Outline

- Introduction
- Background
- Distributed DBMS Architecture
- Distributed Database Design
- Distributed Query Processing
- Distributed Transaction Management
- Building Distributed Database Systems (RAID)
- Mobile Database Systems
- Privacy, Trust, and Authentication
- Peer to Peer Systems
Outline

1. Assuring privacy in data dissemination
2. Privacy-trust tradeoff
3. Privacy metrics
3. Privacy Metrics

Problem
- How to determine that certain degree of data privacy is provided?

Challenges
- Different privacy-preserving techniques or systems claim different degrees of data privacy
- Metrics are usually ad hoc and customized
  - Customized for a user model
  - Customized for a specific technique/system
- Need to develop uniform privacy metrics
  - To confidently compare different techniques/systems
Requirements for Privacy Metrics

- Privacy metrics should account for:
  - Dynamics of legitimate users
    - How users interact with the system?
      - E.g., repeated patterns of accessing the same data can leak information to a violator
  - Dynamics of violators
    - How much information a violator gains by watching the system for a period of time?
  - Associated costs
    - Storage, injected traffic, consumed CPU cycles, delay
Proposed Approach

A. Anonymity set size metrics
B. Entropy-based metrics
A. Anonymity Set Size Metrics

- The larger set of indistinguishable entities, the lower probability of identifying any one of them
  - Can use to "anonymize" a selected private attribute value within the domain of its all possible values

"Hiding in a crowd"

"Less" anonymous (1/4)

"More" anonymous (1/n)
Anonymity Set

- Anonymity set $A$
  
  $A = \{(s_1, p_1), (s_2, p_2), \ldots, (s_n, p_n)\}$
  
  - $s_i$: subject $i$ who might access private data
    
    or: $i$-th possible value for a private data attribute
  
  - $p_i$: probability that $s_i$ accessed private data
    
    or: probability that the attribute assumes the $i$-th possible value
Effective Anonymity Set Size

- Effective anonymity set size is

\[
L = |A| \sum_{i=1}^{\lvert A \rvert} \min(p_i, \frac{1}{\lvert A \rvert})
\]

- Maximum value of \( L \) is \( |A| \) iff all \( p_i \)'s are equal to \( 1/|A| \)
- \( L \) below maximum when distribution is skewed
  - skewed when \( p_i \)'s have different values

- Deficiency:
  \( L \) does not consider violator’s *learning* behavior
B. Entropy-based Metrics

- Entropy measures the randomness, or uncertainty, in private data
- When a violator gains more information, entropy decreases
- Metric: Compare the current entropy value with its maximum value
  - The difference shows how much information has been leaked
Dynamics of Entropy

- Decrease of system entropy with attribute disclosures (capturing dynamics)

- When entropy reaches a threshold (b), data evaporation can be invoked to increase entropy by controlled data distortions.
- When entropy drops to a very low level (c), apoptosis can be triggered to destroy private data.
- Entropy increases (d) if the set of attributes grows or the disclosed attributes become less valuable – e.g., obsolete or more data now available.
Quantifying Privacy Loss

- Privacy loss $D(A,t)$ at time $t$, when a subset of attribute values $A$ might have been disclosed:

$$D(A,t) = H^*(A) - H(A,t)$$

- $H^*(A)$ – the maximum entropy
  - Computed when probability distribution of $p_i$’s is uniform
- $H(A,t)$ is entropy at time $t$

- $w_j$ – weights capturing relative privacy “value” of attributes

$$H(A,t) = \sum_{j=1}^{\left|A\right|} w_j \left( \sum_{\forall i} \left( - p_i \log_2(p_i) \right) \right)$$
Using Entropy in Data Dissemination

- Specify two thresholds for $D$
  - For triggering evaporation
  - For triggering apoptosis

- When private data is exchanged
  - Entropy is recomputed and compared to the thresholds
  - Evaporation or apoptosis may be invoked to enforce privacy
Secure Data Warehouse
Data warehouse is an integrated repository derived from multiple distributed source databases.

Created by replicating or transforming source data to new representation.

Some data can be web-database or regular databases (relational, files, etc.).

Warehouse creation involves reading, cleaning, aggregating, and storing data.

Warehouse data is used for strategic analysis, decision making, market research types of applications.

Open access to third party users.
Examples:

- Human genome databases.
- Drug-drug interactions database created by thousands of doctors in hundreds of hospitals.
- Stock prices, analyst research.
- Teaching material (slides, exercises, exams, examples).
- Census data or similar statistics collected by government.
Ideas for Security

- Replication
- Aggregation and Generalization
- Exaggeration and Mutilation
- Anonymity
- User Profiles, Access Permissions
Anonymity

One can divulge information to a third party without revealing where it came from and without necessarily revealing the system has done so.

- User privacy and warehouse data privacy.
- User does not know the source of data.
- Warehouse system does not store the results and even the access path for the query.
- Separation of storage system and audit query system.
- Non-intrusive auditing and monitoring.
- Distribution of query processing, logs, auditing activity.
- Secure multi-party computation.
- Mental poker (card distribution).
Equivalent Views

- Witness (Permission Inference)
  User can execute query $Q$ if there is an equivalent query $Q'$ for which the user has permission. Security is on result and not computation.

- Create views over mutually suspicious organizations by filtering out sensitive data.
Similarity Depends on Application

- Two objects might be similar to a K-12 student, but not a scientist.
- 1999 and 1995 annual reports of the CS department might be similar to a graduate school applicant, but not to a faculty applicant.

**Goal:** Use ideas of replication to provide security by using a variety of similarity criterion

**Goal:** Different QoS to match different classes of users.
Similarity Based Replication

SOME DEFINITIONS:

- **Distinct functions** used to determine how similar two objects are (Distinct Preserving Transformations).
- **Precision**: fraction of retrieved data as needed (relevant) for the user query.
- **False Positive**: object retrieved that is similar to the data needed by query, but it is not.
- **False Negative**: object is needed by the query, but not retrieved.
Access Permission

- Information permission (system-wide)
  - (employee salary is releasable to payroll clerks and cost analyst).

- Physical permission (local)
  - (cost analysts are allowed to run queries on the warehouse).
Cooperation Instead of Autonomy in Warehouse

- In UK, the Audit Commission estimated losses of the order of $2 billion.
- Japanese Yakuza made a profit of $7 billion.
- A secure organization needs to secure data, as well as its interpretation.
  (Integrity of data OK, but the benefit rules were interpreted wrong and misapplied.)
  ⇒ Interpretation Integrity
Extensions to the SQL Grant/Revoke Security Model

- Limitation is a generalization of revoke.
- Limitation Predicates should apply to only paths (reduces chance of inadvertent & malicious denial of service).
- One can add either limitation or reactivation, or both.
- Limitation respects lines of authority.
- Flexibility can be provided to limitation.
Aggregation and Generalization

- Summaries, Statistics
  - (over large or small set of records)
  - (various levels of granularity)
- Graphical image with numerical data.
- Reduce the resolution of images.
- Approximate answers
  - (real-time vs. delayed quotes, blood analysis results)
- Inherit access to related data.
Dynamic

- Authenticate users dynamically and provides access privileges.
  - Mobile agent interacts with the user and provides authentication and personalized views based on analysis and verification.
- Rule-based interaction session.
- Analysis of the user input.
- Determination of the user’s validity and creating a session id for the user and assignment of access permission.
Exaggeration and Misleading

- Give low or high range of normal values. Initially (semantically normal).
- Partially incorrect or difficult to verify data. Quality improves if security is assured.
- Give old data, check damage done, give better data.
- Projected values than actual values.
User Profile

- User profiles are used for providing different levels of security.
- Each user can have a profile stored at the web server or at third party server.
- User can change profile attributes at run-time.
- User behavior is taken into account based on past record.
- Mobile agent accesses the web page on behalf of the user and tries to negotiate with web server for the security level.
User Profile

- Personal category
  - personal identifications; name, dob, ss, etc.

- Data category
  - document content; keywords
  - document structure; audio/video, links
  - source of data

- Delivery data – web views, e-mail

- Secure data category
Static

- Predefined set of user names, domain names, and access restrictions for each
  - (restricted & inflexible)
- Virtual view, Materialized view, Query driven
- Build user profiles and represent them
  - past behavior
  - feedback
  - earlier queries
  - type, content and duration