Tatouage des bases de données

École thématiques Masse de données distribuées – Les Houches

David Gross-Amblard

Laboratoire Le2i-CNRS
Université de Bourgogne, Dijon, France
http://ufrsciencestech.u-bourgogne.fr/~gadavid

École MDD / Les Houches / 20.5.2010

Plan

Plan

Définition

Tatouage (watermarking)

Altération volontaire et imperceptible d'un document (électronique) pour y dissimuler une information le concernant

Définition

Tatouage (watermarking)

Altération volontaire et imperceptible d'un document (électronique) pour y dissimuler une information le concernant

On laisse de côté le tatouage visible...

Définition

Tatouage (watermarking)

Altération volontaire et imperceptible d'un document (électronique) pour y dissimuler une information le concernant

Applications

- Incrustation de méta-données (marque : paramètre d'acquisition)
- Preuve de propriété (marque : identité du propriétaire)
- Forte valeur commerciale ou scientifique des données
  - licence : 50\$ km^2
  - reproduction : 500\$/km^2
  - licence complète : \$\$\$
- Traçabilité des copies (*fingerprinting*, marque : identité de receveur)

Définition

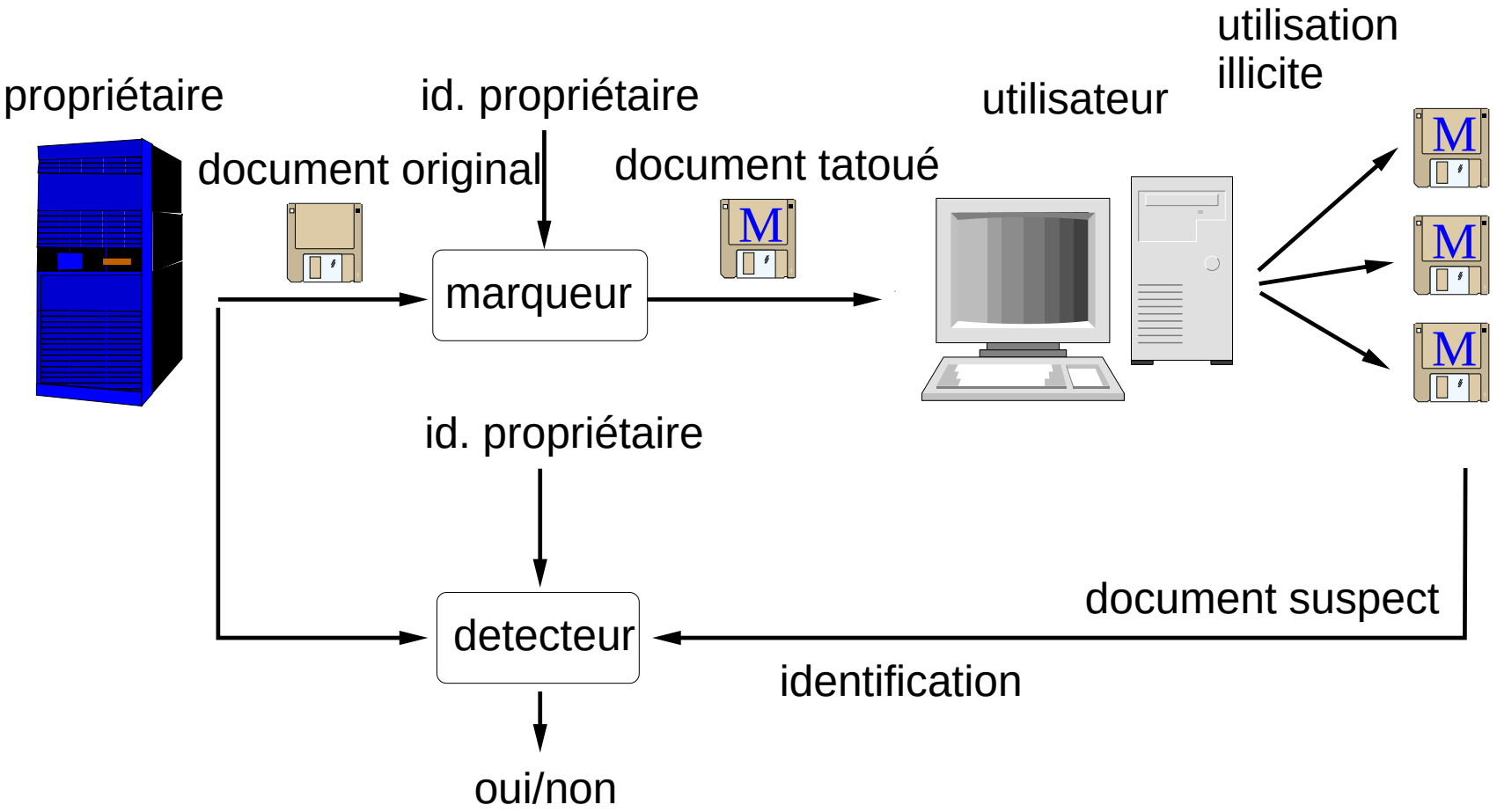
Tatouage (watermarking)

Altération volontaire et imperceptible d'un document (électronique) pour y dissimuler une information le concernant

Vocabulaire

- Information : présence/absence de la marque, marque éventuellement constituée de plusieurs bits (message)
- Capacité : nb. de bits de message dissimulés
- Tatouage imperceptible : préserve la qualité du document

Protocole de tatouage



Paire d'algorithmes : marqueur et détecteur (extracteur)

Exemple sur images

image originale

image tatouée

qualité : rapport signal/bruit avec original (PSNR)

marque imperceptible

Contexte malveillant

Tatouage robuste

- Robustesse : aux manipulations naturelles / aux attaques expertes (algo. public)
- Attaque réussie : marque non détectée, et donnée de bonne qualité

Tatouage fragile

- Marque altérée à la moindre modification
- Authenticité du document (localiser les falsifications)
- Pas traité ici

Contexte malveillant

Tatouage robuste

- Robustesse : aux manipulations naturelles / aux attaques expertes (algo. public)
- Attaque réussie : marque non détectée, et donnée de bonne qualité

Tatouage fragile

- Marque altérée à la moindre modification
- Authenticité du document (localiser les falsifications)
- Pas traité ici

Contexte malveillant

Tatouage robuste

- Robustesse : aux manipulations naturelles / aux attaques expertes (algo. public)
- Attaque réussie : marque non détectée, et donnée de bonne qualité

Tatouage fragile

- Marque altérée à la moindre modification
- Authenticité du document (localiser les falsifications)
- Pas traité ici

Attaque des Bavarois

changement d'échelle

JPEG 10%, smoothing 0%

rotation

Attaques

Compromis limitant propriétaire et attaquant

↑ force tatouage	↑ robustesse	↓ qualité
↑ force attaque	↑ succès	↓ qualité

Glossaire des attaques

- Bruit aléatoire
- Sous-ensemble
- Mélange
- Sur-tatouage
- Collusion
- Descente de gradient
- Inversion
- (votre proposition ici...)
- Pas de preuve de protocole

Attaques

Compromis limitant propriétaire et attaquant

↑ force tatouage	↑ robustesse	↓ qualité
↑ force attaque	↑ succès	↓ qualité

Glossaire des attaques

- Bruit aléatoire
- Sous-ensemble
- Mélange
- Sur-tatouage
- Collusion
- Descente de gradient
- Inversion
- (votre proposition ici...)
- Pas de preuve de protocole

Autres critères

Critères

- Tatouage aveugle : original inutile lors de la détection
- Taux de faux-positifs (faible)
- Complexité en temps (durée de vie des données)

Plan

Tatouer une base de données

id produit	stock	cat.
211	45	pizza
111	427	pizza
221	90	meuble
113	235	bio
223	74	laitage
331	125	meuble
9010221	61	chaussette
900001	249	chaussette

- Données numériques
- Données catégorielles
- Clé primaire
- Dépendances fonctionnelles
- Utilisation : données brutes et requêtes

Spécificité du tatouage des bases de données

Multimédia	Bases de données
Peu de structure	Très structuré : - schéma / types - clés primaires
Document unique	Grande collection
Utilisation unique	Interrogations variées langage de requêtes
Qualité psycho-perceptive PSNR	Dépend de l'application : description formelle
Document fixe	Mises à jour

Schéma classique de tatouage

A. Choisir une propriété $\mathcal{F}$ modifiable dans chaque n-uplet

- peu d'impact sur la qualité
- peu sensible aux transformations classiques
- bits de poids faible des nombres (ex. *distorsion* = 2 bits)

B. Calculer un identifiant *Id* robuste pour chaque n-uplet

- robuste aux altérations
- limiter les ambiguïtés (n-uplets ayant même identifiant)
- clé primaire

C. Introduire une dépendance secrète entre *Id* et $\mathcal{F}$

Schéma classique de tatouage : données numériques

Qualité : précision des nombres

A. Choisir une propriété $\mathcal{F}$ modifiable dans chaque n-uplet

- peu d'impact sur la qualité
- peu sensible aux transformations classiques
- bits de poids faible des nombres (ex. *distorsion* = 2 bits)

B. Calculer un identifiant *Id* robuste pour chaque n-uplet

- robuste aux altérations
- limiter les ambiguïtés (n-uplets ayant même identifiant)
- clé primaire

C. Introduire une dépendance secrète entre *Id* et $\mathcal{F}$

choix pseudo-aléatoires paramétrés par clé secrète K_p du propriétaire

Sécurité

Générateur pseudo-aléatoire cryptographique G

- Engendre une séquence de nombres facilement
- Suite statistiquement indifférentiable d'experiences aléatoires indépendantes identiquement distributées
- Etant donnée une partie de la séquence, difficile de prédire la suite
- La séquence est entièrement déterminée par un nombre *graine* (une même graine donne la même séquence)
- Fonctions $G.entierSuivant()$, $G.bitSuivant()$

Tatouage d'une BD numérique/Agrawal,Kiernan [1]

clé primaire P

données modifiables

clé secrète $\mathcal{K}$

Générateur G

G initialisé avec clé.$\mathcal{K}$

Quel n-uplet marquer ?

$G.entierSuivant()$ mod $periode = 0$?

id produit	stock	cat.
211	45	pizza
111	427	pizza
221	90	meuble
113	235	bio
223	74	laitage
331	125	meuble
9010221	61	chaussette
900001	249	chaussette

Tatouage d'une BD numérique/Agrawal,Kiernan [1]

clé primaire P

données modifiables

clé secrète $\mathcal{K}$

Générateur G

G initialisé avec clé.$\mathcal{K}$

Quel n-uplet marquer ?

$G.entierSuivant()$ mod $periode = 0$?

id produit	stock	cat.
211	45	pizza
111	427	pizza
221	90	meuble
113	235	bio
223	74	laitage
331	125	meuble
9010221	61	chaussette
900001	249	chaussette

Tatouage d'une BD numérique/Agrawal,Kiernan [1]

clé primaire P

données modifiables

clé secrète $\mathcal{K}$

Générateur G

G initialisé avec clé.$\mathcal{K}$

Quel n-uplet marquer ?

$G.entierSuivant()$ mod $periode = 0$?

id produit	stock	cat.
211	45	pizza
111	427	pizza
221	90	meuble
113	235	bio
223	74	laitage
331	125	meuble
9010221	61	chaussette
900001	249	chaussette

Tatouage d'une BD numérique

clé primaire P

données modifiables

vue en binaire

clé secrète $\mathcal{K}$

Générateur G

quel bit tatouer ?

$G.entierSuivant()$ mod $distorsion = ?$

id produit	stock	cat.
211	0101101	pizza
111	110101011	pizza
221	01011010	meuble
113	11101011	bio
223	1001010	laitage
331	1111101	meuble
9010221	111101	chaussette
900001	11111001	chaussette

Tatouage d'une BD numérique

clé primaire P

données modifiables

clé secrète $\mathcal{K}$

Générateur G

substituer par la marque $G.bitSuivant()$

id produit	stock	cat.
211	010110X0	pizza
111	110101011	pizza
221	01011010	meuble
113	11101011	bio
223	10010X10	laitage
331	1111101	meuble
9010221	111101	chaussette
900001	1111100X0	chaussette

Détection du tatouage

G initialisé avec clé.$\mathcal{K}$

localiser les n-uplets $G.entierSuivant()$ mod $periode$ $= 0$?

id produit	stock	cat.
211	44	pizza
111	427	pizza
221	90	meuble
113	235	bio
223	74	laitage
331	125	meuble
9010221	61	chaussette
900001	248	chaussette

Détection du tatouage

quel bit a été tatoué ?
$G.entierSuivant()$
mod $distorsion = ?$

id produit	stock	cat.
211	010110 0	pizza
111	110101011	pizza
221	01011010	meuble
113	11101011	bio
223	10010 1 0	laitage
331	1111101	meuble
9010221	111101	chaussette
900001	1111100 1	chaussette

Détection du tatouage

le bit caché correspond-t-il à $G.bitSuivant()$?

id produit	stock	cat.
211	010110 0	pizza
111	110101011	pizza
221	01011010	meuble
113	11101011	bio
223	10010 1 0	laitage
331	1111101	meuble
9010221	111101	chaussette
900001	1111100 1	chaussette

Sur beaucoup de n-uplets, taux de correspondance $>> \frac{1}{2}$: suspect !

(analyse : binomiale et loi des grands nombres)

Méthode d'Agrawal et Kiernan

Avantages

- Robuste (bruit, sous-ensemble, mélange), mais pas arrondi
- Aveugle
- Incrémentale
- Efficace
- Si absence de clé primaire : bits de poids fort, mais attention...[6, 8]

Plan

Généralisation

Autres types de données

données	qualité	synchronisation	cible
flux numériques	précision	fenêtre poids forts	poids faibles [16]
flux XML	edit-distance	typage	réécritures [10]
géométriques	précision	signal / topologie	positions des points [12, 5, 4, 3]
topographiques	précision, angles	poids fort du centroïde	longueur principale [9]
catégorielles	sémantique	clé primaire	substitution [17]
textuelles	distance sémantique	structure de la phrase	synonymes, généralisation [2, 18]

Généralisation

Préserver le résultat de requêtes importantes (jointure, agrégats

- Méthode gloutonne, optimisation [15, 14]
- Analyse des dépendances données / requêtes [7, 11]

Complément

Sur `http://ufrsciencestech.u-bourgogne.fr/~gadavid/mdd`

- Démo de la méthode Agrawal et Kiernan (Java)
- Démo du tatouage géographique (Java/OpenJump/Watergoat)
- Un (brouillon de) cours plus complet (remarques appréciées)
- Quelques articles de référence

Merci.

Complément

Sur `http://ufrsciencestech.u-bourgogne.fr/~gadavid/mdd`

- Démo de la méthode Agrawal et Kiernan (Java)
- Démo du tatouage géographique (Java/OpenJump/Watergoat)
- Un (brouillon de) cours plus complet (remarques appréciées)
- Quelques articles de référence

Merci.

Références bibliographiques I

Rakesh Agrawal, Peter J. Haas, and Jerry Kiernan.
Watermarking Relational Data : Framework, Algorithms and Analysis.
VLDB J., 12(2) :157–169, 2003.

Mikhail J. Atallah, Victor Raskin, Christian Hempelmann, Mercan Karahan, Radu Sion, Umut Topkara, and Katrina E. Triezenberg.
Natural language watermarking and tamperproofing.
In Petitcolas [13], pages 196–212.

Cyril Bazin, Jean-Marie Le Bars, and Jacques Madelaine.
A blind, fast and robust method for geographical data watermarking.
In *ASIACCS '07 : Proceedings of the 2nd ACM symposium on Information, computer and communications security*, pages 265–272, New York, NY, USA, 2007. ACM.

Oliver Benedens.
Affine invariant watermarks for 3d polygonal and nurbs based models.
In Josef Pieprzyk, Eiji Okamoto, and Jennifer Seberry, editors, *ISW*, volume 1975 of *Lecture Notes in Computer Science*, pages 15–29. Springer, 2000.

Oliver Benedens.
Robust watermarking and affine registration of 3d meshes.
In Petitcolas [13].

Références bibliographiques II


F. Cayre, C. Fontaine, and T. Furon.
Watermarking security : Theory and practice.
IEEE Transactions on Signal Processing, 53(10) :3976–3987, 2005.
special issue "Supplement on Secure Media III".


David Gross-Amblard.
Query-Preserving Watermarking of Relational Databases and XML Documents.
In *Symposium on Principles of Databases Systems (PODS)*, pages 191–201, 2003.


Julien Lafaye.
An analysis of database watermarking security.
In *IAS*, pages 462–467. IEEE Computer Society, 2007.


Julien Lafaye, Jean Béguec, David Gross-Amblard, and Anne Ruas.
Invisible graffiti on your buildings : Blind and squaring-proof watermarking of geographical databases.
In Dimitris Papadias, Donghui Zhang, and George Kollios, editors, *SSTD*, volume 4605 of *Lecture Notes in Computer Science*, pages 312–329. Springer, 2007.


Julien Lafaye and David Gross-Amblard.
XML streams watermarking.
In *IFIP WG 11.3 Working Conference on Data and Applications Security (DBSEC)*, 2006.

Références bibliographiques III


Julien Lafaye, David Gross-Amblard, Camélia Constantin, and Meryem Guerrouani.
Watermill : An optimized fingerprinting system for databases under constraints.
IEEE Trans. Knowl. Data Eng. (TKDE), 20(4) :532–546, 2008.


R. Ohbuchi, H. Masuda, and M. Aono.
Watermarking 3D polygonal models.
In *ACM Multimedia'97*, 1997.


Fabien A. P. Petitcolas, editor.
Information Hiding, 5th International Workshop, IH 2002, Noordwijkerhout, The Netherlands, October 7-9, 2002, Revised Papers, volume 2578 of *Lecture Notes in Computer Science*. Springer, 2003.


Mohamed Shehab, Elisa Bertino, and Arif Ghafoor.
Watermarking relational databases using optimization-based techniques.
IEEE Trans. Knowl. Data Eng. (TKDE), 20(1) :116–129, 2008.


Radu Sion, Mikhail Atallah, and Sunil Prabhakar.
Protecting rights over relational data using watermarking.
IEEE Trans. Knowl. Data Eng. (TKDE), 16(12) :1509–1525, December 2004.

Références bibliographiques IV


Radu Sion, Mikhail J. Atallah, and Sunil Prabhakar.
Resilient rights protection for sensor streams.
In Mario A. Nascimento, M. Tamer Özsu, Donald Kossmann, Renée J. Miller, José A. Blakeley, and K. Bernhard Schiefer, editors, *VLDB*, pages 732–743. Morgan Kaufmann, 2004.


Radu Sion, Mikhail J. Atallah, and Sunil Prabhakar.
Rights protection for categorical data.
IEEE Trans. Knowl. Data Eng., 17(7) :912–926, 2005.


Umut Topkara, Mercan Topkara, and Mikhail J. Atallah.
The hiding virtues of ambiguity : quantifiably resilient watermarking of natural language text through synonym substitutions.
In Sviatoslav Voloshynovskiy, Jana Dittmann, and Jessica J. Fridrich, editors, *MM&Sec*, pages 164–174. ACM, 2006.