Inference-Proof Materialized Views

Doctoral Examination

Marcel Preuß

Information Systems and Security (ISSI)

Technische Universität Dortmund, Germany

August 24, 2016
Context of this Work
Inference-Proof Data Publishing

**Nowadays:** Data publishing is ubiquitous
- Governments and companies provide data
- People share data about their private lives

**But:** Original data often contains sensitive (personal) information
- Set up a confidentiality policy
- Release “secure views” instead of original data
  - Do not reveal any confidential information
  - Consider adversary’s abilities to infer information
Framework and Goal

**Framework:** Relational model relying on first-order logic

- Complete original instance \( r \) (definite knowledge: \(+/−\))
- Confidentiality policy \( psec \) of potential secrets
  \( (\exists X) R(X, c) \) s.t. each variable \( X \) occurs only once
- Adversary is aware of policy and protection mechanism

**Goal:** Enforce policy **efficiently** by weakened view on \( r \) s.t.

- Weakened view \( weak(r, psec) \) contains only true knowledge
- Inference-proofness from adversary’s point of view:
  For each \( \Psi \in psec \) there is a “secure” alternative instance \( r^\Psi \)
  - \( r^\Psi \) does **not satisfy** \( \Psi \)
  - \( r^\Psi \) is **indistinguishable** from original instance \( r \)
    \( \rightarrow weak(r^\Psi, psec) = weak(r, psec) \)
Confidentiality by Weakening
Construction of Weakened Views

**Stage 1:** Disjoint disjunction templates \((\text{independent of } r)\)

- Partition the policy \(\text{psec}\) into disjoint clusters \(C_1, \ldots, C_q\) (inducing disjunction templates) of a certain minimum size
- If necessary: Construct additional potential secrets

**Stage 2:** Weakened view \(\text{weak}(r, \text{psec})\) \((\text{dependent on } r)\)

- Keep each tuple of \(r\) not satisfying any \(\Psi \in C_i\)
- Introduce each disjunction \(\bigvee_{\Psi \in C_i} \Psi\) satisfied by \(r\)
- Knowledge not satisfying kept tuples or disjuncts is negative

\(\rightarrow\) Three classes of knowledge: +, ∨, −
Inference-Proofness by Isolation

**Structure** of weakened views:

\[
\begin{align*}
+ & \quad R(c_1), R(c_2), \ldots, R(c_p) \\
& \quad \downarrow R(c_i) \not\models_{DB} \Psi_{j,\ell} \\
\lor & \quad \Psi_{1,1} \lor \ldots \lor \Psi_{1,k_1} \quad \ldots \quad \Psi_{m,1} \lor \ldots \lor \Psi_{m,k_m} \\
& \quad \uparrow \neg R(d_i) \not\models_{DB} \neg \Psi_{j,\ell} \\
- & \quad \neg R(d_1), \neg R(d_2), \neg R(d_3), \ldots
\end{align*}
\]

(definite knowl.)

Hence: For each \(\Psi \in \Psi_{i,1} \lor \ldots \lor \Psi_{i,k_i}\) alternative instance \(r^{\Psi}\) with

\[
\begin{align*}
& \quad r^{\Psi} \not\models_M \Psi \quad \checkmark \\
& \quad r^{\Psi} \models_M +, \lor, - \quad \leadsto \text{indistinguishability by construction of weakened views} \quad \checkmark
\end{align*}
\]

but: \(r^{\Psi} \models_M \Psi_{i,1} \lor \ldots \lor \Psi_{i,k_i}\)
About the Clustering of Policy Elements

Desired properties for disjoint disjunction templates
- Credibility of all disjuncts $\implies$ confidentiality
- Semantically meaningful $\implies$ availability
- Certain length $\implies$ level of confidentiality/availability

Desired properties for disjoint clustering of policy elements
- Consider (high-level) specification of admissible clusters
  $\implies$ Depends on application scenario
- Each cluster must have a certain (minimum) size $k^*$
- Minimize number of additional potential secrets

Clustering problem is NP-hard for $k^* \geq 3$ (Reduction of X3C)
Efficient Clustering for $k^* = 2$ (1)

Model all admissible clusters within simple and undirected Indistinguishability Graph $G = (V, E)$ with:

- $V := \{ \Psi \in \text{psec} \mid \Psi \text{ is to be clustered} \}$
- $E := \{ \{\Psi, \Psi'\} \mid \Psi \lor \Psi' \text{ is admissible} \}$
Efficient Clustering for $k^* = 2$ (2)

Compute **maximum matching** on indistinguishability graph

- Matching: Subset of pairwise vertex-disjoint edges
- Induces set of disjoint and admissible disjunction templates
Efficient Clustering for $k^* = 2$ (3)

How to handle policy elements not covered by the matching?

- Pair with *additional* (artificial) potential secrets
- Minimum number of these due to maximum matching
Inference-Proofness under A Priori Knowledge
Introducing A Priori Knowledge

Usually: Adversary also has some a priori knowledge \( \textit{prior} \)

Challenge for inference-proofness: “secure” alternative instance \( r^{\Psi} \)
- \( r^{\Psi} \) does \textbf{not satisfy} \( \Psi \)
- \( r^{\Psi} \) is \textbf{indistinguishable} from original \( r \)
- \( r^{\Psi} \) satisfies \textit{prior}

Assumed \textit{prior}: “Single Premise TGDs” of the form

\[
\Gamma := (\forall X) \left[ R(X, c_1) \Rightarrow (\exists Y) R(X, Y, c_2) \right] \quad \text{s.t.}
\]
- each \( X \) occurs only once in \( \text{prem}(\Gamma) \) and
- each \( X, Y \) occurs only once in \( \text{concl}(\Gamma) \)
Confidentiality Compromising Dependencies

**Semantics** of Single Premise TGDs: (also via transitive chains)

- Existent DB-Tuple $\Rightarrow$ Existence of other DB-Tuple
- Non-Existent DB-Tuple $\Rightarrow$ Non-Existence of other DB-Tuple

**Broken isolation** in weakened views:

\[
\begin{align*}
+ & \quad R(c_1), R(c_2), \ldots, R(c_p) \quad \text{(definite knowl.)} \\
\lor & \quad \Psi_{1,1} \lor \ldots \lor \Psi_{1,k_1} \ldots \Psi_{m,1} \lor \ldots \lor \Psi_{m,k_m} \\
- & \quad \neg R(d_1), \neg R(d_2), \neg R(d_3), \ldots \quad \text{(definite knowl.)}
\end{align*}
\]
Re-Establishing Sufficient Isolation (1)

Handling of dependency $\Gamma$ interfering with policy elements

- Add policy elements protecting $\text{prem}(\Gamma)$ and $\text{concl}(\Gamma)$
  $\rightarrow$ Do not reveal satisfaction-status of premise or conclusion

- Attention: New policy elements $\rightsquigarrow$ further interferences

Problem: Disjunctions do not always guarantee distortion of non-satisfaction of conclusions

Only escape: Resort to distortion by complete refusal 😞
Re-Establishing Sufficient Isolation (2)

Inference-channel within disjunctive knowledge:

\[ \Psi_1 \lor \Psi_2 \models_{DB} \text{prem} (\Gamma_1)[\sigma_1] \Rightarrow \text{concl} (\Gamma_1)[\sigma_1] \]
\[ \models_{DB} \text{prem} (\Gamma_2)[\sigma_2] \Rightarrow \text{concl} (\Gamma_2)[\sigma_2] \]

How to eliminate this kind of inference-channel?

- Partitioning of prior s.t. \( \Gamma_1 \) and \( \Gamma_2 \) in same partition, if
  - their conclusions imply the same \( \Psi \) (under some \( \sigma_1, \sigma_2 \)) or
  - they can possibly form a transitive chain
- Do not construct disjunction, if all disjuncts imply a premise of the same partition
Conclusion & Future Work
Conclusion & Future Work

Main contributions:

- Confidentiality by cooperative weakening without lies
- Even if adversary employs Single Premise TGDs
- Efficient computation for disjunctions of length $k^* = 2$
- Without prior: Confidentiality level can provably be varied

Possible future work:

- Clustering algorithm for $k^* \geq 3$ ($\rightarrow$ Reasonable heuristic)
- More expressive classes of a priori knowledge
- Proof for different levels of confidentiality under prior
- Model $k$-anonymity/$\ell$-diversity within weakening approach
Backup Slides
Confidentiality by Weakening: Example (1)

Policy: \( \text{psec} = \{ \Psi_1 = R(a, b, c), \ \Psi_2 = R(a, b, d) \} \)

Complete original instance \( r \):

<table>
<thead>
<tr>
<th>+</th>
<th>−</th>
<th>( R(a, b, c) ), ( R(a, c, c) ), ( R(b, a, c) )</th>
</tr>
</thead>
</table>
| \( (a, b, c) \) | \( (a, a, a) \) | \( (\forall X)(\forall Y)(\forall Z) \) [ \( (X \equiv a \land Y \equiv b \land Z \equiv c) \lor \) \( (X \equiv a \land Y \equiv c \land Z \equiv c) \lor \) \( (X \equiv b \land Y \equiv a \land Z \equiv c) \lor \) \( \neg R(X, Y, Z) \) ]
| \( (a, c, c) \) | \( (a, a, b) \) | \                                 |
| \( (b, a, c) \) | :             | \                                 |
| \( (a, b, d) \) | :             | \                                 |
| \( : \) | :             | \                                 |

Obviously: \( r \) satisfies \( \Psi_1 \) (\( \rightarrow \) to be weakened)
Confidentiality by Weakening: Example (2)

Disjunction template: \( \Psi_1 \lor \Psi_2 = R(a, b, c) \lor R(a, b, d) \)

Weakened view \( \text{weak}(r, psec) \):

<table>
<thead>
<tr>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (a, b, c) )</td>
<td>( (a, a, a) )</td>
</tr>
<tr>
<td>( (a, c, c) )</td>
<td>( (a, a, b) )</td>
</tr>
<tr>
<td>( (b, a, c) )</td>
<td>:</td>
</tr>
<tr>
<td>( (a, b, d) )</td>
<td>:</td>
</tr>
</tbody>
</table>

Disjunctive knowledge:

\[
R(a, b, c) \lor R(a, b, d)
\]

Achievement: \( \text{weak}(r, psec) \) does neither imply \( \Psi_1 \) nor \( \Psi_2 \)
Isolation within Disjunctive Knowledge

Policy of only ground atoms: Isolation due to disjoint clustering

But: Existential quantification in policy can break up isolation

- Consider: $\Psi_1 \lor \Psi_2$ with $\Psi_1 \models_{DB} \Psi_2$
- Then: $\Psi_1 \lor \Psi_2 \models_{DB} \Psi_2$ reveals validity of $\Psi_2$
- Also harmful, if $\Psi_1$ and $\Psi_2$ stem from different disjunctions

How to re-establish isolation?

- Only weakest sentences of $psec$ may occur in disjunctions
  - No implication between disjuncts
- Stronger policy elements still implicitly protected
Experimental Evaluation for $k^* = 2$

About the prototype implementation

- Criterion for admissible disjunctions: “Interchangeability”
- “Boost”-library for maximum matchings on general graphs

Lessons learned from 5 experiment setups

- Algorithm efficiently handles input instances of realistic size
- Size and structure of $psec$ and $prior$ crucial for runtime
- Low number of additional potential secrets and refusals
  But: Admissibility criterion should fit to application scenario
- Parallelization: Doubling threads nearly halves runtime
- Clustering is significantly faster with matching heuristic
  → Only slight loss of availability