Secure Data Outsourcing with Adversarial Data Dependency Constraints

BigDataSecurity 2016

Boxiang Dong Wendy Hui Wang Jie Yang

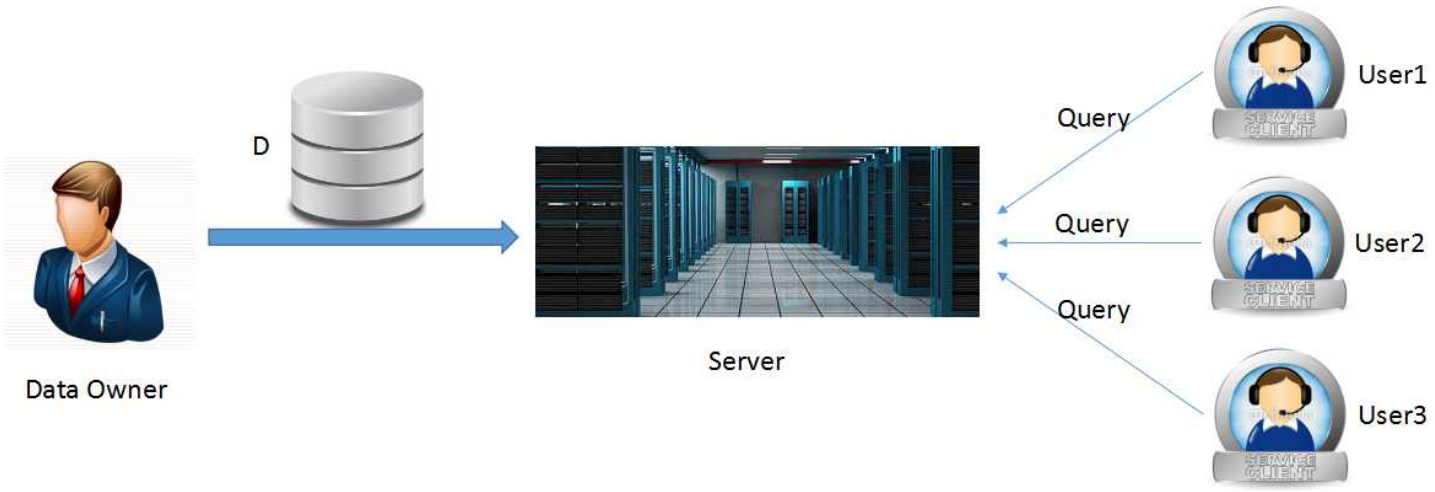
Department of Computer Science
Stevens Institute of Technology

School of Software Engineering
South China University of Technology

April 9, 2016

Database-as-a-Service (DaS)

Database as a Service:

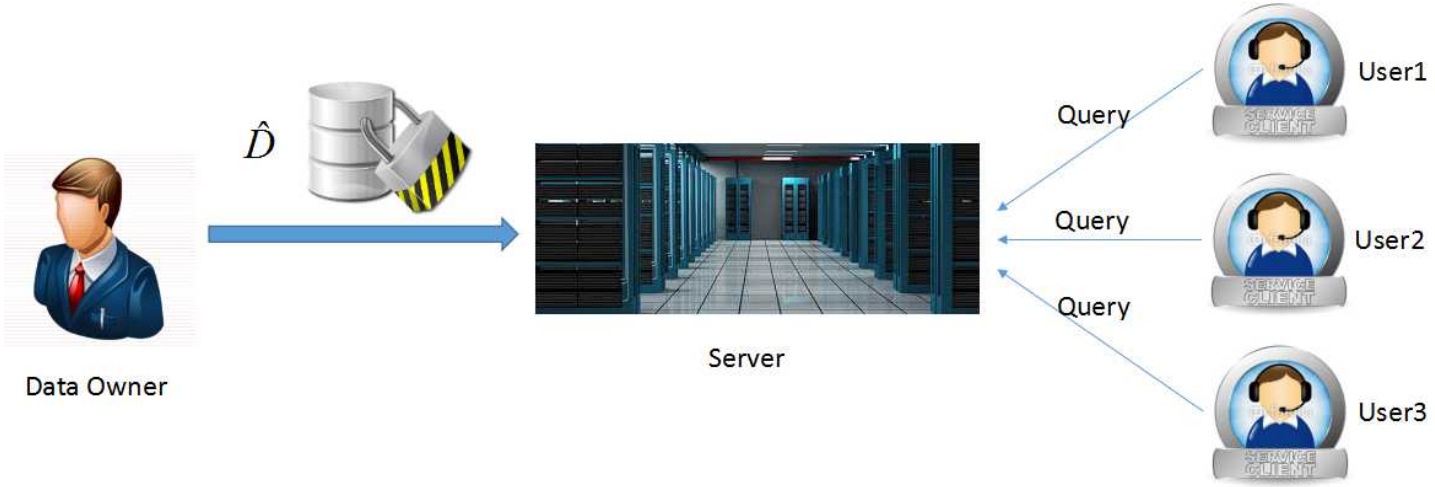


- Weak data owner
- Computationally powerful service provider (e.g. cloud)
- DaS enables the data owner to outsource the database services to a third party server.

Data Security Issue

Security The outsourced data may contain important and sensitive information.

Solution The data owner encrypts the data before outsourcing.



Security Constraint

Security constraint $\Pi_Y \sigma_C$

- Y is a set of attributes.
- C is a conjunction of equalities of $A = B$ or $A = a$.

Basic encryption $\bar{D}$ Encrypt the sensitive values specified by the security constraint.

NM	SEX	AGE	DC	DS
Alice	F	53	CPD5	HIV
Carol	F	30	VPI8	Cancer
Ela	F	24	VPI8	Cancer

(a) The original dataset D

$S_1 : \Pi_{DS} \sigma_{NM='Alice'}$

$S_2 : \Pi_{DS} \sigma_{NM='Ela'}$

NM	SEX	AGE	DC	DS
Alice	F	53	CPD5	α
Carol	F	30	VPI8	Cancer
Ela	F	24	VPI8	γ

(b) The basic encryption $\bar{D}$

: sensitive data

FD Attack

Functional dependency (FD) $X \rightarrow Y$ if $r_1[X] = r_2[X]$, then $r_1[Y] = r_2[Y]$.

FD attack Infer the encrypted sensitive value based on the FD.

NM	SEX	AGE	DC	DS
Alice	F	53	CPD5	HIV
Carol	F	30	VPI8	Cancer
Ela	F	24	VPI8	Cancer

(a) The original dataset D

$FD : DC \rightarrow DS$

$S_1 : \Pi_{DS}\sigma_{NM='Alice'}$

$S_2 : \Pi_{DS}\sigma_{NM='Ela'}$

NM	SEX	AGE	DC	DS
Alice	F	53	CPD5	α
Carol	F	30	VPI8	Cancer
Ela	F	24	VPI8	γ

(b) The unsafe basic encryption $\bar{D}$

: sensitive data

: inference channel

Naive Solutions

1. Encrypt all the data values.

NM	SEX	AGE	DC	DS
β	δ	ϵ	ζ	α
η	δ	θ	ι	γ
λ	δ	μ	ι	γ

(encryption overhead: 13)

2. Encrypt all values of the attributes that involve a FD.

NM	SEX	AGE	DC	DS
Alice	F	53	ζ	α
Carol	F	30	ι	γ
Ela	F	24	ι	γ

(encryption overhead: 4)

: sensitive data
: additional encrypted data

Encryption Overhead Amount of encrypted non-sensitive values.

Drawbacks Large encryption overhead.

- Incur high encryption cost.
- Reduce the data useability.

Related Work

Encryption in DaS model

- Searchable encryption [SWP00]: can not defend against FD attack.
- Homomorphic encryption [SV10]: inefficient.

Inference attack in Multi-level Security Database

- Database-design time [CHKP07, SO91]: over-encrypt the data.
- Query-time [BFJ00]: not applicable to our scenario.

K-anonymity

- Suppression and generalization [Swe02, WL11]: can not defend against FD attack.

Goal

Design a scheme.

- Robust against FD attack
- Efficiency
- Low encryption overhead

NM	SEX	AGE	DC	DS
Alice	F	53	CPD5	α
Carol	F	30	ι	Cancer
Ela	F	24	VPI8	γ

(encryption overhead: 1)

: sensitive data
: additional encrypted data

Sensitive/Evidence Records

$FD : X \rightarrow Y$.

For all records with the same (x, y) values,

Sensitive record $S(x, y)$

- $r[Y]$ is sensitive.
- $r[X]$ is not sensitive.

Evidence record $E(x, y)$

- $r[Y]$ is not sensitive.
- $r[X]$ is not sensitive.

RID	NM	SEX	AGE	DC	DS
r_1	Alex	M	36	VPI8	Cancer
r_2	Bob	M	53	VPI8	Cancer
r_3	Carol	F	30	VPI8	Cancer
r_4	Ela	F	24	VPI8	γ
r_5	Amy	F	20	VPI8	γ

$S : \Pi_{DS}\sigma_{AGE<30}$

$S(VPI8, Cancer) = \{r_4, r_5\}$

$E(VPI8, Cancer) = \{r_1, r_2, r_3\}$

: sensitive data

: sensitive records

: evidence records

Encryption for One Single SC

Pick the scheme which has smaller encryption overhead.

Scheme 1 Pick $A \in X$, encrypt $r[A]$ for $r \in S(x, y)$.

Scheme 2 Pick $A \in X \cup Y$, encrypt $r[A]$ for $r \in E(x, y)$.

RID	NM	SEX	AGE	DC	DS
r_1	Alex	M	36	VPI8	Cancer
r_2	Bob	M	53	VPI8	Cancer
r_3	Carol	F	30	VPI8	Cancer
r_4	Ela	F	24	ι	γ
r_5	Amy	F	20	ι	γ

(Scheme 1: *overhead* $= 2$)

RID	NM	SEX	AGE	DC	DS
r_1	Alex	M	36	ι	Cance
r_2	Bob	M	53	ι	Cance
r_3	Carol	F	30	ι	Cance
r_4	Ela	F	24	VPI8	γ
r_5	Amy	F	20	VPI8	γ

(Scheme 2: *overhead* $= 3$)

: sensitive data
: additional encrypted data

Encryption for Multiple SCs

Theorem (NP-Completeness)

Given a dataset D and $k > 1$ SCs $\mathcal{S}$, the problem of finding the optimal robust scheme that enforces $\mathcal{S}$ on D against the FD attack is NP-complete.

RID	NM	SEX	AGE	DC	DS
r_1	Joe	M	28	CPD5	α
r_2	Alice	F	24	CPD5	α
r_3	Maggy	F	33	CPD5	HIV
r_4	Phil	M	43	CPD5	HIV
r_5	Peter	M	39	CPD5	HIV
r_6	Ray	M	52	CPD5	HIV
r_7	Steve	M	31	CPD5	HIV

$S_1 : \Pi_{DS}\sigma_{AGE<30}$

RID	NM	SEX	AGE	DC	DS
r_1	Joe	M	28	CPD5	HIV
r_2	Alice	F	24	CPD5	α
r_3	Maggy	F	33	CPD5	α
r_4	Phil	M	43	CPD5	HIV
r_5	Peter	M	39	CPD5	HIV
r_6	Ray	M	52	CPD5	HIV
r_7	Steve	M	31	CPD5	HIV

$S_2 : \Pi_{DS}\sigma_{SEX=F}$

: sensitive data

Encryption for Multiple SCs

Theorem (NP-Completeness)

Given a dataset D and $k > 1$ SCs $\mathcal{S}$, the problem of finding the optimal robust scheme that enforces $\mathcal{S}$ on D against the FD attack is NP-complete.

RID	NM	SEX	AGE	DC	DS
r_1	Joe	M	28	CPD5	α
r_2	Alice	F	24	CPD5	α
r_3	Maggy	F	33	CPD5	α
r_4	Phil	M	43	CPD5	HIV
r_5	Peter	M	39	CPD5	HIV
r_6	Ray	M	52	CPD5	HIV
r_7	Steve	M	31	CPD5	HIV

$$S(S_1) = \{r_1, r_2\}$$
$$E(S_1) = \{r_4, r_5, r_6, r_7\}$$

RID	NM	SEX	AGE	DC	DS
r_1	Joe	M	28	CPD5	α
r_2	Alice	F	24	CPD5	α
r_3	Maggy	F	33	CPD5	α
r_4	Phil	M	43	CPD5	HIV
r_5	Peter	M	39	CPD5	HIV
r_6	Ray	M	52	CPD5	HIV
r_7	Steve	M	31	CPD5	HIV

$$S(S_2) = \{r_2, r_3\}$$
$$E(S_2) = \{r_4, r_5, r_6, r_7\}$$

: sensitive data
: sensitive records
: evidence records

Encryption for Multiple SCs

Theorem (NP-Completeness)

Given a dataset D and $k > 1$ SCs $\mathcal{S}$, the problem of finding the optimal robust scheme that enforces $\mathcal{S}$ on D against the FD attack is NP-complete.

Four solutions

Solution 1: encrypt $S(S_1)$ and $S(S_2)$

RID	NM	SEX	AGE	DC	DS
r_1	Joe	M	28	β	α
r_2	Alice	F	24	β	α
r_3	Maggy	F	33	β	α
r_4	Phil	M	43	CPD5	HIV
r_5	Peter	M	39	CPD5	HIV
r_6	Ray	M	52	CPD5	HIV
r_7	Steve	M	31	CPD5	HIV

encryption overhead $= 3$

Solution 2: encrypt $S(S_1)$ and E

RID	NM	SEX	AGE	DC	D
r_1	Joe	M	28	β	
r_2	Alice	F	24	β	
r_3	Maggy	F	33	CPD5	
r_4	Phil	M	43	β	H
r_5	Peter	M	39	β	H
r_6	Ray	M	52	β	H
r_7	Steve	M	31	β	H

encryption overhead $= 6$

: sensitive data
: additional encrypted data

Encryption for Multiple SCs

Theorem (NP-Completeness)

Given a dataset D and $k > 1$ SCs $\mathcal{S}$, the problem of finding the optimal robust scheme that enforces $\mathcal{S}$ on D against the FD attack is NP-complete.

Four solutions

Solution 3: encrypt $E(S_1)$ and $S(S_2)$

RID	NM	SEX	AGE	DC	DS
r_1	Joe	M	28	CPD5	α
r_2	Alice	F	24	β	α
r_3	Maggy	F	33	β	α
r_4	Phil	M	43	β	HIV
r_5	Peter	M	39	β	HIV
r_6	Ray	M	52	β	HIV
r_7	Steve	M	31	β	HIV

encryption overhead $= 6$

Solution 4: encrypt $E(S_1)$ and E

RID	NM	SEX	AGE	DC	
r_1	Joe	M	28	CPD5	
r_2	Alice	F	24	CPD5	
r_3	Maggy	F	33	CPD5	
r_4	Phil	M	43	β	H
r_5	Peter	M	39	β	H
r_6	Ray	M	52	β	H
r_7	Steve	M	31	β	H

encryption overhead $= 4$

: sensitive data
: additional encrypted data

Encryption for Multiple SCs

We design an efficient heuristic algorithm *GMM*:

Do Pick the option with the smallest overhead.

While unsafe against FD attack

RID	NM	SEX	AGE	DC	DS
r_1	Joe	M	28	CPD5	α
r_2	Alice	F	24	CPD5	α
r_3	Maggy	F	33	CPD5	α
r_4	Phil	M	43	CPD5	HIV
r_5	Peter	M	39	CPD5	HIV
r_6	Ray	M	52	CPD5	HIV
r_7	Steve	M	31	CPD5	HIV

$$S(S_1) = \{r_1, r_2\}$$
$$E(S_1) = \{r_4, r_5, r_6, r_7\}$$

RID	NM	SEX	AGE	DC	DS
r_1	Joe	M	28	CPD5	α
r_2	Alice	F	24	CPD5	α
r_3	Maggy	F	33	CPD5	α
r_4	Phil	M	43	CPD5	HIV
r_5	Peter	M	39	CPD5	HIV
r_6	Ray	M	52	CPD5	HIV
r_7	Steve	M	31	CPD5	HIV

$$S(S_2) = \{r_2, r_3\}$$
$$E(S_2) = \{r_4, r_5, r_6, r_7\}$$

: sensitive data
: sensitive records
: evidence records

Encryption for Multiple SCs

Do Pick the option with the smallest overhead.

While unsafe against FD attack

Step 1: encrypt $S(S_1)$

RID	NM	SEX	AGE	DC	DS
r_1	Joe	M	28	β	α
r_2	Alice	F	24	β	α
r_3	Maggy	F	33	CPD5	α
r_4	Phil	M	43	CPD5	HIV
r_5	Peter	M	39	CPD5	HIV
r_6	Ray	M	52	CPD5	HIV
r_7	Steve	M	31	CPD5	HIV

$S(S_2) = \{r_3\}, E(S_2) = \{r_4, r_5, r_6, r_7\}$

Step 2: encrypt $S(S_2)$

RID	NM	SEX	AGE	DC	DS
r_1	Joe	M	28	β	α
r_2	Alice	F	24	β	α
r_3	Maggy	F	33	β	α
r_4	Phil	M	43	CPD5	HI
r_5	Peter	M	39	CPD5	HI
r_6	Ray	M	52	CPD5	HI
r_7	Steve	M	31	CPD5	HI

encryption overhead $= 3$

: sensitive data
: additional encrypted data

Experiment Setup

- Environment

  Language Java
  Testbed 2.4*GHz* Intel Core *i*5 CPU, 4*GB* RAM, Mac OS X 10.9

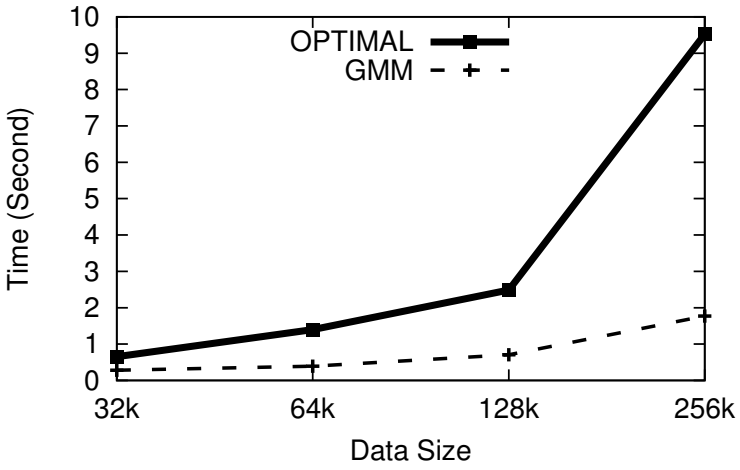
- Datasets:

  Adult UCI machine learning repository
  Orders TPC-H benchmark

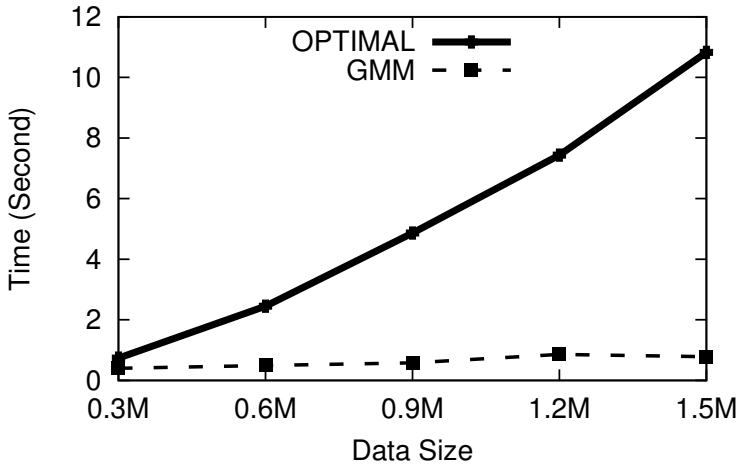
- Approaches

  GMM Our heuristic approach
  OPTIMAL The exhaustive search algorithm

Time Performance

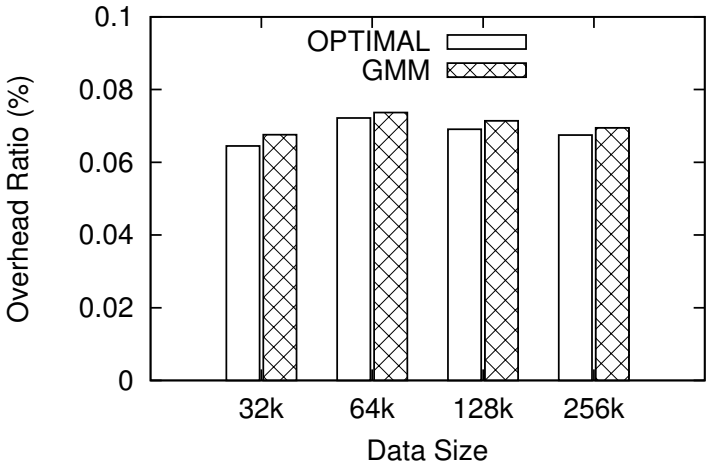


(a) Adult dataset

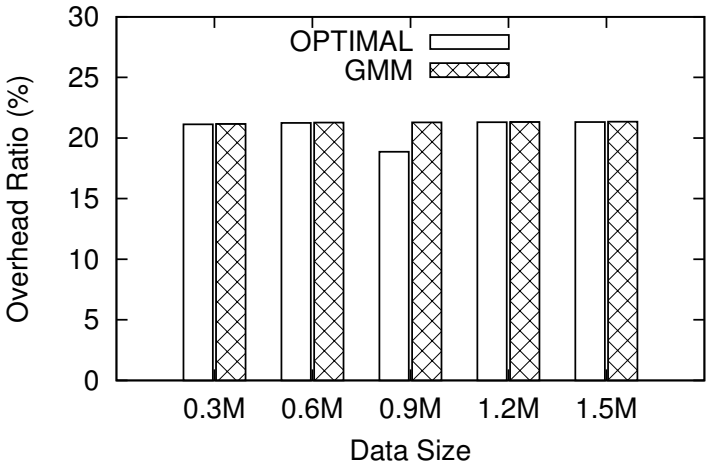


(b) Orders dataset

Encryption Overhead



(a) Adult dataset



(b) Orders dataset

Conclusion

A scheme against FD-based attack in the DaS model based on encryption.

- Formalize the FD attack.
- Prove that finding an optimal scheme with minimal overhead is NP-complete.
- Design efficient heuristic approaches to construct robust schemes with small overhead.

References I

[BFJ00] Alexander Brodsky, Csilla Farkas, and Sushil Jajodia.
Secure databases: Constraints, inference channels, and monitoring disclosures.
Knowledge and Data Engineering, IEEE Transactions on, 12(6):900–919, 2000.

[CHKP07] Laura Chiticariu, Mauricio A Hernández, Phokion G Kolaitis, and Lucian Popa.
Semi-automatic schema integration in clio.
In *Proceedings of the 33rd international conference on Very large data bases*, pages 1326–1329, 2007.

[SO91] T-A Su and Gultekin Ozsoyoglu.
Controlling fd and mvd inferences in multilevel relational database systems.
Knowledge and Data Engineering, IEEE Transactions on, 3(4):474–485, 1991.

[SV10] Nigel P Smart and Frederik Vercauteren.
Fully homomorphic encryption with relatively small key and ciphertext sizes.
In *Public Key Cryptography (PKC)*, pages 420–443. 2010.

[Swe02] Latanya Sweeney.
k-anonymity: A model for protecting privacy.
International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):557–570, 2002.

[SWP00] Dawn Xiaoding Song, David Wagner, and Adrian Perrig.
Practical techniques for searches on encrypted data.
In *Proceedings of IEEE Symposium on Security and Privacy*, pages 44–55, 2000.

[WL11] Hui Wang and Ruilin Liu.
Privacy-preserving publishing microdata with full functional dependencies.
Data & Knowledge Engineering, 70(3):249–268, 2011.

Q & A

Thank you!

Questions?