male, Female	2
Major	Bio, CS, EE, Psy, ...	50
Class	1978, 1979, 1980, 1981	4
SAT	310, 320, 330, ..., 790, 800	50
GP	0.0, 0.1, 0.2, ..., 3.9, 4.0	41

Statistical inference attacks

- **direct attack**
  - finding sensitive information directly with queries that yield only a few records
- **indirect attacks** seeks to infer the final result based on a number of intermediate statistical results
  - **sum**
  - **count**
  - **median**
  - **tracker attack:**

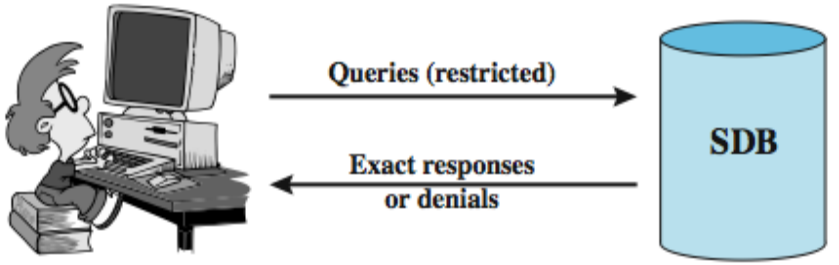
  means finding sensitive information by using additional queries that each produce a small result

Basic controls for statistical inference attacks

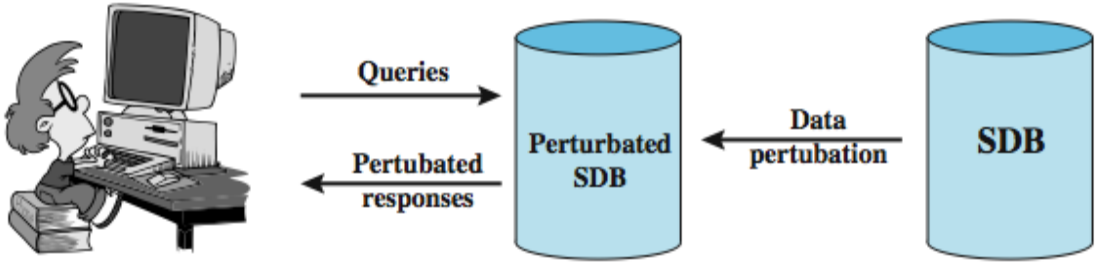
In general there are three types of controls:

- **query restriction** (suppression)
  - reject query without response (data withheld)
- **perturbation** (data or output) (concealing)
  - provide an inexact answer to the query
- **query analysis,** i.e. track what the user knows
  - keeping track on previous queries (query history)
  - maintain a record for each user of earlier queries
  – this method is extremely costly

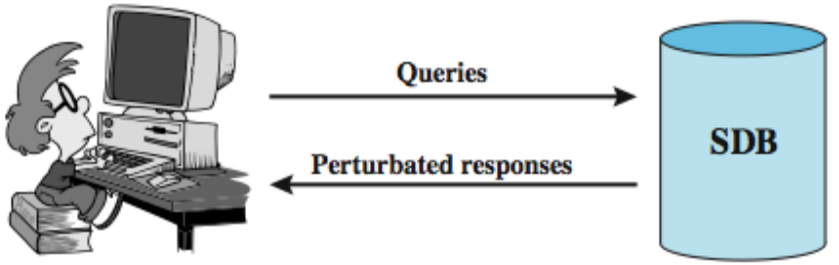
Protecting against inference in SDB's



(a) Query set restriction



(b) Data perturbation



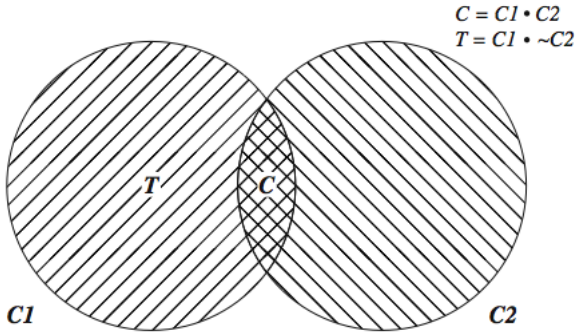
(c) Output perturbation

Control methods for statistical inference attacks

- **Query Size Restriction**
  - $k \leq |X(C)| \leq N - k$
- **combining results:** present values in ranges
  - combining rows or columns
  - rounding
- **random sample**
  - compute the result on a random sample of the database
- **random data perturbation**
  - add an error term e

Tracker attack example

- count(C) = 1 is forbidden due to query size restriction
- divide queries into parts
  - C = C1.C2
  - count(C) = count(C1) - count (C1. ~*C2)*
- combination is called a tracker
- each part acceptable
- overlap is desired result



Other query restrictions

- query set overlap control
  - limit overlap between new and previous queries
  - has problems and overheads
- partitioning
  - records are clustered into a number of mutually exclusive groups
  - only allow queries on entire groups
- query denial and information leakage
  - denials can leak information
  - to counter must track queries from user

Perturbation

- add noise to statistics generated from data
  - will result in differences in statistics
- data perturbation techniques
  - data swapping
  - generate statistics from underlying probability distribution of attributes
- output perturbation techniques
  - random-sample query (based on a subset)
  - statistic adjustment of result (random or not)
- perturbation techniques may result in loss of accuracy in results