Lecture notes for 6.4, 6.5, & 6.8 data mining

Name	Grade	Druggie	Wage
Sue	63	Y	7.75
Bob	59	Y	8.00
Elle	95	N	7.75
Mark	56	Y	6.25
Steve	89	N	8.50
Jane	61	N	15.00

Example 1:

Query 1: FETCH druggie WHERE name="Mark"
Suppose blocked for privacy

Query 2: FETCH minmax(grade) WHERE druggie=false
Query 3: FETCH minmax(grade) WHERE druggie=true
Query 4: FETCH grade WHERE name="Mark"
Discover that Mark is a druggie

Query 5: FETCH grade WHERE name="Jane"
Inconclusive... data exposure depends on context, prior knowledge, actual data

Example 2:

Query 1: FETCH average(wage) WHERE name=Jane
Supposed blocked... DBMS will not return data from single records for privacy

Query 2: FETCH count(names)
Query 3: FETCH average(wage)
Query 4: FETCH average(wage) WHERE name =/= Jane
Can compute Jane's wage from these three values

Example 3:

Query 1: FETCH grade WHERE name=Sue
Suppose blocked... not allowed to see individual student's grade

Query 2: FETCH grade WHERE name=Sue OR druggie=Y
Query 3: FETCH grade WHERE name=Sue OR druggie=N
Only the grade 63 appears in both sets, so learn Sue's grade

Sensitive data

- Entire database (military DB)
- Data in some, not all, DB fields or records
- Existence of some DB fields or records

Why sensitive

- Classified data
- From a sensitive source (e.g. A spy)
- Administratively set as sensitive (for whatever reason)
- Discloses sensitive information when combined with previously disclosed information

Types of disclosures

- Exact data
- Data range (may suffice to know that friends salary is above some threshold)
- Negative result (number of felonies is not zero)
- Existence of data (do not want employees to know you are monitoring certain aspects)

Inference

- Def: way to derive sensitive data from nonsensitive data
- Aggregate values may reveal information
  - If field sum is zero, any record contributing to the sum has field value zero
  - If count is 1, then any aggregate info (sum, mean, median) is exact value of matching record
  - Mean: example 2 above

Defenses

- Suppression
  - Refuse to return sensitive results
  - Refuse to return results based on small sample
- Concealing
  - Return inexact results (round to nearest 10...)
  - Compute based on random sampling of DB
  - Adjust values by random alteration
- Track what the user knows
  - Refuse to return results that allow inference
  - Extremely difficult

Data mining

- Def: extraction of meaningful information from large data sets
- Identifies patterns, relationships in data
- Correlation may not be causation... does not identify cause & effect