Presentation Outline

1. Inference channel

2. Related work - Overview

3. My research
   (a) Inference channels in MLS relational databases
       • Problem description
       • Data-dependent disclosure
       • Data-independent disclosure
       • Contributions
   (b) Inference channels in semi-structured databases
       • Problem description and contributions
   (c) Inference channels in numeric databases
       • Problem description and contributions

4. Future research
Secure Databases: Constraints, Inference
Channels and Data Disclosure

Dissertation Defense
Csilla Farkas

Dissertation Directors:
Dr. Sushil Jajodia
Dr. Alexander Brodsky

Center for Secure Information Systems and
Department of Information and Software Engineering
George Mason University, Fairfax, VA 22030-4444
Publications

- A. Bródsky, C. Farkas and S. Jajodia:
  Secure Databases: Constraints, Inference Channels and Monitoring Disclosure
  IEEE Trans. Knowledge and Data Eng., Accepted May 1999

- A. Bródsky, C. Farkas and S. Jajodia:
  Data Disclosure and Inference Channels
  Technical Report, George Mason University, 2000

- A. Bródsky, C. Farkas, D. Wijesekara and S.X. Wang:
  Constraints, Inference Channels and Secure Databases
  Sixth International Conference on Principles of Constraint Programming,
  September 18-22, 2000

- A. Bródsky, C. Farkas and S. Jajodia:
  Information Privacy and the Inference Problem
  IEEE Trans. Knowledge and Data Eng., To be submitted
Inference Channel Problem

Inference channel in databases: a means to infer confidential data from non-confidential data and meta-data.

Inference channel problem: detect and remove inference channels.
Example 1: Inference channel via FD (1)

Employee relation:

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>SALARY ($)</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>Clerk</td>
<td>34,000</td>
<td>3</td>
</tr>
<tr>
<td>Brunnel</td>
<td>Clerk</td>
<td>34,000</td>
<td>5</td>
</tr>
<tr>
<td>Hammer</td>
<td>Director</td>
<td>65,000</td>
<td>10</td>
</tr>
<tr>
<td>Smith</td>
<td>Secretary</td>
<td>28,000</td>
<td>5</td>
</tr>
</tbody>
</table>

Functional dependency: RANK $\rightarrow$ SALARY

Confidential information: Salaries of the employees
Example 1: Inference channel via FD (2)

Query 1: “Name and rank of the employees with 3 years of experience.”

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>Clerk</td>
<td>3</td>
</tr>
</tbody>
</table>

Query 2: “Rank and salary of the employees with 5 years of experience.”

<table>
<thead>
<tr>
<th>RANK</th>
<th>SALARY ($)</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clerk</td>
<td>34,000</td>
<td>5</td>
</tr>
<tr>
<td>Secretary</td>
<td>28,000</td>
<td>5</td>
</tr>
</tbody>
</table>

INFERENCE CHANNEL: Brown’s salary is $34,000
Example 2: Inference channel via domain knowledge (1)

Medical Database (MED-DB):

Domain knowledge: *Raloxifene* and *Herceptin* used to treat breast cancer patients

Confidential information: Diagnosis of patients
Example 2: Inference channel via domain knowledge (2)

Prescription info  +  Domain knowledge

Jane Smith

Raloxifene  Herceptin

9/10/99  9/10/99

Herceptin + Raloxifene
→ Breast cancer

INFERENC CHANNEL: Jane Smith has breast cancer
INFERENEC CHANNEL:

NON-CONFIDENTIAL DATA + CONSTRAINTS
(DATABASE DEP.
DOMAIN KNOWLEDGE)
History of Research

1970s and early 1980s: Inference channels in *statistical databases*

1980 - present: Inference channels in *relational databases*

- Inferences via *queries* conditioned on confidential data
- Inferences raised by combining *database dependencies* with non-confidential data

No known research: Inference channels in *semi-structured databases*
Related Work

Database design time inference detection:

- M. Morgenstern (1988)

Query processing time inference detection:

- D.E. Denning (1985)
- B.M. Thuraisingham (1987)
Related Work - Limitations

- Over-classification $\rightarrow$ Reduces data availability
- Limited expressive power $\rightarrow$ Limited application domain
- Framework only $\rightarrow$ Assurance of protection?
My Principal Contributions

- Introduced **characterization** of disclosure inference algorithms by
  - *Completeness* - confidentiality
  - *Soundness* - data availability

- Developed **disclosure inference algorithms** for variety of
  - *Settings*
  - *Constraints*
  - *Operation modes*
Multilevel Secure Relational Databases

- **Database**: data classified with security levels

- **Users**: assigned security clearances

- **Secrecy requirement**: users gain access - directly or indirectly - to only those data for which they have proper clearances
My Research

- **Queries**: $\Pi_Y \sigma_C$

- **Database Constraints**: Horn-clause constraints

- **Security granularity**:
  - (partial) tuples
  - queries
  - their combinations
Conceptual Architecture: Disclosure Monitor

User's query

Mandatory Access Control

Granted query (result)
Additional disclosed information

Disclosure Inference Engine

Database constr. Previous queries (results)

Permission/Refusal of answer
Data Representation (1)

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>SALARY ($)</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>Clerk</td>
<td>34,000</td>
<td>3</td>
</tr>
<tr>
<td>Brunnel</td>
<td>Clerk</td>
<td>34,000</td>
<td>5</td>
</tr>
<tr>
<td>Hammer</td>
<td>Director</td>
<td>65,000</td>
<td>10</td>
</tr>
<tr>
<td>Smith</td>
<td>Secretary</td>
<td>28,000</td>
<td>5</td>
</tr>
</tbody>
</table>

Projection facts:

Employee[NAME=Brown, RANK=Clerk, EXPERIENCE=3]

Employee[RANK=Clerk, SALARY=34,000, EXPERIENCE=5]

Employee[RANK=Secretary, SALARY=28,000, EXPERIENCE=5]
Data Representation (2)

Query-answer pair (QA-pair):

\[
\text{Employee[NAME=Brown, RANK=Clerk, EXPERIENCE=3],} \\
\Pi_{NAME, RANK, EXPERIENCE} \sigma_{EXPERIENCE=3}
\]

\[
\{ \text{Employee[RANK=Clerk, SALARY=34,000, EXPERIENCE=5],} \\
\text{Employee[RANK=Secretary, SALARY=28,000, EXPERIENCE=5]} \} \\
\Pi_{RANK, SALARY, EXPERIENCE} \sigma_{EXPERIENCE=5}
\]
Data-Dependent Disclosure: Example

Previous queries and answers:

- \( \pi_{NAME,RANK,EXPERIENCE} \sigma_{EXPERIENCE=3} \)

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>SALARY ($)</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>Clerk</td>
<td>-</td>
<td>3</td>
</tr>
</tbody>
</table>

- \( \pi_{RANK,SALARY,EXPERIENCE} \sigma_{EXPERIENCE=5} \)

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>SALARY ($)</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>Clerk</td>
<td>34,000</td>
<td>5</td>
</tr>
<tr>
<td>-</td>
<td>Secretary</td>
<td>28,000</td>
<td>5</td>
</tr>
</tbody>
</table>

Disclosed by using FD:

- \( \pi_{NAME,SALARY} \)

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>SALARY ($)</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>-</td>
<td>3&lt;=,000</td>
<td>-</td>
</tr>
</tbody>
</table>
Data-Dependent Disclosure Inference

Let $\mathcal{D}$ be a set of database constraints, $P_1, \ldots, P_n$ be sets of projection facts over attribute sets $X_1, \ldots, X_n$, and $PF$ be a projection fact over $Y$. We say that the set of QA-pairs

$$\mathcal{P} = \{(P_1, \Pi X_1 \sigma_{C_1}), \ldots, (P_n, \Pi X_n \sigma_{C_n})\}$$

data-dependently discloses $(PF, \Pi Y \sigma_C)$, denoted as

$$\mathcal{P} \models_\mathcal{D} (PF, \Pi Y \sigma_C),$$

if for every $r$ over $R$ that satisfies $\mathcal{D}$,

$$P_i \subseteq \Pi X_i \sigma_{C_i}(r) \text{ for all } i = 1, \ldots, n$$

implies

$$PF \in \Pi Y \sigma_C(r).$$
Data-Independent Disclosure: Example

Previous queries:

- $\Pi_{NAME,RANK,EXPERIENCE} \sigma_{EXPERIENCE = 3}$
  
<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>SALARY ($)</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>-</td>
<td>3</td>
</tr>
</tbody>
</table>

- $\Pi_{RANK,SALARY,EXPERIENCE} \sigma_{EXPERIENCE = 5}$
  
<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>SALARY ($)</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>x</td>
<td>x</td>
<td>5</td>
</tr>
</tbody>
</table>

Disclosed by using FD:

- $\Pi_{NAME,SALARY}$
  
<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>SALARY ($)</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>-</td>
<td>x</td>
<td>-</td>
</tr>
</tbody>
</table>
Data-Independent Disclosure Inference: Example

New-Employee relation:

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>SALARY ($)</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>Clerk</td>
<td>34,000</td>
<td>4</td>
</tr>
<tr>
<td>Brunnel</td>
<td>Clerk</td>
<td>34,000</td>
<td>5</td>
</tr>
<tr>
<td>Hammer</td>
<td>Director</td>
<td>65,000</td>
<td>10</td>
</tr>
<tr>
<td>Smith</td>
<td>Secretary</td>
<td>28,000</td>
<td>5</td>
</tr>
</tbody>
</table>

Previous queries:

- \( \Pi_{\text{name}, \text{rank}, \text{experience}} \sigma_{\text{experience}=3} \)

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>SALARY ($)</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

- \( \Pi_{\text{rank}, \text{salary}, \text{experience}} \sigma_{\text{experience}=5} \)

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>SALARY ($)</th>
<th>EXPERIENCE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>Clerk</td>
<td>34,000</td>
<td>5</td>
</tr>
<tr>
<td>-</td>
<td>Secretary</td>
<td>28,000</td>
<td>5</td>
</tr>
</tbody>
</table>
Data-Independent Disclosure Inference

Let \( \mathcal{D} \) be a set of database constraints and \( \Pi_{X_1}\sigma_{C_1}, \ldots, \Pi_{X_n}\sigma_{C_n} \) queries over \( R \). We say that the set of queries

\[
\mathcal{P} = \{ \Pi_{X_1}\sigma_{C_1}, \ldots, \Pi_{X_n}\sigma_{C_n} \}
\]

data-independently (or existentially) discloses the query \( \Pi_Y\sigma_C \) under \( \mathcal{D} \), denoted as \( \mathcal{P} \sim_{\mathcal{D}} \Pi_Y\sigma_C \), if there exist

1. \( r \) over \( R \) that satisfies \( \mathcal{D} \),
2. sets \( P_1 \subseteq \Pi_{X_1}\sigma_{C_1}(r), \ldots, P_n \subseteq \Pi_{X_n}\sigma_{C_n}(r) \), and
3. \( PF \in \Pi_Y\sigma_C(r) \)

such that \( \{ (P_1, \Pi_{X_1}\sigma_{C_1}), \ldots, (P_n, \Pi_{X_n}\sigma_{C_n}) \} \vdash_{\mathcal{D}} (PF, \Pi_Y\sigma_C) \)
Contributions: Assurance

Disclosure Inference Algorithms are evaluated by:

Completeness: the algorithm generates all disclosed information
(no possible inference remains undetected)

Soundness: all generated information is indeed disclosed
(maximal data availability)

Soundness + completeness = security + data availability
Contributions: Data-Dependent Disclosure Inference

- **Classified objects**: (partial) tuples, selection-projection queries or their combinations

- **Decidability result**: data-dependent disclosure is decidable, i.e., given a set $\mathcal{D}$ of database constraints and a set $\mathcal{P}$ of QA-pairs
  - whether $\mathcal{P} \models_{\mathcal{D}} (PF, \Pi Y \sigma C)$
  - whether $\mathcal{P} \models_{\mathcal{D}} S$

- Developed **sound** and **complete** Data-Dependent Disclosure Inference Algorithm
Contributions: Data-Independent Disclosure Inference

- **Classified objects**: selection-projection queries

- **Decidability result**: if neither the queries nor the constraints involve constants then data-independent disclosure is decidable, i.e., given a set of queries $\mathcal{P} = \{ \Pi X_1 \sigma_C, \ldots, \Pi X_n \sigma_C \}$, a set of Horn-clause constraints $\mathcal{D}$, and a query $\Pi Y \sigma_C$

  - whether $\mathcal{P} \leadsto_{\exists \mathcal{D}} \Pi Y \sigma_C$

- **Developed**:

  - **Sound and complete** Constant-free Data-Independent Disclosure Inference Algorithm

  - **Complete** General Data-Independent Disclosure Inference Algorithm
Privacy Information Flow Model

![Diagram of information flow between entities such as Patient, Physician, Health Insurance Company, and Pharmacy, with annotations for individual privacy requirements (IPR) and information flow.]
Contributions: Inference in Semi-Structured Databases

- Privacy Information Flow Model - express privacy requirements
- Privacy Mediator Architecture - enforce the privacy requirements
- Sound and complete Inference Algorithm