Hippocratic Databases and Fine Grained Access Control

Li Xiong

CS573 Data Privacy and Security
Review

- Anonymity - an individual (or an element) not identifiable within a well-defined set
- Confidentiality - information is accessible only to those authorized to have access
- Access control - control which principles have access to which resources
- Privacy - the right of individuals to determine for themselves when, how and to what extent information about them is communicated to others.
From Access Control to Hippocratic Databases and Fine Grained Access Control

- Access control - control which principles have access to which resources
- Traditional database security provided by access control
  - Control which user have access to which table
- We need to re-architect database systems to include responsibility for the privacy of data.
Hippocratic databases (Agrawal ‘02)

- A vision, inspired by the Hippocratic Oath, of databases that preserve privacy
- Key privacy principles
- A strawman design for a Hippocratic database
- Technical challenges
Hippocratic Oath

“And about whatever I may see or hear in treatment, or even without treatment, in the life of human beings – things that should not ever be blurted out outside – I will remain silent, holding such things to be unutterable.”
Traditional Databases

- **Fundamental to a database system is**
  1. Ability to manage persistent data.
  2. Ability to access a large amount of data efficiently.

- **Universal capabilities of a database system**
  1. Support for at least one data model.
  2. Support for certain high-level languages that allow the user to define the structure of data, access data, and manipulate data.
  3. Transaction management, the capability to provide correct, concurrent access to the database by many users at once.
  4. Access control, the ability to deny access to data by unauthorized users and the ability to check the validity of the data.
  5. Resiliency, the ability to recover from system failures without losing data.
Hippocratic Databases

- Hippocratic databases require all the capabilities provided by current database systems
- Different focus
- Need to rethink data definition and query languages, query processing, indexing and storage structures, and access control mechanisms
Hippocratic Databases vs. Statistical Databases

- Hippocratic databases vs. Statistical databases
  - Hippocratic databases share the goal of preventing disclosure of private information
  - but the class of queries for Hippocratic databases is much broader.
- Hippocratic databases vs. traditional access control
  - Hippocratic databases requires more complex privacy policy management and more fine-grained access control
Privacy Regulations

United States Privacy Act of 1974 requires federal agencies to

1. permit an individual to determine what records pertaining to him are collected, maintained, used, or disseminated;
2. permit an individual to prevent records pertaining to him obtained for a particular purpose from being used or made available for another purpose without his consent;
3. permit an individual to gain access to information pertaining to him in records, and to correct or amend such records;
4. collect, maintain, use or disseminate any record of personally identifiable information in a manner that assures that such action is for a necessary and lawful purpose, that the information is current and accurate for its intended use, and that adequate safeguards are provided to prevent misuse of such information;
5. permit exemptions from the requirements with respect to the records provided in this Act only in those cases where there is an important public policy need for such exemption as has been determined by specific statutory authority; and
6. be subject to civil suit for any damages which occur as a result of willful or intentional action which violates any individual’s right under this Act.
Privacy Regulations

Recent privacy documents

- 1996 Health Insurance Portability and Accountability Act (HIPAA)
- 1999 Gramm-Leach-Bliley Financial Services Modernization Act
- 2000 Personal Information Protection and Electronic Documents Act (PIPEDA)
- 2003 Personal Information Protection Act (PIPA)
Guidelines

- **Collection**
- **Retention**
- **Use**
- **Disclosure**

*Example: Grad student information at the university*
Ten Founding Principles

1. **Purpose Specification.** For personal information stored in the database, the purposes for which the information has been collected shall be associated with that information.

2. **Consent.** The purposes associated with personal information shall have consent of the donor of the personal information.

3. **Limited Collection.** The personal information collected shall be limited to the minimum necessary for accomplishing the specified purposes.

4. **Limited Use.** The database shall run only those queries that are consistent with the purposes for which the information has been collected.

5. **Limited Disclosure.** The personal information stored in the database shall not be communicated outside the database for purposes other than those for which there is consent from the donor of the information.
Ten Founding Principles

6. **Limited Retention.** Personal information shall be retained only as long as necessary for the fulfillment of the purposes for which it has been collected.

7. **Accuracy.** Personal information stored in the database shall be accurate and up-to-date.

8. **Safety.** Personal information shall be protected by security safeguards against theft and other misappropriations.

9. **Openness.** A donor shall be able to access all information about the donor stored in the database.

10. **Compliance.** A donor shall be able to verify compliance with the above principles. Similarly, the database shall be able to address a challenge concerning compliance.
Strawman Design

- Use purpose as the central concept
- Use scenario
  - Mississippi is an on-line bookseller who needs to obtain certain minimum personal information to complete a purchase transaction. This information includes name, shipping address, and credit card number.
  - Mississippi also needs an email address to notify the customer of the status of the order.
  - Mississippi uses the purchase history of customers to offer book recommendations on its site.
  - It also publishes information about books popular in the various regions of the country (purchase circles).
The Characters

- Name: Alice
- Privacy fundamentalist
- Does not want Mississippi to retain any information once her purchase transaction is complete.
The Characters

- Name: Bob
- Privacy pragmatist
- Likes the convenience of providing his email and shipping address only once by registering at Mississippi.
- Also likes recommendations but he does not want his transactions used for purchase circles.
The Characters

- Name: Mallory
- Employee with questionable ethics
- The database and privacy officer must ensure that she is not able to obtain more information that she is supposed to.
Privacy meta data

- Privacy meta data defines for each purpose, and for each piece of information collected for that purpose:
  - Authorized-users: set of users (applications) who can access this information
  - External-recipients: whom the information can be given out to
  - Retention-period: how long the information is stored
- Privacy-policies table – external recipients and retention period
- Privacy-authorization table – access supporting the policies
Privacy Metadata

Privacy Metadata Schema

<table>
<thead>
<tr>
<th>table</th>
<th>attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>privacy-policies</td>
<td>purpose, table, attribute, { external-recipients }, retention</td>
</tr>
<tr>
<td>privacy-authorizations</td>
<td>purpose, table, attribute, { authorized-users }</td>
</tr>
</tbody>
</table>

Database Schema

<table>
<thead>
<tr>
<th>table</th>
<th>attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>customer</td>
<td>purpose, customer-id, name, shipping-address, email, credit-card-info</td>
</tr>
<tr>
<td>order</td>
<td>purpose, customer-id, transaction-id, book-info, status</td>
</tr>
</tbody>
</table>

Privacy-Policies Table

<table>
<thead>
<tr>
<th>purpose</th>
<th>table</th>
<th>attribute</th>
<th>external-recipients</th>
<th>retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>purchase</td>
<td>customer</td>
<td>name</td>
<td>{ delivery-company, credit-card-company }</td>
<td>1 month</td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>shipping-address</td>
<td>{ delivery-company }</td>
<td>1 month</td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>email</td>
<td>empty</td>
<td>1 month</td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>credit-card-info</td>
<td>{ credit-card-company}</td>
<td>1 month</td>
</tr>
<tr>
<td>purchase</td>
<td>order</td>
<td>book-info</td>
<td>empty</td>
<td>1 month</td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>name</td>
<td>empty</td>
<td>3 years</td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>shipping-address</td>
<td>empty</td>
<td>3 years</td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>email</td>
<td>empty</td>
<td>3 years</td>
</tr>
<tr>
<td>recommendations</td>
<td>order</td>
<td>book-info</td>
<td>empty</td>
<td>10 years</td>
</tr>
<tr>
<td>purchase-circles</td>
<td>customer</td>
<td>shipping-address</td>
<td>empty</td>
<td>1 year</td>
</tr>
<tr>
<td>purchase-circles</td>
<td>order</td>
<td>book-info</td>
<td>{ aggregated-all }</td>
<td>1 year</td>
</tr>
</tbody>
</table>
## Privacy Metadata

### Privacy-Authorizations Table

<table>
<thead>
<tr>
<th>purpose</th>
<th>table</th>
<th>attribute</th>
<th>authorized-users</th>
</tr>
</thead>
<tbody>
<tr>
<td>purchase</td>
<td>customer</td>
<td>customer-id</td>
<td>all</td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>name</td>
<td>{ shipping, charge, customer-service }</td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>shipping-address</td>
<td>{ shipping }</td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>email</td>
<td>{ shipping, customer-service }</td>
</tr>
<tr>
<td>purchase</td>
<td>customer</td>
<td>credit-card-info</td>
<td>{ charge }</td>
</tr>
<tr>
<td>purchase</td>
<td>order</td>
<td>customer-id</td>
<td>all</td>
</tr>
<tr>
<td>purchase</td>
<td>order</td>
<td>transaction-id</td>
<td>all</td>
</tr>
<tr>
<td>purchase</td>
<td>order</td>
<td>book-info</td>
<td>{ shipping }</td>
</tr>
<tr>
<td>purchase</td>
<td>order</td>
<td>status</td>
<td>{ shipping, customer-service }</td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>customer-id</td>
<td>all</td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>name</td>
<td>{ registration, customer-service }</td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>shipping-address</td>
<td>{ registration }</td>
</tr>
<tr>
<td>registration</td>
<td>customer</td>
<td>email</td>
<td>{ registration, customer-service }</td>
</tr>
<tr>
<td>recommendations</td>
<td>order</td>
<td>customer-id</td>
<td>{ mining }</td>
</tr>
<tr>
<td>recommendations</td>
<td>order</td>
<td>transaction-id</td>
<td>{ mining }</td>
</tr>
<tr>
<td>recommendations</td>
<td>order</td>
<td>book-info</td>
<td>{ mining }</td>
</tr>
<tr>
<td>purchase-circles</td>
<td>customer</td>
<td>customer-id</td>
<td>{ olap }</td>
</tr>
<tr>
<td>purchase-circles</td>
<td>customer</td>
<td>shipping-address</td>
<td>{ olap }</td>
</tr>
<tr>
<td>purchase-circles</td>
<td>order</td>
<td>customer-id</td>
<td>{ olap }</td>
</tr>
<tr>
<td>purchase-circles</td>
<td>order</td>
<td>book-info</td>
<td>{ olap }</td>
</tr>
</tbody>
</table>
Data Collection

- Matching privacy policy with user preferences
  - Privacy Constraint Validator checks whether the business’s privacy policy is acceptable to the user
  - Example: If Alice required a 2 week retention period, the database would reject the transaction

- Data insertion
  - Data is inserted with the purpose for which it may be used
Queries

- Submitted to the database along with their purpose. Example: recommendations
- Before query execution: Attribute Access Control checks privacy-authorizations table for a match on purpose, attribute and user.
  - Mallary (customer service) queries creditcard-info with “purchase”
  - authorized-users: charge
Queries

- During query execution: Record Access Control ensures that only records whose purpose attribute includes the query’s purpose will be visible to the query.
  - E.g. queries with “recommendations” will see Bob’s books but not Alice’s
  - Alice’s purpose attribute: purchase
Queries

- After query execution: Query Intrusion Detector is run on the query results to spot queries whose access pattern is different from the usual access pattern for queries with that purpose and by that user.
- An audit trail of all queries is maintained for external privacy audits, as well as addressing challenges regarding compliance.
Other Features

- Data Retention Manager deletes data items that have outlived their purpose.
- Data Collection Analyzer examines the set of queries for each purpose to determine if any information is being collected but not used. (Limited Collection).
- DCA determines if data is being kept for longer than necessary. (Limited Retention)
- DCA determines if people have unused (unnecessary) authorizations to issue queries with a given purpose. (Limited Use)
- Encryption Support allows some data items to be stored in encrypted form to guard against snooping.
P3P and Hippocratic Databases

- Platform for Privacy Preferences (P3P)
- A P3P policy describes the purpose of the collection of information along with intended recipients and retention period.
- The sites policy is programmatically compared to a user’s privacy preferences
- Very few implementations
New Challenges – Language

- P3P language insufficient
  - Developed for web shopping ⇒ language restricted
  - P3P is a good starting for a language which can be used in a wider variety of environments such as finance, insurance, and health care
  - Difficult to find balance between expressibility and usability
- Work is being done to arrange purposes in a hierarchy rather than the flat space that P3P uses
New Challenges – Efficiency

- What type of performance hit will integrated privacy checking entail?
- Some techniques from multilevel secure databases will apply
- Storage of purpose – space versus efficiency
Challenges – Limited Collection

- **Access Analysis**: Analyze the queries for each purpose and identify attributes that are collected for a given purpose but not used.
  - Problem: Necessity of one attribute may depend on others

- **Granularity Analysis**: Analyze the queries for each purpose and numeric attribute and determine the granularity at which information is needed – (data generalization?)

- **Minimal Query Generation**: Generate the minimal query that is required to solve a given problem.
New Challenges – Others

- Compliance
  - Query auditing and compliance checking
- Limited retention
  - How to delete a record not only from the table, but logs w/o affecting recovery
  - How to support historical analysis
- Openness
  - How to allow Alice to find out what databases have information about her?
Conclusion

- Presented a vision, inspired by the Hippocratic Oath, of databases that preserve privacy
- Enunciated key privacy principles
- Discussed a strawman design for a Hippocratic database
- Identified technical challenges
Limiting disclosure in Hippocratic databases (Lefevre ‘04)

- One approach to implement the privacy policy enforcement for Hippocratic databases and in general fine-grained access control
- Support of privacy policies
- Support of cell-level access control
  - Table semantics
  - Query semantics
Implementation Architecture
Policy definition

- A policy meta-language for defining privacy policy rules
- A policy is a set of rules \(<\text{data}, \text{purpose-recipient pair}, \text{condition}>\)
  - E.g. \(<\text{address}, \text{solicitation-charity}, \text{optin = yes}>\)
- Potential difficulties in translating from high-level policy to meta specifications
Access control

- **Table semantics (independent of queries)**
  - For each table, define a view for each purpose-recipient pair
    - Prohibited values are replaced with null based on the policy constraints
  - Queries are evaluated against the view

- **Query semantics (take queries into account)**
  - For the table in the FROM clause, define a view for the querying purpose-recipient pair
  - Result tuples that are null in all columns are discarded
Example

<table>
<thead>
<tr>
<th>P#</th>
<th>Name</th>
<th>Age</th>
<th>Address</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alice Adams</td>
<td>10</td>
<td>1 April Ave.</td>
<td>111-1111</td>
</tr>
<tr>
<td>2</td>
<td>Bob Blaney</td>
<td>20</td>
<td>2 Brooks Blvd.</td>
<td>222-2222</td>
</tr>
<tr>
<td>3</td>
<td>Carl Carson</td>
<td>30</td>
<td>3 Cricket Ct.</td>
<td>333-3333</td>
</tr>
<tr>
<td>4</td>
<td>David Daniels</td>
<td>40</td>
<td>4 Dogwood Dr.</td>
<td>444-4444</td>
</tr>
</tbody>
</table>

Figure 2: Full data table of patient information.

<table>
<thead>
<tr>
<th>P#</th>
<th>P#</th>
<th>Name</th>
<th>Age</th>
<th>Address</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

Figure 3: Patient choices for disclosure of information to charities for solicitation.
### Example

<table>
<thead>
<tr>
<th>P#</th>
<th>Name</th>
<th>Age</th>
<th>Address</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alice Adams</td>
<td>10</td>
<td>1 April Ave.</td>
<td>111-1111</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>3 Cricket Ct.</td>
<td>333-3333</td>
</tr>
<tr>
<td>4</td>
<td>David Daniels</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 4: Privacy-enforced table of patient information, using table semantics.

### Figure 5: Comparing Table Semantics and Query Semantics for a simple projection

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice Adams</td>
<td>10</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>David Daniels</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice Adams</td>
<td>10</td>
</tr>
<tr>
<td>David Daniels</td>
<td>-</td>
</tr>
</tbody>
</table>
Query Modification

- Query modification algorithms to enforce the privacy conditions at cell-level

SELECT Phone FROM Patients

SELECT
CASE WHEN EXISTS
  (SELECT phone_choice FROM PatientChoices
   WHERE Patient.P# = PatientChoices.P# AND PatientChoices.Phone_Choice = 1)
THEN phone ELSE null END
FROM patients
WHERE EXISTS
  (SELECT ID_Choice FROM PatientChoices
   WHERE Patient.P# = PatientChoices.P# AND PatientChoices.Phone_Choice = 1)
Overhead and scalability

![Chart showing overhead and scalability with data table size and elapsed time in seconds. The chart compares unmodified, modified internal, and modified external multiple versions, with data table size in millions of records.]
Impact of Record Filtering

![Graph showing the impact of record filtering on elapsed time with different selectivity levels. The graph compares Modified Internal, Modified External Multiple, and Unmodified methods.](image)
References

- Hippocratic databases, Agrawal, 2002
- Limiting disclosure in Hippocratic databases, LeFevre, 2003