Security and Privacy Technology Enablers for Electronic Healthcare

Tyrone Grandison PhD
IBM Healthcare Center of Excellence
Almaden Research Center
San Jose, California
tyron@us.ibm.com

2007 AMIA Spring Congress | May 22-24, 2007
Introduction

- Caveat
  - As medical information move to electronic platforms, policy and social education programs must be augmented by appropriate, corresponding technology\(^1\).

- Objectives
  - Define the addressable.
  - Define the current major problems.
  - Outline technological solutions to each of these problems.

Scope of Current Technical Enablers

- **The Problem Space**
  - Tightly Coupled Complex Systems
  - Each Silo’ed System has its own Protection Mechanisms
  - Conflicting Priorities and Policies
  - New (and changing) Technology

- **Solution Requirements**
  - Reduce the complexity and work-load in integrating and deploying systems, i.e. allow systems to worry about their core function and leverage security and privacy controls in the data system.
  - Do not impact the performance/efficiency of the currently running system
  - Enable the current (clinical) workflow and do not require it to change.
What the Future Holds for Healthcare?

- Bernd Blobel, Head, German National eHealth Competence Center, University of Regensburg Medical Center, Regensburg, Germany
System Requirements – Current & Future

- Appropriate security and privacy services
- Openness
- Flexibility
- Scalability
- Portability
- User acceptance
- Service orientation
- Distribution at Internet level
- Lawfulness
- Based on standards
- Service-oriented interoperability

– Bernd Blobel, Head, German National eHealth Competence Center, University of Regensburg Medical Center, Regensburg, Germany
Current Major Problems

- Policy-based Private Data Management.
  - How does one enforce data disclosure policies and patient preferences?
  - How does one enable privacy-preserving data mining?

- Secure Information Exchange
  - How does one selective share the minimum amount of data necessary for a task?
  - How does one de-identify data for information exchange?

- Efficient Data Access Tracking
  - How do you efficiently track access and disclosure?
  - How do you protect data sent to outsourced agents?
Technology Solutions

- Policy-based Private Data Management.
  - Active Enforcement
  - Privacy-Preserving Data Mining
- Secure Information Exchange
  - Sovereign Information Sharing
  - Optimal $k$-anonymization (de-identification)
- Efficient Data Access Tracking
  - Compliance Auditing
  - Database Watermarking
**Hippocratic Database Active Enforcement**

- **Privacy Policy:** Organizations define a set of policies describing who may access data (users or roles), for what purposes data may be accessed (purposes) and to whom data may be disclosed (recipients).

- **Consent:** Data subjects are given control, through opt-in and opt-out choices, over who may see their data and under what circumstances.

- **Active Enforcement:** Intercepts and rewrites incoming queries to comply with policies, subject choices, and context.

- **Efficiency:** Rewritten queries benefit from all of the optimizations and performance enhancements provided by the underlying engine (e.g. parallelism).

- **Advantages:**
  - Cell-level access and disclosure control.
  - Application modification not required.
  - Database agnostic; does not require changes to the database engine.

---

### Table

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Age</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Adam</td>
<td>25</td>
<td>(111) 111-1111</td>
</tr>
<tr>
<td>3</td>
<td>Bob</td>
<td>-</td>
<td>(333) 333-3333</td>
</tr>
<tr>
<td>4</td>
<td>Daniel</td>
<td>40</td>
<td>-</td>
</tr>
</tbody>
</table>
Privacy-Preserving Data Mining

- Preserves privacy at the individual level, but allows accurate data mining models to be constructed at the aggregate level.
- Adds random noise to individual values to protect data subject privacy.
- EM algorithm estimates original distribution of values given randomized values + randomization function.
- Algorithms for building classification models and discovering association rules on top of privacy-preserved data with only small loss of accuracy.
Sovereign Information Integration

- Autonomous databases for competitive, statutory, or security reasons.
  - Provides selective, minimal sharing on need-to-know basis.
- Example: Which DNA expressions correlate with reactions to certain drugs?
- Algorithms for computing secure joins and join counts without revealing any additional information among the databases.

Minimal Necessary Sharing

<table>
<thead>
<tr>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>u</td>
</tr>
<tr>
<td>u</td>
<td>v</td>
</tr>
<tr>
<td>v</td>
<td>x</td>
</tr>
</tbody>
</table>

- R must not know that S has b & y
- S must not know that R has a & x

Count (R $\nabla S$)

- R & S do not learn anything except that the result is 2.
Optimal $k$-Anonymization

- **Optimal $k$-Anonymization** (Bayardo, Agrawal, 2005)
  - Algorithm finds optimal $k$-anonymizations under two representative cost measures and variations of $k$.

- **Advantages of optimal $k$-anonymization:**
  - **Truthful** - Unlike other disclosure protection techniques that use data scrambling, swapping, or adding noise, all information within a $k$-anonymized dataset is truthful.
  - **Secure** - More secure than other de-identification methods, which may inadvertently reveal confidential information.

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>City</th>
<th>Age</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erica</td>
<td>19 Main Street</td>
<td>San Jose</td>
<td>26</td>
<td>$42,000</td>
</tr>
<tr>
<td>Paul</td>
<td>130 Harry Road</td>
<td>San Jose</td>
<td>42</td>
<td>$88,000</td>
</tr>
<tr>
<td>Mark</td>
<td>4800 17th Street</td>
<td>San Jose</td>
<td>47</td>
<td>$120,000</td>
</tr>
<tr>
<td>Henry</td>
<td>210 Almaden Pkwy</td>
<td>San Jose</td>
<td>28</td>
<td>$50,000</td>
</tr>
</tbody>
</table>

(k=2, on name, address, age)

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>City</th>
<th>Age</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>95131</td>
<td>San Jose</td>
<td>20-29</td>
<td>$42,000</td>
</tr>
<tr>
<td>*</td>
<td>95120</td>
<td>San Jose</td>
<td>40-49</td>
<td>$88,000</td>
</tr>
<tr>
<td>*</td>
<td>95120</td>
<td>San Jose</td>
<td>40-49</td>
<td>$120,000</td>
</tr>
<tr>
<td>*</td>
<td>95131</td>
<td>San Jose</td>
<td>20-29</td>
<td>$50,000</td>
</tr>
</tbody>
</table>
Compliance Auditing

<table>
<thead>
<tr>
<th>ID</th>
<th>Timestamp</th>
<th>Query</th>
<th>User</th>
<th>Purpose</th>
<th>Recipient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2004-02...</td>
<td>Select ...</td>
<td>B. Jones</td>
<td>Marketing</td>
<td>MortgageCo.</td>
</tr>
<tr>
<td>2</td>
<td>2004-02...</td>
<td>Select ...</td>
<td>S. Roberts</td>
<td>Account service</td>
<td>S. Roberts</td>
</tr>
</tbody>
</table>

- **Audits**: Determine whether specified particular data has been accessed in violation of privacy policies or choices.
- **Audit expression**: Auditor specifies the information disclosures that he or she would like to track.
- **Suspicious Queries**: Audit system identifies logged queries that accessed the specified data.
- **Audit Results**: Returns the queries that accessed the specified information and the circumstances of access.
- **Advantages**:
  - Cell-level disclosure auditing.
  - Low storage overhead; reuses existing database infrastructure.
  - Low performance impact; defers computation until audit time.
Watermarking Databases

- Deters data theft and asserts ownership of pirated copies by intentionally introduced pattern in the data.
  - Very unlikely to occur by chance.
  - Hard to find => hard to destroy (robust against malicious attacks).

- Existing watermarking techniques developed for multimedia are not applicable to database tables.
  - Rows in a table are unordered.
  - Rows can be inserted, updated, deleted.
  - Attributes can be added, dropped.

- New algorithm for watermarking database tables.
  - Watermark can be detected using only a subset of the rows and attributes of a table.
  - Robust against updates, incrementally updatable.

---

1. Choose secret key
2. Specify table/attributes to be marked
3. Pseudo randomly select a subset of the rows for marking
   Function of secret key and attribute values

1. Specify secret key
2. Specify table/attributes which should contain marks
3. Identify marked rows/attributes, compare marks with expected mark values
   Requires neither original unmarked data nor the watermark
4. Confirm presence or absence of the watermark
Conclusion

- Technology controls for security and privacy must be used in conjunction with legal policy, organizational requirements and social awareness programs in order to address the current and future problems in medical informatics systems.

- Controls must be moved to the data level in order to:
  - Reduce the complexity in current system.
  - Provide a unified protection framework.
  - Allow the resolution of conflicts at the data level.
  - Scale to future technology without infrastructure modification.

- There is a current set of enablers that would avert breaches and integrate seamlessly into current systems.
Thank You

Questions
Selected References


Slides available at