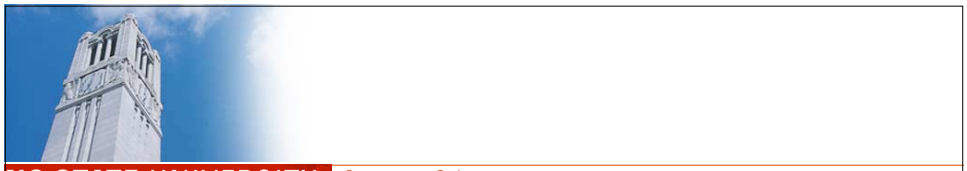


NC STATE UNIVERSITY Computer Science

CSC 405
Introduction to Computer Security

Topic 6.3: Database Inference Control

Outline

- Inference attacks
  - Direct attacks (no inference needed)
  - Indirect attacks via aggregations
  - Tracker attacks
  - Inference via linear systems
  - Inference via database constraints
- Inference control
  - Limited Response Suppression
  - Combining results
  - Random sample
  - Random data perturbation
  - Query analysis

Direct Attacks

Name	Sex	Race	Aid	Fines	Drugs	Dorm
Adams	M	C	5000	45	1	Holmes
Bailey	M	B	0	0	0	Grey
Chin	F	A	3000	20	0	West
Dewitt	M	B	1000	35	3	Grey
Earhart	F	C	2000	95	1	Holmes
Fein	F	C	1000	15	0	West
Groff	M	C	4000	0	3	West
Hill	F	B	5000	10	2	Holmes
Koch	F	C	0	0	1	West
Liu	F	A	0	10	2	Grey
Majors	M	C	2000	0	2	Grey

- Query 1
  - List NAME Where SEX=M ^ DRUGS=1
  - Results: _____

Direct Attacks (Cont'd)

Name	Sex	Race	Aid	Fines	Drugs	Dorm
Adams	M	C	5000	45	1	Holmes
Bailey	M	B	0	0	0	Grey
Chin	F	A	3000	20	0	West
Dewitt	M	B	1000	35	3	Grey
Earhart	F	C	2000	95	1	Holmes
Fein	F	C	1000	15	0	West
Groff	M	C	4000	0	3	West
Hill	F	B	5000	10	2	Holmes
Koch	F	C	0	0	1	West
Liu	F	A	0	10	2	Grey
Majors	M	C	2000	0	2	Grey

- Query 2
  - List NAME where (SEX=M ^ DRUGS=1) v
    (SEX !=M ^ SEX !=F) v (DORM=AYRES)
  - Result=_____

Direct Attacks (Cont'd)

- Protect against direct attacks
  - "*n* items over *k* percent" rule
    - Data should be withheld if *n* items represent over *k%* of the result reported.
    - Adopted by U.S. Census Bureau
    - Intuition: do not reveal results where a small number of records make up a large proportion of the category.
  - Release only statistics
    - Examples: sum, average, count, etc.

Indirect Attacks via Aggregations

Sums of Financial Aid by Dorm and Sex

	Holmes	Grey	West	Total
M	5000	3000	4000	12000
F	7000	0	4000	11000
Total	12000	3000	8000	23000

Female Students Living in Grey

Name
Liu

- Try to infer a sensitive value from a reported sum.
- What can we infer for the female students living in Grey?
  - ___'s financial aid = _____

Indirect Attacks via Aggregations (Cont'd)

Count of Financial Aid by Dorm and Sex

	Holmes	Grey	West	Total
M	1	3	1	5
F	2	1	3	6
Total	3	4	4	11

Male Students Living in Holmes or West

Name	Dorm
Adams	Holmes
Groff	West

- With additional counts, what can we further infer?
  - _____'s financial aid = _____
  - _____'s financial aid = _____

Tracker Attacks

- DBMS protection
  - Allow aggregation of sensitive attributes only when the number of data items that constitute the aggregate is more than a threshold *t*.
- Trackers defeats this protection by using additional queries.

Tracker Attacks (Cont'd)

Name	Sex	Race	Aid	Fines	Drugs	Dorm
Adams	M	C	5000	45	1	Holmes
Bailey	M	B	0	0	0	Grey
Chin	F	A	3000	20	0	West
Dewitt	M	B	1000	35	3	Grey
Earhart	F	C	2000	95	1	Holmes
Fein	F	C	1000	15	0	West
Groff	M	C	4000	0	3	West
Hill	F	B	5000	10	2	Holmes
Koch	F	C	0	0	1	West
Liu	F	A	0	10	2	Grey
Majors	M	C	2000	0	2	Grey

- Query 3
  - Sum ((Sex=F) ^ (Race=C) ^ (Dorm=Holmes))
  - Is this allowed?

Tracker Attacks (Cont'd)

- sum (a^b^c) = sum(a) – sum (a^ ¬ (b^c))
- sum ((Sex=F) ^ (Race=C) ^ (Dorm=Holmes))

  is equivalent to

  sum (Sex=F) –

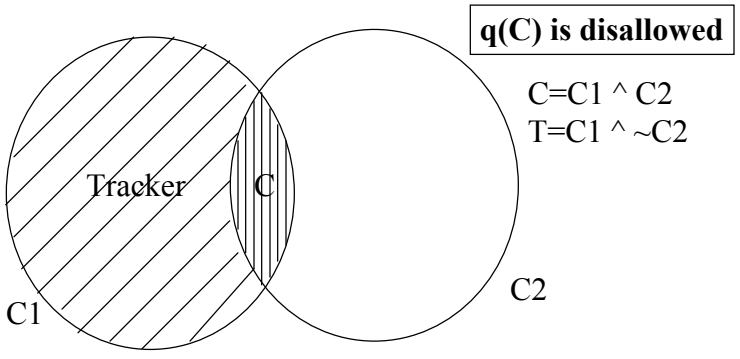
  sum ((Sex=F)^(Race!=C ∨ Dorm != Holmes)

Tracker Attacks (Cont'd)

Name	Sex	Race	Aid	Fines	Drugs	Dorm
Adams	M	C	5000	45	1	Holmes
Bailey	M	B	0	0	0	Grey
Chin	F	A	3000	20	0	West
Dewitt	M	B	1000	35	3	Grey
Earhart	F	C	2000	95	1	Holmes
Fein	F	C	1000	15	0	West
Groff	M	C	4000	0	3	West
Hill	F	B	5000	10	2	Holmes
Koch	F	C	0	0	1	West
Liu	F	A	0	10	2	Grey
Majors	M	C	2000	0	2	Grey

- Count ((Sex=F) ^ (Race=C) ^ (Dorm=Holmes))
  = _____ – _____ = _____

Tracker Attacks (Cont'd)



q(C)=q(C1) – q(T)

Inference via Linear Systems

- Generalization of the Tracker attacks
- We can get a sequence of linear equations through a sequence of queries
  - Variables: sensitive values
  - Q1 = c1 + c2 + c3 + c4 + c5
  - Q2 = c1 + c2 + c4
  - Q3 = c3 + c4
  - Q4 = c4 + c5
  - Q5 = c2 + c5
  - C5 = ((Q1 – Q2) – (Q3 – Q4))/2.

Inference via Database Constraints

- Integrity constraints
- Database dependencies
- Key integrity

Integrity Constraints

- C=A+B
- A=public, C=public, and B=secret
- B can be calculated from A and C, i.e., secret information can be calculated from public data

Database Dependencies

- Knowledge about the database could be used to make inference
  - Functional dependencies
  - Multi-valued dependencies
  - Join dependencies
  - etc.

Functional Dependency

- FD: A → B, that is for any two tuples in the relation, if they have the same value for A, they must have the same value for B.
- Example: FD: Rank → Salary
- Secret information: Name and Salary together
  - Query1: Name and Rank
  - Query2: Rank and Salary
  - Combine answers for query1 and 2 to reveal Name and Salary together

Inference Controls

- Two ways
  - Suppression
    - Sensitive data values are *not provided*
    - Query is rejected without response
  - Concealing
    - The answer provided is *close to* but not exactly the actual value.
  - Both can be applied to either queries or individual items within the database.

Limited Response Suppression

- Suppression technique
- Eliminate low-frequency elements
  - Not always work.

Student by Dorm and Sex

	Holmes	Grey	West	Total
M	--	3	--	5
F	2	--	3	6
Total	3	4	4	11

What are the suppressed values?

Combining Results

- Suppression techniques
  - Combine rows or columns to protect sensitive values.
  - Present results in ranges
  - Rounding

Students by Sex and Drug Use

Sex	Drug Use			
	0	1	2	3
M	1	1	1	2
F	2	2	2	0

Sex	Drug Use	
	0 or 1	2 or 3
M		
F		

Random Sample

- Concealing technique
  - Use random sample of the database to answer queries.
  - The same sample set should be chosen for equivalent queries.
    - Prevent averaging attacks

Random Data Perturbation

- Concealing technique
  - Perturb the values of the database by a small error.
  - Statistical measures such as sum and mean will be close.
  - Easier than random sample.

Query Analysis

- Suppression technique
  - Decide whether a result should be provided through analyzing queries and their implications.
  - Need to maintain a query history
  - Difficult to know what a user knows from out-of-band ways.

Methodologies of Inference Control

- Suppress obviously sensitive information
  - Easy to do, but tend to be over restrictive
- Track what user knows
  - Very expensive
    - Query history
  - Cannot deal with conspiracy
- Disguise data
  - Sacrifice the quality of data

Conclusions

- No general technique is available to solve the problem
- Need assurance of protection
- Hard to incorporate outside knowledge