

# Exposing Minimal Data Sets to Rule-Based Data Analyzers

N. ANCIAUX<sup>1</sup>   B. NGUYEN<sup>1,2</sup>   M. VAZIRGIANNIS<sup>3</sup>

<sup>1</sup>Projet SMIS  
INRIA

<sup>2</sup>PRiSM Laboratory  
University of Versailles St-Quentin

<sup>3</sup>Athens University of Economics and Business

7eme Journée Optimeo, Versailles, 7 juin 2011

# Outline I

- 1 Introduction
  - Context
  - Privacy Principles
  - Contributions
- 2 Related Works
- 3 Model and Scenario
  - Scenario
  - Model
  - Example
- 4 Formalization
  - Hypergraph Model
  - Hardness Proofs
- 5 Heuristic and Experimentation
  - The *ME* Algorithm
  - Example
  - Experimentation

# Context I

## Official Digital Information Storage and Use

### Increasing e-commerce and e-gov

- Individuals receive more and more official documents in digital format (employers, banks, insurances, civil authorities, hospitals, schools, services, etc.)
- These documents are digitally signed with the providers certificate (authenticity and origin).
- Legal data retention policies impose to keep these documents for long periods, if not indefinitely.

# Context II

## Official Digital Information Storage and Use

### Personal document management

The management of such personal documents has become a business. Several solutions exist :

- Encrypted Server Side storage : Digicoffre, Adminium, Securibox, etc.
- Hardware protected portable device storage : PDS, Nori, Personal Data Ecosystem, etc.

# Context III

## Official Digital Information Storage and Use

### Increasing data requests

- Personal (and signed) information is requested by third parties (banks, insurance, governmental services) to determine the particular benefits a given individual can be offered, given specific circumstances.
- Such requests can be processed automatically, if appropriate and certified information is provided by users (online loans, revenue tax, social services benefits, etc.).

The information *requested* is not necessarily the information *needed*.

For instance, when asking for a loan, a bank may want to check whether a person has sufficient income and is young enough, *OR* may simply want to ask if the person has sufficient assets.

# Information Disclosure

## The risk

### Inevitable privacy disclosure risk

There is far too much risk in sharing *a lot of or all* one's information with many different third parties. No server is immune to an attack (See [IBM]), and to a privacy breach.

### Our approach

We do not study how to secure a server: we study what information can be stored to limit the impact of a disclosure if (when) it takes place.

# Limited Collection

## Principle and Paradox

### The *Limited Collection* Principle

In order to control information leakage, many countries[Dir95, fECOD80, GAO89]adopt the *Limited Collection* approach for data management : a process must only collect (and store) data necessary to complete its purpose.

### The *Limited Collection* Paradox

In order to limit the collection of personal information pertaining to a given individual with a view to achieve a given process, servers must collect personal information to determine which information is useful or not for the treatment.

### Example

If a bank wants to know if a person is *either* young enough and with a good enough salary *or* has sufficient assets, it will collect *all* this information, and only use what is necessary.

# Minimal Exposure

A reverse implementation of Limited Collection

## Turning Limited Collection around

Our contribution is to propose a reverse implementation of Limited Collection, by giving enough knowledge to individuals so that they decide *themselves* which information to share, and with what benefits.

## The Minimal Exposure Problem

Given a set of (signed) personal data, and a set of benefits that are obtained using this data, solving the Minimal Exposure (*ME*) problem means finding the *minimal* set of data to disclose to obtain all the benefits.

Note that minimality can be computed in many different ways. We study the issue assuming that we can compute the benefit to privacy by removing any element independently.



# Minimal Exposure

## Our contribution

### Contribution

- Modelization of the *ME* problem in the context of decision trees
- Complexity analysis
- Heuristics and experimentation

# Limited Retention in Law

- In most countries Limited Retention is considered by law as a fundamental privacy right. The process follows several steps :
  - IS must only collect the data it needs for a given task (Limited Collection problem)
  - IS must destroy any data that becomes useless (Data Destruction problem)
  - Users can selectively destroy data (erroneous, compromising or highly sensitive) (Selective Deletion problem)
- European and French laws [Leg78, Dir95, War05]
- OECD[fECoD80] and UN[GAO89]

# Limited Retention in Database Systems

- Hypocratic DB [AKSX02]
- Privacy in pervasive computing [KRB09]
- Privacy Preserving Data Mining : This is different from *LR*.  
Objective is to protect all the raw data (via perturbation, deletion, k-anon etc. → trade vs. utility) or to protect the results of the algorithm (i.e. hide some classes, but not raw data → minimize global loss of information). Neither of these aspects correspond to law.
- Our objective is to get rid of maximal useless (raw) data (and be “legal”).

# Overview I

## Actors

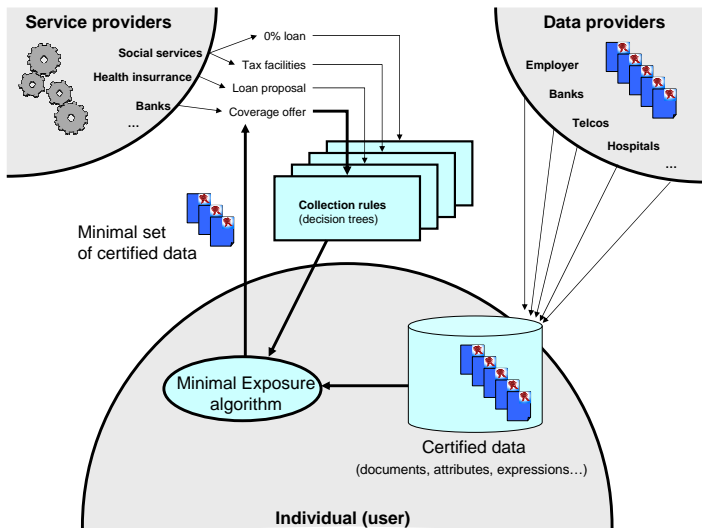
### Scenario

The scenario involves three actors :

- The Individual (user) that owns and stores his personal data.
- Data Provider (state, companies, employers, etc.) that produce the data sent to the Individual.
- Service Providers (banks, social services, health insurance, etc.) that provide services to the Individual, if the individual can exhibit specific “official” data.

# Overview II

## Diagram



# Model I

## Data Model

All the data produced, stored and analyzed is assumed to be a set of *attribute* $\theta$  *value* triples, where  $attribute \in [a_1, \dots, a_n]$ ,  $\theta \in \{=, >, <, \geq, \leq, \neq\}$  and  $value \in dom(attribute) \cup \emptyset$ . To simplify the rest of the discussion, except stated otherwise, we note  $f_i$  one of these triples, and  $\mathcal{F} = \{f_i\}$  (facts).

A user  $u$  can provide data  $D_u$  such that a certain number of  $f_i$  are *true*.

## Privacy Model : Exposure metric

We use a simple cardinality metric to compute the exposure (privacy) of a given set of data. The more triples sent to the service provider, the greater the exposure. Note that results are similar on any metric that can be computed for each value independently.

# Model II

## Classifier : Decision tree

We consider a set of  $p$  conjunctive rules, each leading to a certain class  $C_j$ . We use a decision trees to model a rule based classifier : applied on some individuals it gives a YES/NO answer for a given benefit (class). A decision tree is modeled as a logical formula in DNF : 1 decision tree is simply the disjunction of all the rules leading to the same class. Finaly, since there are multiple classes (benefits) to which an individual can belong, we use a forest of decision trees.

## Running Example

We use the running example of an individual who wants to subscribe to a bank loan. In this first example, an individual will disclose either his attribute value or not.

$inc > 30.000 \wedge ass > 50.000 \rightarrow HigherLoan$

$col > 200.000 \wedge PR = NO \rightarrow HigherLoan$

$TR > 10\% \wedge mar = TRUE \wedge ch > 0 \rightarrow 0\%Rate$

$PR = NO \wedge edu = Univ \wedge age < 30 \rightarrow 0\%Rate$

$inc > 30.000 \wedge mar = TRUE \wedge ch > 0 \rightarrow LongerLoan$

$ass > 50.000 \wedge PR = NO \wedge claim < 5000 \rightarrow LongerLoan$

$ass > 50.000 \wedge TR > 10\% \wedge mar = TRUE \wedge ch > 0 \rightarrow LowerInsurance$

$inc > 30.000 \wedge PR = NO \wedge edu = Univ \wedge age < 30 \rightarrow LowerInsurance$



# Running Example

The example is summarized by the 4 decision trees :

$$(f_1 \wedge f_2) \vee (f_3 \wedge f_4) \rightarrow C_1$$

$$(f_5 \wedge f_6 \wedge f_7) \vee (f_4 \wedge f_8 \wedge f_9) \rightarrow C_2$$

$$(f_1 \wedge f_6 \wedge f_7) \vee (f_2 \wedge f_4 \wedge f_{10}) \rightarrow C_3$$

$$(f_2 \wedge f_5 \wedge f_6 \wedge f_7) \vee (f_1 \wedge f_4 \wedge f_8 \wedge f_9) \rightarrow C_4$$

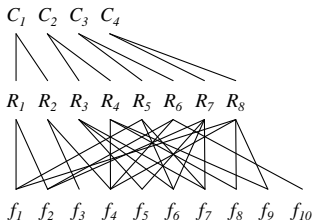
# ME-Hypergraph Model I

We use a labeled multi-hyper(pseudo)graph (i.e. with loops) model to capture the *ME* problem.

- labels appear when we consider that several different antecedants in the rules can lead to the same class.
- Pseudo in the case a single fact is an antecedent,
- multi if there are several similar sets that are the antecedents of different classes.

# ME-Hypergraph Model II

$$\begin{array}{l}
 \underbrace{(f_1 \wedge f_2)}_{R_1} \vee \underbrace{(f_3 \wedge f_4)}_{R_2} \rightarrow C_1 \\
 \underbrace{(f_5 \wedge f_6 \wedge f_7)}_{R_3} \vee \underbrace{(f_4 \wedge f_8 \wedge f_9)}_{R_4} \rightarrow C_2 \\
 \underbrace{(f_1 \wedge f_6 \wedge f_7)}_{R_5} \vee \underbrace{(f_2 \wedge f_4 \wedge f_{10})}_{R_6} \rightarrow C_3 \\
 \underbrace{(f_2 \wedge f_5 \wedge f_6 \wedge f_7)}_{R_7} \vee \underbrace{(f_1 \wedge f_4 \wedge f_8 \wedge f_9)}_{R_8} \rightarrow C_4
 \end{array}$$



# Hypergraph Modelisation of the problem

## Model

We consider for a given user  $u$  the edge-labelled hypergraph  $H_{ME}(V, E, L)$  where  $V = \{v_i\} = f_i \subset \mathcal{F}$  represent the vertices and  $E = [e_j]$  the hyperedges of  $H_{ME}$  such that *all the antecedents of a given rule  $r_j$  are incident to  $e_j$  and we label  $e_j$  with the consequent  $C_\lambda$  of rule  $r_j$* . We note  $L$  the edge labelling function. Note that if two edges connect the same vertices, then their labels will be different.

## Parameters used for hypergraph generation

- $F$  the set of facts,  $R$  the set of rules,  $C$  the set of classes,  $E_F$  the set of edges between  $F$  and  $R$ ,  $E_C$  the set of edges between  $C$  and  $R$ .  $(F \cup C, R, E_F \cup E_C)$  is a bipartite graph.
- $d_F, d_C$  the average out-degree of nodes  $F$  and  $C$
- $d_{RF}, d_{RC}$  the average out-degree of nodes  $R$  to each subset of the partition  $F \cup C$ . Note that we consider that  $d_{RC} = 1$ .

# Minimality problem

## Problem

**Hypergraph Minimal Exposure Problem :** Given  $H_{ME}$  a labeled muti-hyper(pseudo)graph, find

$$H_{ME}^{min}(V_m, E_m, L) \subset H_{ME}$$

such that :

- 1  $V_m \subset V = f_i \subset \mathcal{F}$
- 2  $L(E_m) = L(E)$
- 3  $\|V_m\|$  is minimal.

## Hardness

The Hypergraph Minimal Exposure Problem is NP-Hard.

# Decision problem

## Problem

**Hypergraph  $n$ -Exposure Decision Problem** : Given  $H_{ME}$  a labeled multi-hyper-(pseudo)-graph,  $n \in \mathbb{N}$  does there exist

$$H_{ME}^{min}(V_m, E_m, L) \subset H_{ME}$$

such that :

- 1  $V_m \subset V = \mathcal{F}$
- 2  $L(E_m) = L(E)$
- 3  $\|V_m\| \leq n.$

## Hardness

The Hypergraph  $n$ -Exposure Decision Problem is NP-Complete.

# Hardness Proofs

## ME-PL Definitions

### Definitions

- Let  $A$  represent a set of boolean attributes  $[a_1, \dots, a_n]$
- Let  $D_u(A) \in \{0, 1\}^n$  represent a truth assignment to attributes  $A$  for user  $u$ . We note  $D_u(a_i) = \{0, 1\}$ .
- Let  $Exposure(D_u(A)) = \sum_i(D_u(a_i))$ .
- Let  $W$  represent a forest of decision trees all satisfied by the assignment  $D_u(A)$ .  $W = \bigwedge_i(\bigvee_j(\bigwedge_k P_{i,j,k}))$  where  $P_{i,j,k} \in [a_1, a_n]$  and the truth value of  $P_{i,k,j}$  is  $D_u(P_{i,j,k})$ .

# Hardness Proofs

## ME-PL Problem Statement

### Minimal Exposure Problem

Computing the Minimal Exposure of  $D_u(A)$  means finding  $D'_u(A) \in \{0, 1\}^n$  such that  $\forall a_i, D'_u(a_i) = 1 \rightarrow D_u = 1$  and  $Exposure(D'_u)$  is minimal.

### $n$ -Exposure Problem

$D_u(A)$  is  $n$ -Exposable  $\Leftrightarrow \exists D'_u(A) \in \{0, 1\}^n$  such that  $\forall a_i, D'_u(a_i) = 1 \rightarrow D_u = 1$  and  $Exposure(D'_u) \leq n$ .



# Hardness Proofs

Reduction from the Min Weighted CNF Satisfiability Problem MIN WSAT[CM80]

## Max Weighted Satisfiability

Given clauses  $C_1, \dots, C_m$  on  $n$  variables  $x_1, \dots, x_n$  with non-negative weights  $w(x)$ , we wish to compute a truth assignment that both satisfies the clauses and minimizes the sum of the weights of the variables set to 1.

## Approximation [MPT03]

MIN WSAT belongs to 0-DAPX.

## Reduction

- $n$ -Exposure is clearly in NP.
- MAX WSAT is a special (simpler) case of  $n$ -Exposure where there are  $n$  classes but where *every* rule is a singleton.

# The *ME* Algorithm

---

**Algorithm 1** The *ME* algorithm

*Input* :  $f_i$  a set of true facts,  $R$  a set of rules

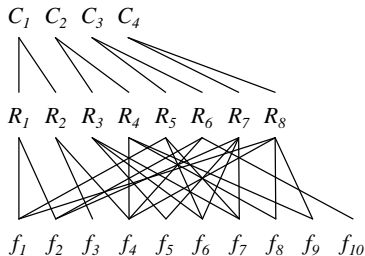
*Output* :  $Res \subset f_i$  such that  $R(Res) = R(f_i)$

---

- 1:  $S \leftarrow f_i$
  - 2:  $Res \leftarrow \emptyset$
  - 3: **while** a fact can be removed from  $S$  and  $R(S) = R(f_i)$  **do**
  - 4:   **for all** fact  $f_i$  that can be removed from  $S$  **do**
  - 5:     Compute  $nb_i$  the number of facts that need to be kept to avoid reducing  $R(S)$
  - 6:   **end for**
  - 7:   Remove fact  $f_i$  from  $S$  such that  $nb_i$  is minimal
  - 8:   Compute the set of facts that must be kept, and add it to  $Res$
  - 9: **end while**
-

# Example

step	$f_1$	$f_2$	$f_3$	$f_4$	$f_5$	$f_6$	$f_7$	$f_8$	$f_9$	$f_{10}$
1	7	7	2	5	4	6	6	4	4	3
2	$\infty$	$\infty$	-	5	5	6	6	5	5	4
3	$\infty$	$\infty$	-	5	7	$\infty$	$\infty$	5	5	-
4	$\infty$	$\infty$	-	-	$\infty$	$\infty$	$\infty$	5	5	-
5	$\infty$	$\infty$	-	-	$\infty$	$\infty$	$\infty$	-	5	-
Min. Exposure	$\infty$	$\infty$	-	-	$\infty$	$\infty$	$\infty$	-	-	-



# Hypergraph Generator

## Parameters

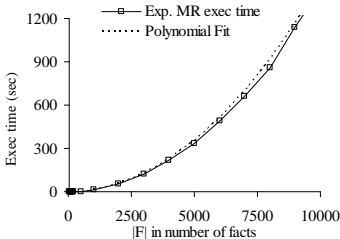
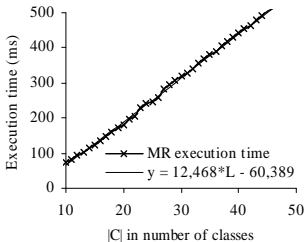
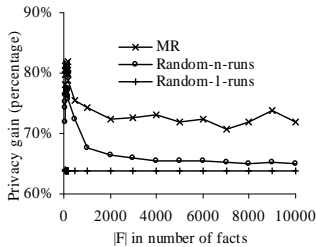
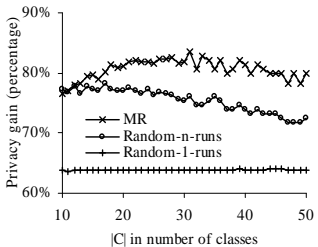
$$|E_F| = \sum_{f_i} d_{f_i} = |F| \times d_F = |R| \times d_{RF}$$

$$|E_C| = \sum_{C_i} d_{C_i} = |C| \times d_C = |R| \times d_{RC} = |R|$$

### Hypergraph Generator

- We use a quadruplet as generator for a hypergraph :  
 $(|F|, d_F, |C|, d_C)$ . In our example :  
 $|C| = 4, |F| = 10, d_F = 2.4, d_C = 2$  which can be used to compute  $|R| = 8, d_{RF} = 3$ .
- On real data (Association Rules on market baskets), we found :  
 $|F| = 134, d_F = 2, |C| = 20, d_C = 6, |R| = 120, d_{RF} = 2.2$ .

# Performance analysis vs. baseline random algorithm



# Conclusion and Future Work

## Expressivity

Although not presented here, our model already captures data anonymization through generalization (e.g. do not send the exact data value if it is not needed).

## Performance

Compare with heuristics for an existing similar problem (MIN WSAT or Vertex Cover).

## Data Mining Extensions

- Produce minimal data sets with regards to other types of data mining operations (Association Rule Mining).
- Introduce data mining algorithms into the privacy metric, and extend the analysis to metrics that can not be computed in isolation.

# References I




-  R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, *Hippocratic databases*, Proceedings of the 28th international conference on Very Large Data Bases, VLDB Endowment, 2002, p. 154.
-  P. Camerini and F. Maffioli, *Weighted satisfiability problems and some implications*, Optimization Techniques (K. Iracki, K. Malanowski, and S. Walukiewicz, eds.), Lecture Notes in Control and Information Sciences, vol. 23, Springer Berlin / Heidelberg, 1980, 10.1007/BFb0006601, pp. 170–175.
-  EU Directive, *95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of such Data*, Official Journal of the EC **23** (1995).

## References II

-  Organisation for Economic Co-operation and Development, *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, Organisation for Economic Co-operation and Development; Washington, DC: OECD Publications and Information Center, september 1980.
-  UN GAOR, *Guidelines for the regulation of computerized personal data files*, GA Res. 44th Sess., Supp. No. 49, UN Doc. A/44/49 at 211, 1989.
-  IBM, *Force Threat Reports. IBM Internet Security Systems X-Force, trend and risk report (2009)*.
-  S. Kurkovsky, O. Rivera, and J. Bhalodi, *Classification of Privacy Management Techniques in Pervasive Computing*, International Journal of Ubiquitous Service, Science and Technology **1** (2009), no. 1.



## References III

-  Legifrance, *Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*, 1978.
-  J. Monnot, V.T. Paschos, and S. Toulouse, *Approximation polynomiale des problèmes np-difficiles: optima locaux et rapport différentiel*, Hermès science publications, 2003.
-  J. Warner, *Right to Oblivion: Data Retention from Canada to Europe in Three Backward Steps*, The U. Ottawa Law & Technology Journal **2** (2005), 75.