technische universität dortmund

Inference-Proof Materialized Views

Doctoral Examination

Marcel Preuß

Information Systems and Security (ISSI)

Technische Universität Dortmund, Germany

August 24, 2016

technische universität dortmund

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

technische universität dortmund

Framework and Goal

Framework: Relational model relying on first-order logic

- Complete original instance r (definite knowledge: $+/-$)
- Confidentiality policy *psec* of potential secrets
  $(\exists \boldsymbol{X})\, R(\boldsymbol{X}, \boldsymbol{c})$ 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^{Ψ}
    - r^{Ψ} does **not satisfy** Ψ
    - r^{Ψ} is **indistinguishable** from original instance r
      $\rightarrow weak(r^{\Psi}, psec) = weak(r, psec)$

Confidentiality by Weakening

technische universität dortmund

Construction of Weakened Views

Stage 1: Disjoint disjunction templates *(independent of r)*

- Partition the policy *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 $weak(r, psec)$ *(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

→ Three classes of knowledge: $+$, $\vee$, $-$

Inference-Proofness by Isolation

Structure of weakened views:

$+$ $\boxed{R(\boldsymbol{c}_1),\ R(\boldsymbol{c}_2),\ \ldots,\ R(\boldsymbol{c}_p)}$ (definite knowl.)

↓ $R(\boldsymbol{c}_i) \not\models_{DB} \Psi_{j,\ell}$

$\bigvee$ $\boxed{\Psi_{1,1} \vee \ldots \vee \Psi_{1,k_1} \quad \ldots \quad \Psi_{m,1} \vee \ldots \vee \Psi_{m,k_m}}$ $\Psi_{i,j} \not\models_{DB} \Psi_{\bar{i},\bar{j}}$

↑ $\neg R(\boldsymbol{d}_i) \not\models_{DB} \neg\Psi_{j,\ell}$

$-$ $\boxed{\neg R(\boldsymbol{d}_1),\ \neg R(\boldsymbol{d}_2),\ \neg R(\boldsymbol{d}_3),\ \ldots}$ (definite knowl.)

Hence: For each $\Psi \in \Psi_{i,1} \vee \ldots \vee \Psi_{i,k_i}$ alternative instance r^{Ψ} with

- $r^{\Psi} \not\models_M \Psi$ ✓ (but: $r^{\Psi} \models_M \Psi_{i,1} \vee \ldots \vee \Psi_{i,k_i}$)
- $r^{\Psi} \models_M +, \vee, -$ $\rightsquigarrow$ indistinguishability by construction of weakened views ✓

About the Clustering of Policy Elements

Desired properties for disjoint disjunction templates

- Credibility of all disjuncts $\rightsquigarrow$ confidentiality
- Semantically meaningful $\rightsquigarrow$ availability
- Certain length $\rightsquigarrow$ level of confidentiality/availability

Desired properties for disjoint clustering of policy elements

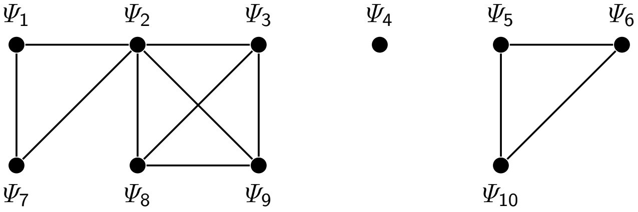
- Consider (high-level) specification of admissible clusters
  $\rightarrow$ Depends on application scenario
- Each cluster must have a certain (minimum) size k^*
- Minimize number of additional potential secrets

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

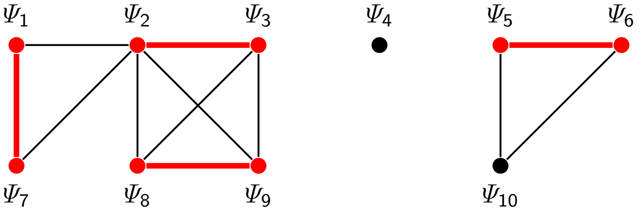
- $V := \{ \Psi \in psec \mid \Psi \text{ is to be clustered} \}$
- $E := \{ \{\Psi, \Psi'\} \mid \Psi \vee \Psi' \text{ is admissible} \}$



Efficient Clustering for $k^* = 2$ (2)

Compute **maximum matching** on indistinguishability graph

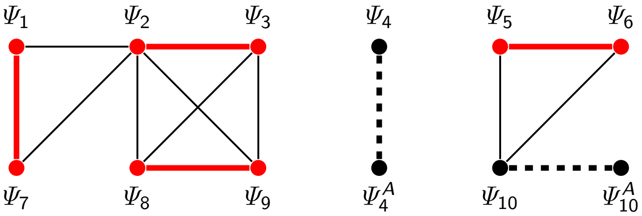
- Matching: Subset of pairwise vertex-disjoint edges
- Induces set of disjoint and admissible disjunction templates



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 *prior*

Challenge for inference-proofness: "secure" alternative instance r^{Ψ}

- r^{Ψ} does **not satisfy** Ψ } (already known)
- r^{Ψ} is **indistinguishable** from original r } (already known)
- r^{Ψ} **satisfies** *prior*

Assumed *prior*: "Single Premise TGDs" of the form

$$\Gamma \ := \ (\forall \boldsymbol{X})\,[\, R(\boldsymbol{X}, \boldsymbol{c}_1) \Rightarrow (\exists \boldsymbol{Y})\, R(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{c}_2)\,] \quad \text{s.t.}$$

- each X occurs only once in $prem(\Gamma)$ and
- each X, Y occurs only once in $concl(\Gamma)$

Confidentiality Compromising Dependencies

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

- Existent DB-Tuple ⇒ Existence of other DB-Tuple
- Non-Existent DB-Tuple ⇒ Non-Existence of other DB-Tuple

Broken isolation in weakened views:

$+$ $R(\boldsymbol{c}_1),\ R(\boldsymbol{c}_2),\ \ldots,\ R(\boldsymbol{c}_p)$ (definite knowl.)

Dependencies

$\vee$ $\Psi_{1,1} \vee \ldots \vee \Psi_{1,k_1} \quad \ldots \quad \Psi_{m,1} \vee \ldots \vee \Psi_{m,k_m}$ Dependencies

Dependencies

$-$ $\neg R(\boldsymbol{d}_1),\ \neg R(\boldsymbol{d}_2),\ \neg R(\boldsymbol{d}_3),\ \ldots$ (definite knowl.)

Re-Establishing Sufficient Isolation (1)

Handling of dependency Γ interfering with policy elements

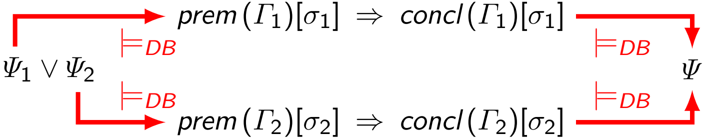
- Add policy elements protecting *prem*(Γ) and *concl*(Γ)
  $\rightarrow$ Do not reveal satisfaction-status of premise or conclusion
- Attention: New policy elements $\rightsquigarrow$ further interferences

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:



How to eliminate this kind of inference-channel?

- Partitioning of *prior* s.t. Γ_1 and Γ_2 in same partition, if
  - their conclusions imply the same Ψ (under some σ_1, σ_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 $k^* \geq 3$ $\quad$ ($\rightarrow$ Reasonable heuristic)
- More expressive classes of a priori knowledge
- Proof for different levels of confidentiality under *prior*
- Model k-anonymity/ℓ-diversity within weakening approach

Backup Slides

technische universität dortmund

Confidentiality by Weakening: Example (1)

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

Complete original instance r:

+	−
(a,b,c)	(a,a,a)
(a,c,c)	(a,a,b)
(b,a,c)	$\vdots$
	(a,b,d)
	$\vdots$

$\Longrightarrow$

$R(a,b,c),\ R(a,c,c),\ R(b,a,c)$

$$(\forall X)(\forall Y)(\forall Z)\ [\ (X \equiv a \wedge Y \equiv b \wedge Z \equiv c) \vee (X \equiv a \wedge Y \equiv c \wedge Z \equiv c) \vee (X \equiv b \wedge Y \equiv a \wedge Z \equiv c) \vee \neg R(X,Y,Z)\ ]$$

Obviously: r satisfies Ψ_1 ($\rightarrow$ to be weakened)

technische universität dortmund

Confidentiality by Weakening: Example (2)

Disjunction template: $\Psi_1 \vee \Psi_2 \;=\; R(a,b,c) \;\vee\; R(a,b,d)$

Weakened view $weak(r, psec)$:

+	−
~~(a,b,c)~~	(a,a,a)
(a,c,c)	(a,a,b)
(b,a,c)	$\vdots$
	~~(a,b,d)~~
	$\vdots$

$\Longrightarrow$

$R(a,c,c),\ R(b,a,c)$

$R(a,b,c) \vee R(a,b,d)$

$$(\forall X)(\forall Y)(\forall Z)\ [\ (X \equiv a \wedge Y \equiv b \wedge Z \equiv c) \vee (X \equiv a \wedge Y \equiv b \wedge Z \equiv d) \vee (X \equiv a \wedge Y \equiv c \wedge Z \equiv c) \vee (X \equiv b \wedge Y \equiv a \wedge Z \equiv c) \vee \neg R(X,Y,Z)\ ]$$

Disjunctive knowledge:
$R(a,b,c) \vee R(a,b,d)$

Achievement: $weak(r, psec)$ does **neither** imply Ψ_1 **nor** Ψ_2

technische universität dortmund

Isolation within Disjunctive Knowledge

Policy of only ground atoms: Isolation due to disjoint clustering

But: Existential quantification in policy can break up isolation

- Consider: $\Psi_1 \vee \Psi_2$ with $\Psi_1 \models_{DB} \Psi_2$
- Then: $\Psi_1 \vee \Psi_2 \models_{DB} \Psi_2$ reveals validity of Ψ_2 ↯
- Also harmful, if Ψ_1 and Ψ_2 stem from different disjunctions

How to re-establish isolation?

- Only weakest sentences of *psec* may occur in disjunctions
  $\rightarrow$ No implication between disjuncts
- Stronger policy elements still implicitly protected

technische universität dortmund

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
  $\rightarrow$ Only slight loss of availability