technische universität dortmund

Inference-Proof Data Publishing by Minimally Weakening a Database Instance

Joachim Biskup **Marcel Preuß**

Information Systems and Security (ISSI)

Technische Universität Dortmund, Germany

December 18, 2014

Context of this Work

Inference-Proof Data Publishing

Nowadays: Data publishing is ubiquitous

- Governments and companies provide data
- People share data about their private lifes

But: Original data often contains sensitive (personal) information

- Set up a confidentiality policy
- Release only "secure views" of original data
  - Do not reveal any information to be protected
  - Consider adversary's abilities to infer information

Framework and Goal

Framework: Relational model relying on first-order logic

- Complete original database instance r
- Confidentiality policy $psec$
  - Each potential secret $\Psi \in psec$ is a ground atom (for now)
  - Adversary is aware of policy and protection mechanism

Goal: Enforce policy by creating weakened instance $weak(r, psec)$

- Replace definite information of r by disjunctions
- 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
    $\rightarrow weak(r^{\Psi}, psec) = weak(r, psec)$

Inference-Proof Weakenings

technische universität dortmund

Case Study 1: Given Setting

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

Original instance r:

$+$	$-$
(a,b,c)	(a,a,a)
(a,c,c)	(a,a,b)
(b,a,c)	(a,a,c)
	$\vdots$

$$
\begin{array}{l}
R(a,b,c),\ R(a,c,c),\ R(b,a,c) \\
(\forall X)(\forall Y)(\forall Z)\ [\\
(X \equiv a \ \wedge\ Y \equiv b \ \wedge\ Z \equiv c) \ \vee \\
(X \equiv a \ \wedge\ Y \equiv c \ \wedge\ Z \equiv c) \ \vee \\
(X \equiv b \ \wedge\ Y \equiv a \ \wedge\ Z \equiv c) \ \vee \\
\neg R(X,Y,Z) \qquad\qquad\qquad]
\end{array}
$$

Obviously: r satisfies Ψ_1 and Ψ_2

Case Study 1: Weakening

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

Weakening $weak(r, psec)$:

+	−
~~(a, b, c)~~	(a,a,a)
~~(a, c, c)~~	(a,a,b)
(b,a,c)	(a,a,c)
	⋮

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

$R(b,a,c)$

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

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

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

Case Study 2: Given Setting

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

Original instance r:

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

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

$$\begin{array}{l}(\forall X)(\forall Y)(\forall Z)\ [\\ (X \equiv a\ \wedge\ Y \equiv b\ \wedge\ Z \equiv c)\ \vee\\ (X \equiv a\ \wedge\ Y \equiv c\ \wedge\ Z \equiv c)\ \vee\\ (X \equiv b\ \wedge\ Y \equiv a\ \wedge\ Z \equiv c)\ \vee\\ \neg R(X,Y,Z)\qquad\qquad\qquad\qquad]\end{array}$$

Obviously: r satisfies Ψ_1, but not Ψ_2

Case Study 2: Weakening

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

Weakening $weak(r, psec)$:

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

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

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

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

$$
\begin{aligned}
&(\forall X)(\forall Y)(\forall Z)\ [\\
&(X \equiv a \wedge Y \equiv b \wedge Z \equiv c)\ \vee\\
&(X \equiv a \wedge Y \equiv b \wedge Z \equiv d)\ \vee\\
&(X \equiv a \wedge Y \equiv c \wedge Z \equiv c)\ \vee\\
&(X \equiv b \wedge Y \equiv a \wedge Z \equiv c)\ \vee\\
&\neg R(X,Y,Z)\ ]
\end{aligned}
$$

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

Case Study 3: The Easy Case

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

Original instance r:

$+$	$-$
(a,b,c)	(a,a,a)
(a,c,c)	(a,a,b)
(b,a,c)	(a,a,c)
	$\vdots$

Nothing to weaken!

Neither Ψ_1 nor Ψ_2 need to be protected.

$\rightarrow\ weak(r, psec) := r$

Obviously: r does **neither** satisfy Ψ_1 **nor** Ψ_2

Clustering of Non-Simple Policies (1)

How to deal with **non-simple policies** of an arbitrary size?

- Partition the policy into a set of disjoint clusters
- For each cluster C: Construct disjunction $\bigvee_{\Psi \in C} \Psi$

How to achieve only **meaningful disjunctions**?

- Declare a set of admissible clusters
  $\rightarrow$ Employ high level languages such as SQL
- Only admissible clusters allowed in final disjoint clustering

Clustering of Non-Simple Policies (2)

How to balance **availability and confidentiality** requirements?

- Size of cluster C
  induces length of disjunction $\bigvee_{\Psi \in C} \Psi$
- Length of disjunction $\bigvee_{\Psi \in C} \Psi$
  induces number of alternative instances
  protecting a policy element of cluster C

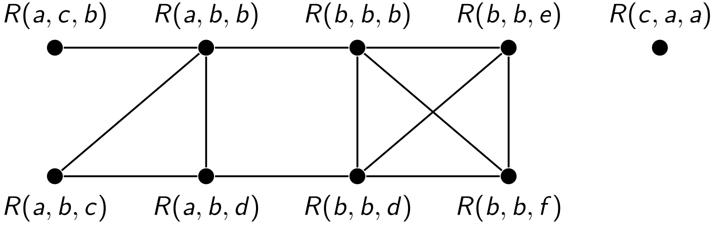
In the following: Goal is to **maximize availability**

- Keep size of clusters as small as possible
- Only one alternative instance per potential secret required
  $\rightarrow$ Clusters of size 2 comply with security definition

Preparing the Clustering Algorithm

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

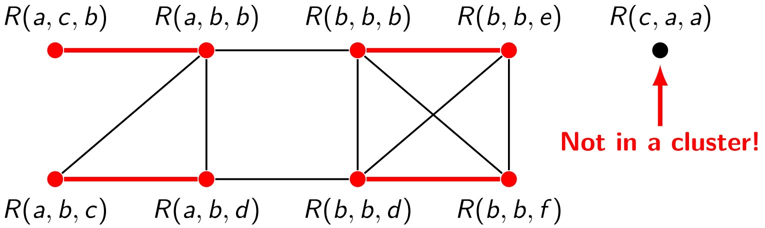
- $V := psec$
- $E := \{\, \{\Psi_1, \Psi_2\} \mid \Psi_1 \vee \Psi_2 \text{ is admissible} \,\}$



First Idea for Clustering Algorithm

Compute **maximum matching** M on indistinguishability-graph G

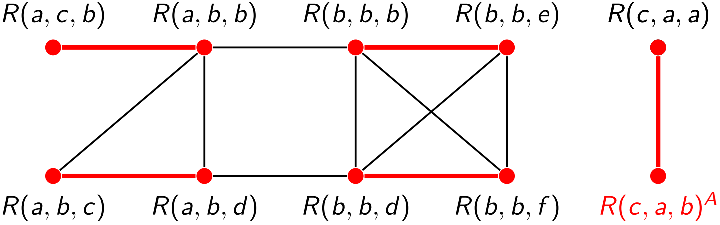
- $M \subseteq E$ is a matching on G, if each pair of different matching edges of M is disjoint
- M is maximum if there is no matching M' with $|M'| > |M|$



Improved Idea for Clustering Algorithm

How to ensure that each policy element is in a cluster?

- Compute a maximum matching M
- For each policy element not covered by M:
  Add **additional** (artificial) potential secret



The Overall Weakening Algorithm

Inputs: Original instance r,
Confidentiality policy *psec*

- **Stage 1:** Clustering of potential secrets (independent of r)
  - Generate indistinguishability-graph $G = (V, E)$ from *psec*
  - Compute maximum matching $M \subseteq E$ on G
  - Construct extended matching M^* based on M
- **Stage 2:** Creation of weakened instance (dependent on r)
  - For each cluster $C \in M^*$: If $\bigvee_{\Psi \in C} \Psi$ is satisfied by r,
    construct disjunction $\bigvee_{\Psi \in C} \Psi$
  - Construct $weak(r, psec)$ (as known from basic case studies)
    $\rightarrow$ Take care of enumeration sequence!

Analysis and Extensions of the Weakening Approach

technische universität dortmund

Sketch of Proof of Inference-Proofness

Consider arbitrary $\Psi \in psec$ of cluster $\{\Psi, \Psi_{ind}\}$

Case 1: Instance r **does not** satisfy $\Psi \vee \Psi_{ind}$

- Construct alternative instance $r^{\Psi} := r$
- r^{Ψ} does not satisfy Ψ (by assumption of this case) ✓
- Obviously: $weak(r^{\Psi}, psec) = weak(r, psec)$ ✓

Case 2: Instance r **does** satisfy $\Psi \vee \Psi_{ind}$

- Construct alternative instance $r^{\Psi} := (r \setminus \{\Psi\}) \cup \{\Psi_{ind}\}$
- Obviously: r^{Ψ} does not satisfy Ψ ✓
- For each cluster: Disjunction satisfied by r^{Ψ} iff satisfied by r
  $\rightsquigarrow weak(r^{\Psi}, psec) = weak(r, psec)$ ✓

Experimental Evaluation of Prototype

Lessons learned from experiments

- Algorithm can handle instances and policies of realistic size
- Runtime of clustering is dominated by matching computation
- Runtime of weakening creation is negligible
- Clustering is significantly faster with matching heuristic
  $\rightarrow$ Slight loss of availability ($\rightarrow$ more unmatched vertices)

technische universität dortmund

Two Extensions Already Considered

Restricted class of **existentially quantified atoms** in policy

- New difficulty: Disjunctions implying confidential knowledge
- Solution: Reduce policy to core of its weakest sentences
  $\rightarrow$ Removed stronger policy elements still implicitly protected

Adversary usually has some **a priori knowledge**

- New difficulty: Alternative instances must satisfy adversary's a priori knowledge to be credible
- Solution (for now): Restrict to ground atoms

technische universität dortmund

Conclusion & Future Work

technische universität dortmund

Conclusion & Future Work

Our contribution:

- Approach creating inference-proof materialized views
- Therefore: Replace some definite information by disjunctions
- Efficient computation (by limiting expressiveness)

Possible future work:

- Employ common database constraints as a priori knowledge
  $\rightarrow$ Equality/Tuple Generating Dependencies
- Guarantee a certain number of $k > 2$ different "secure" alternative instances for each potential secret
- Elaborate connection to k-anonymity/ℓ-diversity