technische universität dortmund

Database Fragmentation with Encryption: Under Which Semantic Constraints and A Priori Knowledge Can Two Keep a Secret?

Joachim Biskup **Marcel Preuß**

Information Systems and Security (ISSI)

Technische Universität Dortmund, Germany

July 15, 2013

Fragmentation with Encryption

Context of Our Contribution

Goal of existing approach: Confidentiality by fragmentation

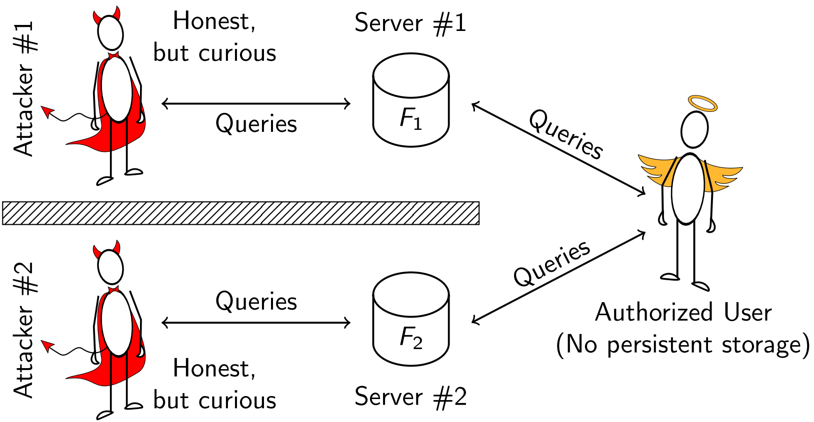
Achievements of this approach

- Formal framework of fragmentation with encryption
- Formal declaration of confidentiality requirements
- Efficient computation of fragmented instances
- Answering queries over fragmented databases

Open problems we solve

- **No formal proof** of "advanced confidentiality"
- Attacker's supposed **a priori knowledge** not considered

Scenario for Working with a Fragmented Database



Fragmentation with Encryption Compliant with Scenario

R	SSN	Name	Illness	HurtBy	Doctor
	1234	Hellmann	Borderline	Hellmann	White
	2345	Dooley	Laceration	McKinley	Warren
	3456	McKinley	Laceration	Dooley	Warren
	3456	McKinley	Concussion	Dooley	Warren

Split columns of r over fragments f_1 and f_2 $\Downarrow$ Add Tuple-IDs to guarantee $f_1 \bowtie f_2 = r$

F_1	<u>tid</u>	SSN	Name	HurtBy	Doctor
	1	e_S^1	Hellmann	e_H^1	White
	2	e_S^2	Dooley	e_H^2	Warren
	3	e_S^3	McKinley	e_H^3	Warren
	4	e_S^4	McKinley	e_H^4	Warren

F_2	<u>tid</u>	SSN	HurtBy	Illness
	1	κ_S^1	κ_H^1	Borderline
	2	κ_S^2	κ_H^2	Laceration
	3	κ_S^3	κ_H^3	Laceration
	4	κ_S^4	κ_H^4	Concussion

"Cleartext attribute": Column in exactly one fragment

"Encrypted attribute": Encrypted values in f_1, crypto-keys in f_2

Hiding Sensitive Values and Associations

R	SSN	Name	Illness	HurtBy	Doctor
	1234	Hellmann	Borderline	Hellmann	White
	2345	Dooley	Laceration	McKinley	Warren
	3456	McKinley	Laceration	Dooley	Warren
	3456	McKinley	Concussion	Dooley	Warren

F_1	tid	SSN	Name	HurtBy	Doctor
	1	e_S^1	Hellmann	e_H^1	White
	2	e_S^2	Dooley	e_H^2	Warren
	3	e_S^3	McKinley	e_H^3	Warren
	4	e_S^4	McKinley	e_H^4	Warren

F_2	tid	SSN	HurtBy	Illness
	1	κ_S^1	κ_H^1	Borderline
	2	κ_S^2	κ_H^2	Laceration
	3	κ_S^3	κ_H^3	Laceration
	4	κ_S^4	κ_H^4	Concussion

fulfills set of confidentiality constraints

$$\mathcal{C} = \{ \quad c_1 = \{\texttt{SSN}\}, \quad c_3 = \{\texttt{Name}, \texttt{HurtBy}\}, \quad c_2 = \{\texttt{Name}, \texttt{Illness}\}, \quad c_4 = \{\texttt{Illness}, \texttt{HurtBy}\} \quad \}$$

technische universität dortmund

Inference-Proofness of Fragmentation

Inference-Proofness under A Priori Knowledge

Notion of **inference-proofness**:

Rational attacker cannot **deduce secret information** from

1. Accessable data
2. His (supposed) a priori knowledge
3. His knowledge about the security mechanism

How to analyze inference-proofness?

- First-order logic modelling of attacker's knowledge
- Formal proof within logic-oriented modelling

Logic-Oriented Modelling of Fragmentation (1)

Suppose: Attacker knows

1. Tuples of outsourced fragment instance f_1
2. Schema $\langle R|A_R|SC_R\rangle$ of original instance r and Knowledge about the world in general
3. Process of fragmentation (algorithm) and Fragment schemas $\langle F_1|A_{F_1}|SC_{F_1}\rangle$ and $\langle F_2|A_{F_2}|SC_{F_2}\rangle$

But: Attacker is curious about hidden original instance r (or hidden instance f_2, respectively)

technische universität dortmund

Logic-Oriented Modelling of Fragmentation (2)

Attacker can infer about r and f_2:

- Cleartext columns of f_1 also valid for r
- Which columns of r and f_2 are hidden from him
  - Columns only stored in r and f_2
  - Encrypted columns of f_1 useless without keys from (hidden) f_2
- Impact of unique Tuple-IDs ...

This knowledge must be modelled as first-order logic sentences!

Logic-Oriented Modelling of Confidentiality Constraints

Confidentiality constraints as potential secrets

- Consider confidentiality constraint $c_i = \{a_{i_1}, \ldots, a_{i_\ell}\}$
- Protect *all* constant combinations possible for $a_{i_1}, \ldots, a_{i_\ell}$
- Leads to first-order formula with free and ∃-quantified variables

Example:

$$c_2 = \{\texttt{Name}, \texttt{Illness}\}$$

$$\downarrow$$

$$\Psi_2((X_N, X_I)) = (\exists X_S)(\exists X_H)(\exists X_D)\, R\,(X_S, X_N, X_I, X_H, X_D)$$

The Impact of A Priori Knowledge: Survey

Until now: Attacker's **a priori knowledge** has been neglected

- Knowledge about semantic database constraints SC_R
- Knowledge about the world in general

Survey of the following results

- No inference-proofness under arbitrary a priori knowledge ↯
- Inference-proofness under constrained a priori knowledge ✓

Goal: Algorithm to construct an inference-proof fragmentation
Complying with attacker's a priori knowledge

Harmful A Priori Knowledge: Example (1)

Attacker's view on r based on f_1:

R	SSN	Name	Illness	HurtBy	Doctor
	?	Hellmann	?	?	White
	?	Dooley	?	?	Warren
	?	McKinley	?	?	Warren
	?	McKinley	?	?	Warren

Suppose attacker knows a priori:
"All patients of psychiatrist `White` suffer from `Borderline`."

As a first-order logic sentence:

$$(\forall X_S)(\forall X_N)(\forall X_I)(\forall X_H)\,[\,R(X_S, X_N, X_I, X_H, \texttt{White}) \Rightarrow (X_I \equiv \texttt{BLine})\,]$$

Attacker's updated view on r violates $c_2 = \{\texttt{Name}, \texttt{Illness}\}$:

R	SSN	Name	Illness	HurtBy	Doctor
	?	Hellmann	Borderline	?	White

Harmful A Priori Knowledge: Example (2)

Attacker's updated view on original instance r:

R	SSN	Name	Illness	HurtBy	Doctor
	?	Hellmann	Borderline	?	White
	?	Dooley	?	?	Warren
	?	McKinley	?	?	Warren
	?	McKinley	?	?	Warren

Suppose attacker knows a priori:
"All patients suffering from `Borderline` have hurt themselves."

As a first-order logic sentence:
$(\forall X_S)(\forall X_N)(\forall X_H)(\forall X_D)\,[\,R(X_S, X_N, \texttt{BLine}, X_H, X_D) \Rightarrow (X_N \equiv X_H)\,]$

Attacker's updated view on r violates $c_3 = \{\texttt{Name}, \texttt{HurtBy}\}$:

R	SSN	Name	Illness	HurtBy	Doctor
	?	Hellmann	Borderline	Hellmann	White

About Harmful Information Flows

Attacker's updated view on r:

R	SSN	Name	Illness	HurtBy	Doctor
	?	Hellmann	Borderline	Hellmann	White

$$(\forall X_S)(\forall X_N)(\forall X_I)(\forall X_H)\,[\,R(X_S, X_N, X_I, X_H, \texttt{White}) \Rightarrow (X_I \equiv \texttt{BLine})\,]$$

- **Harmful constant flow:**
  BLine (constant of formula) $\rightarrow$ Illness (hidden value)
- **Exposed association:** Name $\leftrightarrow$ Illness

$$(\forall X_S)(\forall X_N)(\forall X_H)(\forall X_D)\,[\,R(X_S, X_N, \texttt{BLine}, X_H, X_D) \Rightarrow (X_N \equiv X_H)\,]$$

- **Harmful equality flow:**
  Name (available value of f_1) $\rightarrow$ HurtBy (hidden value)
- **Exposed association:** Name $\leftrightarrow$ HurtBy

technische universität dortmund

Alternative Fragmentation of Example Instance

R	SSN	Name	Illness	HurtBy	Doctor
	1234	Hellmann	Borderline	Hellmann	White
	2345	Dooley	Laceration	McKinley	Warren
	3456	McKinley	Laceration	Dooley	Warren
	3456	McKinley	Concussion	Dooley	Warren

F_1	tid	SSN	Illness	HurtBy	Doctor
	1	e_S^1	Borderline	e_H^1	White
	2	e_S^2	Laceration	e_H^2	Warren
	3	e_S^3	Laceration	e_H^3	Warren
	4	e_S^4	Concussion	e_H^4	Warren

F_2	tid	SSN	HurtBy	Name
	1	κ_S^1	κ_H^1	Hellmann
	2	κ_S^2	κ_H^2	Dooley
	3	κ_S^3	κ_H^3	McKinley
	4	κ_S^4	κ_H^4	McKinley

fulfills set of confidentiality constraints

$$\mathcal{C} = \{ \quad c_1 = \{\mathtt{SSN}\}, \quad c_3 = \{\mathtt{Name}, \mathtt{HurtBy}\}, \quad c_2 = \{\mathtt{Name}, \mathtt{Illness}\}, \quad c_4 = \{\mathtt{Illness}, \mathtt{HurtBy}\} \quad \}$$

A Priori Knowledge under Alternative Fragmentation

Attacker's view on r based on f_1:

R	SSN	Name	Illness	HurtBy	Doctor
	?	?	Borderline	?	White
	?	?	Laceration	?	Warren
	?	?	Laceration	?	Warren
	?	?	Concussion	?	Warren

Suppose attacker knows a priori:

1. $(\forall X_S)(\forall X_N)(\forall X_I)(\forall X_H)\ [R(X_S, X_N, X_I, X_H, \texttt{White}) \Rightarrow (X_I \equiv \texttt{BLine})]$
2. $(\forall X_S)(\forall X_N)(\forall X_H)(\forall X_D)\ [R(X_S, X_N, \texttt{BLine}, X_H, X_D) \Rightarrow (X_N \equiv X_H)]$

A Priori Knowledge is harmless (though premises satisfied)

1. Association `Doctor` $\leftrightarrow$ `Illness` already known from f_1
2. For neither X_N nor X_H a constant is known

Inference-Proofness from Attacker's Point of View

For each (instantiated) potential secret $\Psi(\boldsymbol{v})$:
Existence of alternative instance r' over $\langle R|A_R|SC_R\rangle$ possible

- r' is indistinguishable from original instance r
  - r' and f_1 induce f_2' s.t. r', f_1 and f_2' form a fragmentation
  - r' must satisfy a priori knowledge
- r' does **not** satisfy $\Psi(\boldsymbol{v})$

Construction of Alternative Instance r': Example

Attacker's view on r:

R	SSN	Name	Illness	HurtBy	Doctor
	1234	Hellmann	Borderline	Hellmann	White
	2345	Dooley	Laceration	McKinley	Warren
	3456	McKinley	Laceration	Dooley	Warren
	3456	McKinley	Concussion	Dooley	Warren

`SSN`, `Name`, `HurtBy` are modifiable

Can `Hellmann` ↔ `Borderline` be deduced?

→ Possible alternative view on r:

R	SSN	Name	Illness	HurtBy	Doctor
	9999	Smith	Borderline	Smith	White
	8888	Miller	Laceration	Jones	Warren
	7777	Jones	Laceration	Miller	Warren
	7777	Jones	Concussion	Miller	Warren

Consistent with f_1 and with a priori knowledge

Sufficient Condition for Inference-Proofness

Suppose: A priori knowledge is set of first-order logic sentences
From constrained class of implicational sentences

Theorem: A Fragmentation is inference-proof, if

- Partitioning of r into modifiable and non-modifiable columns
  - Each cleartext-column known from f_1 is non-modifiable
  - Modifiable columns: Subset of columns of f_2
  - Each confident. constraint overlaps with a modifiable column
- A priori knowledge: No information flow...
  - From constants of a priori knowledge to modifiable columns ($\rightarrow$ Eliminates harmful constant flows)
  - Between modifiable and non-modifiable columns ($\rightarrow$ Eliminates harmful equality flows)

Creation of Inference-Proof Fragmentation

About the Creation of Appropriate Fragmentations

Given input:

- Schema $\langle R|A_R|SC_R\rangle$ of original instance
- Set $\mathcal{C}$ of confidentiality constraints
- Attacker's a priori knowledge *prior*

Task: Create an inference-proof fragmentation

- Can be modelled as Binary Integer Linear Program
- Possible goal: Minimize number of "encrypted attributes"
- Wanted fragmentation exists, if solver outputs feasible solution

Conclusion and Future Work

Conclusion and Future Work

Our contribution:

- Extension of existing fragmentation approach by
  - Logic-oriented modelling
  - Attacker's a priori knowledge
- Within modelling: Formal proof of inference-proofness
- Method for computing inference-proof fragmentations

Possible future work:

- Extending feasible a priori knowledge
  $\rightarrow$ Sufficient & necessary condition
- Analysis not relying on perfect encryption algorithm