Data Disclosure Technology for the Grid

Tyrone W A Grandison PhD
Data Disclosure

- The release of information for processing
- Security & privacy concerns critical for Grid
- One way to address this:
  - Hippocratic Databases
GOAL
Create a new generation of information systems that protect the privacy, security, and ownership of data while not impeding the information flow.

Disclosure Control
Database-enforced disclosure control at cell-level of an organization’s data policy and user preferences

Privacy Preserving Data Analytics
Preserve privacy at the individual level, while still building accurate data mining models at the aggregate level

Sovereign Information Sharing
Selective, minimal sharing across database separated due to statutory, competitive or security reasons

Compliance Auditing
Audit if data has been disclosed in violation of the specified policy to help enterprises cope with increasing pace of legislations
What is Hippocratic Database technology?
Technology that facilitates automated non-intrusive, high-performance, fine-grained data disclosure (at the database level).

Functional Components:

- Active Enforcement – enabling the database to reveal only data compliant with policy.
- Eunomia Compliance Auditing – enabling verification and monitoring of compliance with policy (e.g., legislation, privacy, security).
- Sovereign Information Integration – enabling two parties to securely and privately share common information without a third party involved.
- Order Preserving Encryption – enabling the protection of data from medium theft, while allowing encrypted data to be usable.
- Watermarking Databases – deterring data theft and asserting ownership of pirated copies.
- Privacy Preserving Data Mining – preserving privacy at the individual level, while enabling the construction of accurate data mining models at the aggregate level.
- BA k-anonymity – enable data release, while preventing linking attacks and preserving data integrity.
Active Enforcement

- Database-enforced disclosure control at cell-level of an organization’s data policy and user preferences.
- Applications do not require any modification.
- Database agnostic, does not require any change in the database engine.

Implementation intercepts and rewrites incoming queries to factor in policy, user choices, and context (e.g. purpose).
Rewritten queries benefit from all the optimizations and performance enhancements provided by underlying engine (e.g. parallelism).
Demo available for HIPAA enforcement

### Table

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Age</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Adams</td>
<td>25</td>
<td>111-1111</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>333-3333</td>
</tr>
<tr>
<td>4</td>
<td>Daniels</td>
<td>40</td>
<td>-</td>
</tr>
</tbody>
</table>

### Graphs

- **Query Execution Time (seconds)**: Comparison between original and rewritten queries with varying application selectivity.

- **Choice Selectivity**: Comparison between original and rewritten queries with varying choice selectivity.

### Diagram

- **Policy Creation**
- **User Preferences & Data Collection**
- **Application Data Retrieval**
- **Installation Policy Parser**
- **Negotiation User Preferences & Policy Matching**
- **Enforcement SQL Interface**

**DATABASE**

- **Installed Policy**
- **User Data**
Compliance Auditing

- **Problem**: Many laws require companies to account for their handling of sensitive or secured information.

- **How does a company ensure and show that it has complied with such requirements?**
  - Existing database systems and tools provide only rudimentary query logging which is rarely sufficient.
  - Other add-on applications can also log query results, but this has a huge performance impact and still does not reveal certain disclosures of sensitive information.

- **Solution**: An efficient auditing system that tracks disclosures down to the cell level in the database.
  - Allow determining precisely who accessed designated data, for what purpose, when it was accessed, and what changes were made.
  - With minimal impact on the company’s operations.
Eunomia In-House Model

- Query with purpose, recipient
- Updates, inserts, delete
- Database triggers track updates to base tables
- Generate audit record for each query
- Audit if some data has been disclosed in violation of the specified policy
- Audit queries specify what potential data disclosures need monitoring
- Identify logged queries which have accessed the compromised data
- Analyze circumstances of the violation
- Make necessary corrections to procedures, policies, security

<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>Jane</td>
<td>Current</td>
<td>Ours</td>
</tr>
<tr>
<td>2</td>
<td>2004-02…</td>
<td>Select</td>
<td>John</td>
<td>Telemarketing</td>
<td>public</td>
</tr>
</tbody>
</table>
Eunomia Outsourced Model

Company

Almost zero impact on current process

Eunomia Audit Application

Data Reconstruction

DB2

Query Logs

Backlog Tables

Secure communication

Audit reports, alerts

Queries

Recovery log
Separate databases due to statutory, competitive, or security reasons.

- Selective, minimal sharing on need-to-know basis.
- Example: Among those who took a particular drug, how many had adverse reaction and their DNA contains a specific sequence?
  - Researchers must not learn anything beyond counts.
- Algorithms for computing joins and join counts while revealing minimal additional information.

Minimal Necessary Sharing

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

R \( \bowtie \) S

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

Count (R \( \bowtie \) S)

- R & S do not learn anything except that the result is 2.
Privacy Preserving Data Mining

- Insight: Preserve privacy at the individual level, while still building accurate data mining models at the aggregate level.
- Add random noise to individual values to protect privacy.
- EM algorithm to estimate 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.
Conclusion

- Data Disclosure will have to address Security and Privacy concerns for Grid applications
- There is technology that allows this to be done without impeding information flow.
Technical Papers

- R. Agrawal, J. Kiernan, R. Srikant, Y. Xu. Implementing P3P Using Database Technology. 19th Int'l Conf.on Data Engineering(ICDE), Bangalore, India, March 2003.
- R. Agrawal, J. Kiernan, R. Srikant, Y. Xu. Hippocratic Databases. 28th Int'l Conf. on Very Large Databases (VLDB), Hong Kong, August 2002.
- R. Agrawal, J. Kiernan. Watermarking Relational Databases. 28th Int'l Conf. on Very Large Databases (VLDB), Hong Kong, August 2002.
- A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke. Mining Association Rules Over Privacy Preserving Data. 8th Int'l Conf. on Knowledge Discovery in Databases and Data Mining (KDD), Edmonton, Canada, July 2002.

http://www.almaden.ibm.com/software/quest/Publications/ByDate.html