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. Brodsky, **C. Farkas** and S. Jajodia:
  Secure Databases: Constraints, Inference Channels and Monitoring Disclosure
  IEEE Trans. Knowledge and Data Eng., Accepted May 1999
- A. Brodsky, **C. Farkas** and S. Jajodia:
  Data Disclosure and Inference Channels
  Technical Report, George Mason University, 2000
- A. Brodsky, **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. Brodsky, **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:

NAME	RANK	SALARY ($)	EXPERIENCE (years)
Brown	Clerk	34,000	3
Brunnel	Clerk	34,000	5
Hammer	Director	65,000	10
Smith	Secretary	28,000	5

Functional dependency: RANK $\longrightarrow$ 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."

NAME	RANK	EXPERIENCE (years)
Brown	Clerk	3

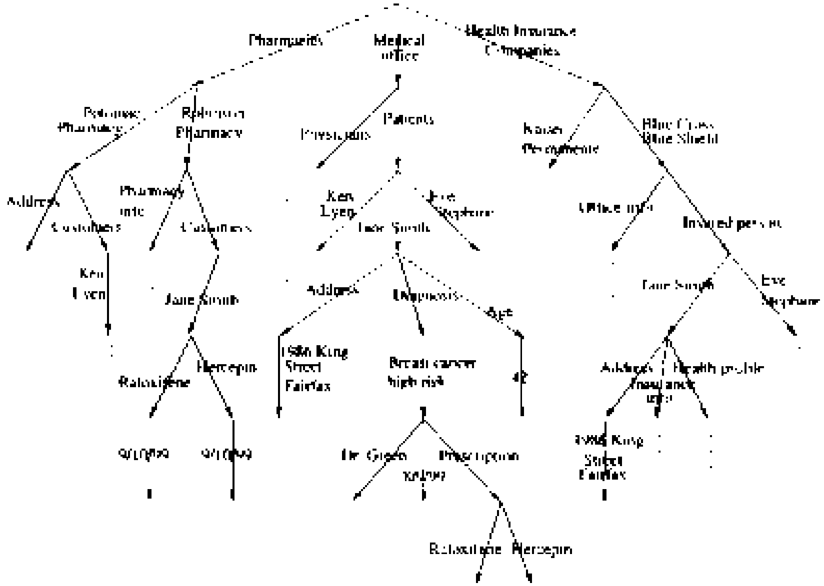
Query 2: "Rank and salary of the employees with 5 years of experience."

RANK	SALARY ($)	EXPERIENCE (years)
Clerk	34,000	5
Secretary	28,000	5

INFERENCE CHANNEL: Brown's salary is $34,000

Example 2: Inference channel via domain knowledge (1)

Medical Database (MED-DB):

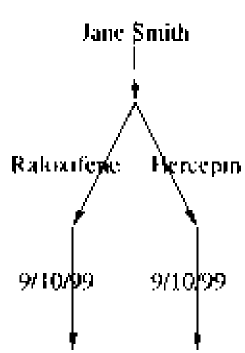


Domain knowledge: *Raloxifene* and *Hercepin* used to treat *breast cancer* patients

Confidential information: *Diagnosis* of patients

Example 2: Inference channel via domain knowledge (2)

Prescription info $+$ Domain knowledge



Hercepin + Ralixofene
$\rightarrow$ Breast cancer

INFERENCE CHANNEL: Jane Smith has breast cancer

INFERENCE 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)
- T. Su and G. Ozsoyoglu (1991)
- T.H. Hinke, H.S. Delugach and A. Chandrasekhar (1995)

Query processing time inference detection:

- D.E. Denning (1985)
- B.M. Thuraisingham (1987)
- S. Mazumdar, D. Stemple and T. Sheard (1988)

Related Work - Limitations

- Over-classification $\longrightarrow$ Reduces data availability
- Limited expressive power $\longrightarrow$ Limited application domain
- Framework only $\longrightarrow$ 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*

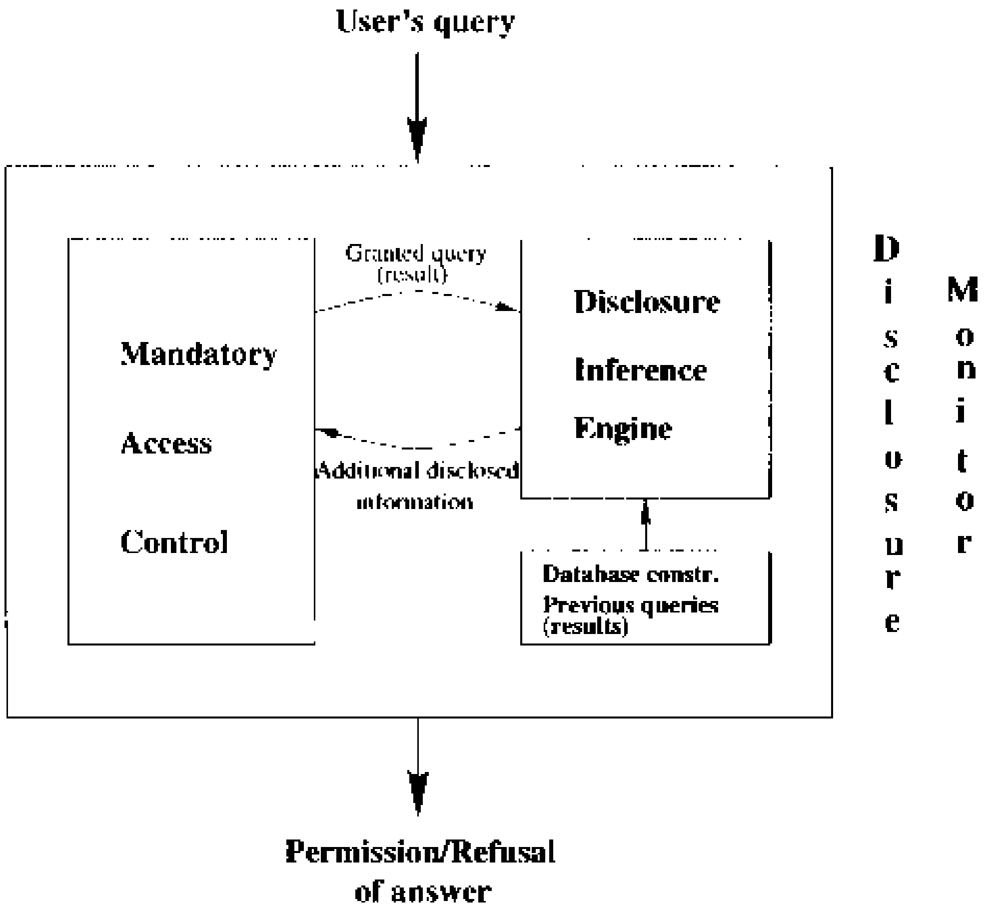
Multilevel Secure Relational Databases

- **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



Data Representation (1)

NAME	RANK	SALARY ($)	EXPERIENCE (years)
Brown	Clerk	34,000	3
Brunnel	Clerk	34,000	5
Hammer	Director	65,000	10
Smith	Secretary	28,000	5

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):

(**Employee**[NAME=Brown,RANK=Clerk, EXPERIENCE=3],
$\Pi_{NAME,RANK,EXPERIENCE}\sigma_{EXPERIENCE=3}$)

({**Employee**[RANK=Clerk,SALARY=34,000, EXPERIENCE=5],
Employee[RANK=Secretary,SALARY=28,000, EXPERIENCE=5] },
$\Pi_{RANK,SALARY,EXPERIENCE}\sigma_{EXPERIENCE=5}$)

Data-Dependent Disclosure: Example

Previous queries and **answers**:

- $\Pi_{NAME,RANK,EXPERIENCE}\sigma_{EXPERIENCE=3}$

NAME	RANK	SALARY ($)	EXPERIENCE (years)
Brown	Clerk	-	3

- $\Pi_{RANK,SALARY,EXPERIENCE}\sigma_{EXPERIENCE=5}$

NAME	RANK	SALARY ($)	EXPERIENCE (years)
-	Clerk	34,000	5
-	Secretary	28,000	5

Disclosed by using **FD**:

- $\Pi_{NAME,SALARY}$

NAME	RANK	SALARY ($)	EXPERIENCE (years)
Brown	-	34,000	-

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* $\mathcal{D}$,

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

implies

$$PF \in \Pi_Y \sigma_C(r)$$

Data-Independent Disclosure: Example

Previous queries:

- $\Pi_{NAME,RANK,EXPERIENCE}\sigma_{EXPERIENCE=3}$

NAME	RANK	SALARY ($)	EXPERIENCE (years)
x	x	-	3

- $\Pi_{RANK,SALARY,EXPERIENCE}\sigma_{EXPERIENCE=5}$

NAME	RANK	SALARY ($)	EXPERIENCE (years)
-	x	x	5

Disclosed by using FD:

- $\Pi_{NAME,SALARY}$

NAME	RANK	SALARY ($)	EXPERIENCE (years)
x	-	x	-

Data-Independent Disclosure Inference: Example

New-Employee relation:

NAME	RANK	SALARY ($)	EXPERIENCE (years)
Brown	Clerk	34,000	4
Brunnel	Clerk	34,000	5
Hammer	Director	65,000	10
Smith	Secretary	28,000	5

Previous queries:

- $\Pi_{NAME,RANK,EXPERIENCE}\sigma_{EXPERIENCE=3}$

NAME	RANK	SALARY ($)	EXPERIENCE (years)
-	-	-	-

- $\Pi_{RANK,SALARY,EXPERIENCE}\sigma_{EXPERIENCE=5}$

NAME	RANK	SALARY ($)	EXPERIENCE (years)
-	Clerk	34,000	5
-	Secretary	28,000	5

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 $\mathcal{P} \rightsquigarrow_{\exists\mathcal{D}} \Pi_Y\sigma_C$, 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 $\{(P_1, \Pi_{X_1}\sigma_{C_1}), \ldots, (P_n, \Pi_{X_n}\sigma_{C_n})\} \models_{\mathcal{D}} (PF, \Pi_Y\sigma_C)$

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

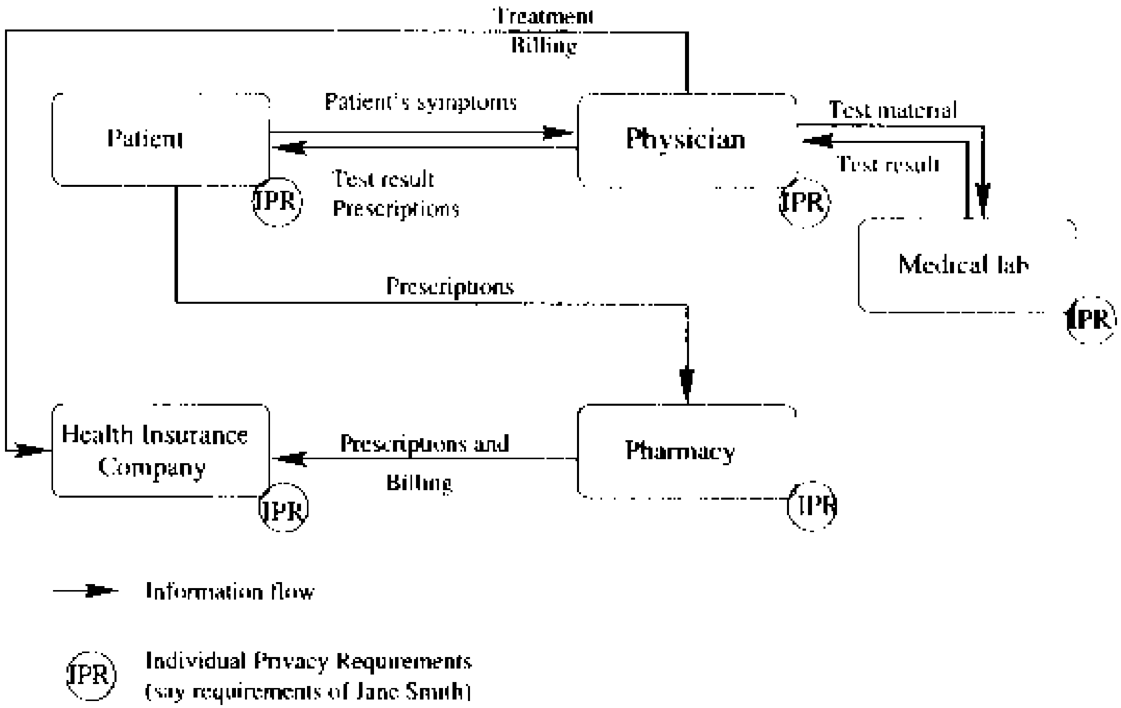
Contributions: Data-Dependent Disclosure Inference

- **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 $\mathcal{P} = \{\Pi_{X_1}\sigma_{C_1}, \ldots, \Pi_{X_n}\sigma_{C_n}\}$, 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



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