590db:Topics in Secure Data Management

Alon Halevy, Dan Suciu

Today's Outline

- Traditional database security
- New data security needs
- Reflections on cryptographic primitives
- Why now ? Why us ?
- The reading list

Data Security

Dorothy Denning, 1982:

- Data Security is the science and study of methods of protecting data (...) from unauthorized disclosure and modification

Traditional Data Security

- Access control
- Security in Statistical Databases

Traditional Data Security

Access control:

- The System R authorization model [Griffith and Wade'76], [Fagin'78]
- grant/revoke privileges to/from users
  - privileges = select/insert/delete/update
- Followed by extensions, improvements, generalizations to the OO data model

Traditional Data Security

Access control in SQL:

- SQL: grant/revoke privileges

  GRANT <privileges> ON <object> TO <users>

  [WITH GRANT OPTION]

  REVOKE <privileges> ON <object> FROM <users> [RESTRICT | CASCADE]

- authorization graph, simple semantics

Traditional Data Security

Access control:

- great success story of the db commuity...
- ...or spectacular failure
  - SAP uses it's own security layer
- Main assumption: data on trusted server
- The real challenge: securing the server

Traditional Data Security

Security in statistical databases:

- Main question:
  - Allow: "Find the average salary"
  - Deny: "Find Alice's salary"
- Research is much harder...
- ...but results are mostly negative
- We will find out next week

New Issues in Secure Data Management

- Today: global sharing of data and services
- Issues:
  - protect data but allow sharing/integration
  - protect data when gets disseminated
  - protect queries, rather than data
  - outsource data processing
- Next slides: a random selection of applications

Latanya Sweeney's Finding

87% of the US population (216 million out of 248 million) are likely to be uniquely identified based on:

zipcode, gender, date-of-birth

Latanya Sweeney's Finding

- In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees
- GIC collects data, and since it's "private", it publishes it:
- GIC(zip, dob, sex, diagnosis, procedure, ...)

Latanya Sweeney's Finding

- Sweeney paid $20 and bought the voter registration list for Cambridge Massachusetts:
- VOTER(name, party, ..., zip, dob, sex)

Latanya Sweeney's Finding

GIC(<u>zip, dob, sex</u>, diagnosis, procedure, ...)

- VOTER(name, party, ..., <u>zip, dob, sex</u>)
- William Weld (former governor) lives in Cambridge, hence is in VOTER
- 6 people in VOTER share his <u>dob</u>
- only 3 of them were man (same <u>sex</u>)
- Weld was the only one in that <u>zip</u>
- Sweeney learned Weld's medical records !

Shared Query Processing

- Alice has a database DB_A
- Bob has a database DB_B
- How can they compute Q(DB_A, DB_B), without revealing their data ?

Shared Query Processing: Example

Alice: I am teaching Databases, and I suspect I have a cheater in my class

Bob: I am teaching Security, and I also suspect a cheater !

Alice: Tell me your suspect's name ! If it's the same as mine, then I'll know for sure he is cheating.

Bob: No. I'm not sure my suspect is cheating. Tell me your suspect's name first, and if they match I'll let you know.

Alice: No.

Private Query Processing

- Scenario: a service offers access to a useful database to customers
- Goal: want to process user queries, but keep queries secret from the engine
- Special case: "Private Information Retrieval", PIR

Database as a Service

Scenario:

- Alice has a database D
- Bob has a database engine, and offers to store and process D, for a fee
- But Alice doesn't trust Bob, and wants to hid the data from him

Access Control Through Encryption

- In DBMS access control enforced by the server
- When data is published, we need to rely on encryption instead
- How can we encrypt the data to enforce multiple policies ?
- How can we be sure nothing else leaks ?
- How can we process the encrypted data efficiently ?

Watermarking

- Want to sell a database instance
- But want to be able to trace the source
- Watermark:
  - small, hidden perturbations in the database that prove its origin
- How can one do that ?
- Variation: fingerprinting

Authenticating Databases

- Alice stores a database and process queries
- Bob submits queries, but doesn't trust Alice
- Both trust Trent, who can sign facts in the database, but doesn't want to sign every single query that Alice answers for Bob
- How can Alice convince Bob that the answer to a query is correct ?
- "Authenticating Data Structures"

A Partial Classification

Gerome's classification, based on Trust Domains

Data	Execution	Query	Client-server DBMS
Data	Query	Execution	Database as a service
Data	Query	Execution	Data publishing

Cryptographic Primitives

- Encryption: symmetric (private), asymmetric (public)
- One-way hash functions
- They are secure, aren't they ?
  - Surprisingly few positive theoretical results !
  - Worse than the P = NP question

Cryptographic Primitives

- Secret Sharing: have a value p
- Compute:
  - v_A = a random number
    - v_B = v_A xor p
- Give v_A to Alice, v_B to Bob
- Neither Alice nor Bob learn anything about p
- Together, they can recover p = v_A xor v_B
- Note: this forms the basis for one approach in PIR

Cryptographic Primitives

- Oblivious Communication:
- Alice holds two values (x,y)
- Bob holds one value, z (z=0, or z=1)
- They communicate…
- … at the end Bob learns either x (when z=0) or learns y (when z=1), Alice learns nothing
- Note: used in multiparty computation

Zero Knowledge Proofs

Jayant graduates, interviews on Wall Street for top job

Vice President (VP): We need to match schemas S1 and S2. Nobody succeeded so far...

Jayant: I matched them already. They are isomorphic !

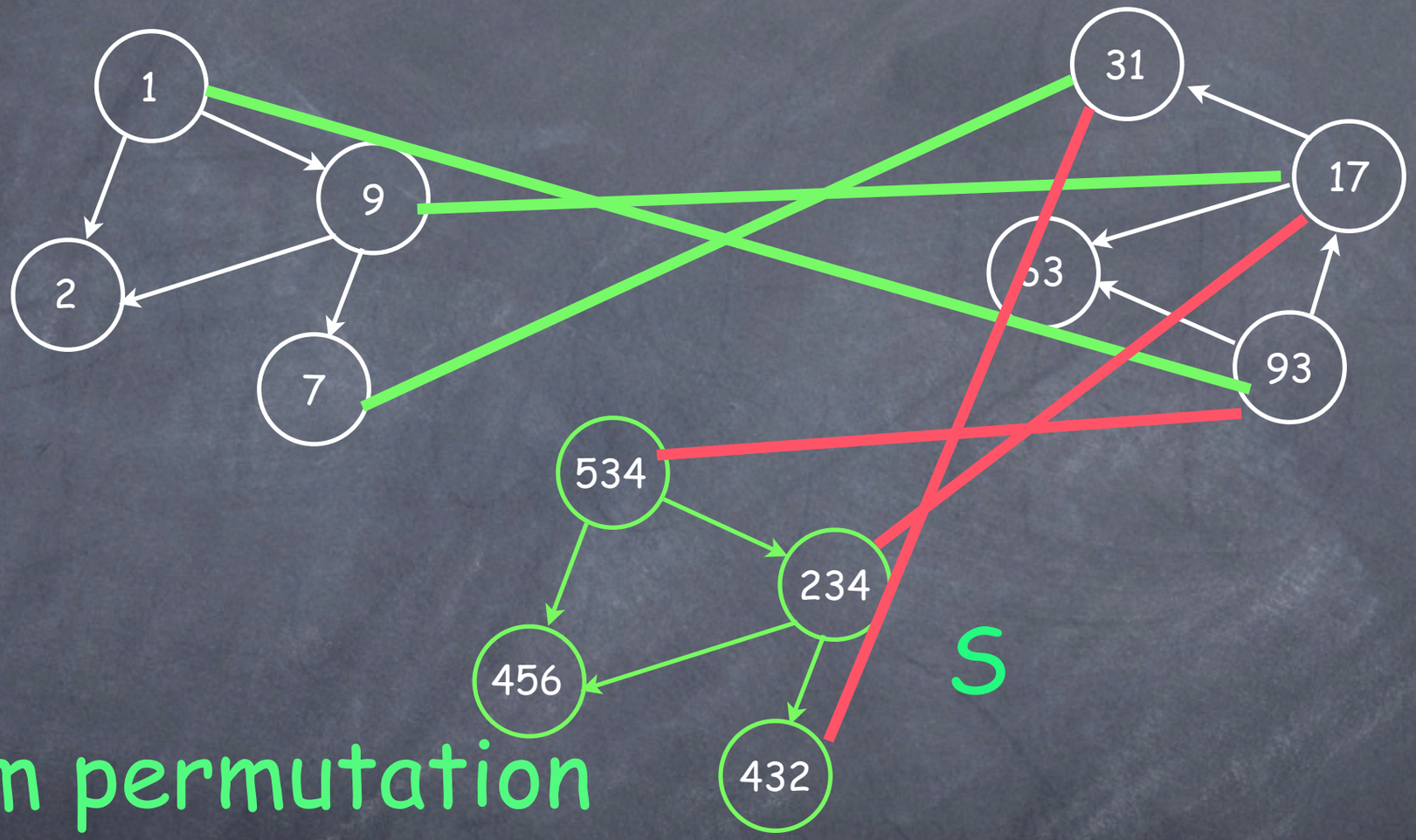
VP: REALLY ? Show me your isomorphism !

Jayant: No. I want the offer first, with a big signup bonus.

VP: No. I want to see first that you found the isomorphism.

Zero Knowledge Proofs

S1



S2

S

Jayant:
random permutation

VP: S2

Jayant: shows isomorphism S -- S2

Repeat several times

Why Now ? Why Us ?

- Security becoming key research topic in data management
- Already renewed interest in the database community (recent SIGMOD/PODS/VLDB papers)
- Mainstream cryptography cannot offer all solutions: need key understanding of data management
- Research needs to be conducted differently: graphs and experiments DO NOT CONVINCE of security !
  - theory is non-optional !

The Reading List

10/14	Security in Statistical DB	Ashish
10/21	Access Control	Ashish
10/28	Multiparty Secure Computation	
11/4	Database as a Service	
11/11	Cryptography in Data	Gerome
11/18	Data Privacy	
11/25	Privacy in Data Mining	
12/2	Data Authentication	
12/9	Watermarking	