



Epistemic Privacy

Alexandre Evfimievski, Ronald Fagin,
David Woodruff

PODS 2008, Vancouver, Canada
9-11 June 2008



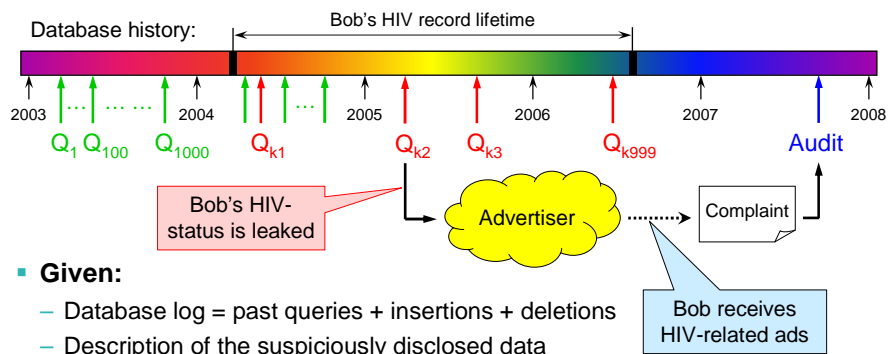
Introduction

- **Data Privacy:** The right to determine when, how and to what extent your personal information is communicated to others. [Westin 1967]
 - Databases store personal records, e.g. of hospital patients
 - Users issue queries
 - Privacy restrictions: (a) privacy policy, (b) patients' preferences, (c) database users' purpose & credentials etc.
- Two types of privacy restrictions
 - **Positive:** what is OK to disclose
 - Easier to enforce, e.g. by secure function evaluation
 - **Negative:** what is NOT OK to disclose (*today's topic*)

Proactive (Online) Enforcement

- User issues a query, the database must answer or deny it “on the fly” while protecting privacy [Kenthapadi *et al.* 2005]
- Assume truthful query answers – no random perturbation
- *A query denial may disclose information!*
- **Example:** Alice asks Bob every year if he is HIV-positive
 - 2006: Bob answers “No” (Only “Yes” is private for Bob.)
 - 2007: Bob answers “No”
 - 2008: Bob refuses to answer – Alice suspects “Yes”!

Retroactive (Offline) Auditing



- **Given:**
 - Database log = past queries + insertions + deletions
 - Description of the suspiciously disclosed data
- **Required:**
 - Find queries that could be the source of the audited data (*today's topic*)

Retroactive Auditing: Motivation

- **Simpler:** Users learn nothing from the auditor
- **Flexible:** No need to foresee / spell out all disclosure cases
 - For complex, unpredictable applications: Healthcare, Police
- **Efficient:** Transfers more computational burden to users
- **Mandatory:** Needed alongside proactive
 - Proactive enforcement does not protect from abuse of privileges and data loss / theft

Talk Outline

- Introduction
- General Framework
- Special Case: Product Distributions
- Conclusion

Privacy and Knowledge

- Let A and B be two properties of the database
 - A is private
 - B is disclosed to the user
- Whether or not B violates the privacy of A — depends on user's prior knowledge:
 - If user knows A in advance, then B does not violate A
 - If user knows "B \Rightarrow A" but not A, then B violates A
- We need to define knowledge

Reasoning About Knowledge

- "Possible worlds" framework:
 - $\Omega = \{\text{all databases}\}$, actual database: $\omega^* \in \Omega$
 - Property A $\equiv \{\omega \in \Omega \mid \omega \text{ satisfies } A\}$
- User : *possibilistic* model
 - Knowledge: set S, $\omega^* \in S$
 - User *knows* property A when $S \subseteq A$
- User : *probabilistic* model
 - Knowledge: distribution $P : \Omega \rightarrow [0, 1]$, $P(\omega^*) > 0$
 - $P[A] \equiv \sum_{\omega \in A} P(\omega)$ is the user's *confidence* that $\omega^* \in A$

Learning a Fact

- A *fact* is a property $B \subseteq \Omega$ such that $\omega^* \in B$
 - E.g. query answer: $B = \{\omega \in \Omega \mid \text{query}(\omega) = \text{query}(\omega^*)\}$
- Learning B: User discards all $\omega \notin B$
 - Possibilistic knowledge: S becomes $B \cap S$
 - Probabilistic knowledge: $P(\omega)$ becomes $P(\omega \mid B)$

Prior Work

- Shannon [1949]: \cong Require $P[A \mid B] = P[A]$ for all distributions P over Ω (“*perfect secrecy*”)
 - Violated by all pairs (A, B) of nontrivial properties
- Miklau & Suciú [2004]: Require $P[A \mid B] = P[A]$ for all P that draw each record of database ω independently
 - **Theorem (M.-S. Criterion):** Property B preserves privacy of $A \Leftrightarrow$ no database record is critical for both A and B
 - Record R is *critical* for property A means: $A(\omega) \neq A(\omega - \{R\})$ for some database ω
 - That is, A and B must depend on disjoint sets of records
 - Overly sensitive, e.g. to aggregate queries

Even if ω is imaginary

The Challenge

Are there *relaxed* privacy definitions which:

- Model user's knowledge (mathematically sound)
- Tolerate more disclosure (practically useful) ?

Our Approach

- Informal privacy definition: **No user can gain confidence in audited property A by learning disclosed fact B.**
 - Arbitrary loss of confidence in A is OK
- *Possibilistic* privacy:
 - Learning B does not flip the user from “not knowing A” to “knowing A”
 - $(B \cap S) \subseteq A \Rightarrow S \subseteq A$ (S is user's knowledge)
- *Probabilistic* privacy:
 - Learning B does not increase the user's probability of A
 - $P[A | B] \leq P[A]$ (P is user's knowledge)

This permits many more queries, some even with *no assumptions on the user!*

Auditor's Knowledge / Assumption

- Auditor's *actual* world is a pair (ω^*, S^*) or (ω^*, P^*)
 - $\omega^* \in \Omega$ is the actual database
 - S^* (set) or P^* (distribution) is the user's prior knowledge
- Auditor's *possible* worlds are pairs (ω, S) or (ω, P)
 - $\omega \in \Omega$ is a possible ω^* given the DB log
 - S or P is a possible S^* or P^* , assuming that $\omega^* = \omega$
- Auditor's *knowledge* is a set K of all (ω, S) or (ω, P) the auditor considers possible; auditor's *assumption* is a superset of K
- Privacy definitions with K
 - **Possibilistic:** $\forall (\omega, S) \in K, \omega \in B: (B \cap S) \subseteq A \Rightarrow S \subseteq A$
 - **Probabilistic:** $\forall (\omega, P) \in K, \omega \in B: P[A | B] \leq P[A]$

Example of Safe A and B

- Consider a database with two records: $\omega^* = \{X, Y\}$
 - Record X = "Bob is HIV-positive" (private / audited)
 - Record Y = "Bob had blood transfusions" (cleared)
- Private property: $A = \{\omega \in \Omega \mid X \in \omega\}$ "X is true"
- Disclosed fact: $B = \{\omega \in \Omega \mid X \in \omega \Rightarrow Y \in \omega\}$ "X \Rightarrow Y"
- Learning B discards some $\omega \in A$, but retains all $\omega \in \neg A$
 - The odds of A can only go down!
 - $\forall P: P[A | B] \leq P[A]$
- So, B preserves the privacy of A
 - Regardless of user's knowledge
 - Despite a shared critical record X

	$Y \in \omega$	$Y \notin \omega$
$X \in \omega$		X
$X \notin \omega$		

$A = \text{red square} = "X"$
 $\neg B = \text{crossed square} = "X \& \neg Y"$

Unrestricted Prior Knowledge

- When is B *always* safe to disclose while protecting A?
 - For all consistent pairs (ω, S) or (ω, P) of the database and the user
- **Proposition:** A, B always safe $\Leftrightarrow A \cap B = \emptyset$ or $A \cup B = \Omega$
 - “ $A \cap B = \emptyset$ ” (“ $B \Rightarrow \neg A$ ”): A is false, and B discloses $\neg A$;
 - “ $A \cup B = \Omega$ ” (“ $\neg B \Rightarrow A$ ”): Learning B discards some $\omega \in A$, but retains all $\omega \in \neg A$
- So, B may be safe for A even though A and B depend on the same (critical) database records

Talk Outline

- Introduction
- General Framework
- Special Case: Product Distributions
- Conclusion

Product Distributions

- Let $\Omega = \{0, 1\}^n$
 - $\omega = (\omega[1], \omega[2], \dots, \omega[n]); \omega[i] = 1 \Leftrightarrow \text{record } \#i \in \text{database}$
- Bit-wise independence — family of *product* distributions:

$$\forall \omega \in \Omega: P(\omega) = \prod_{i=1}^n p_i^{\omega[i]} \cdot (1-p_i)^{1-\omega[i]}$$

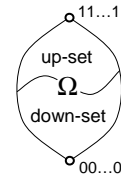
All records are assumed independent

- Equivalent definition (“ \wedge ”, “ \vee ” – bitwise AND, OR):
 - $\forall \omega, \omega' \in \Omega: P(\omega) \cdot P(\omega') = P(\omega \wedge \omega') \cdot P(\omega \vee \omega')$
- Relaxation of independence — *log-supermodular* distributions:
 - $\forall \omega, \omega' \in \Omega: P(\omega) \cdot P(\omega') \leq P(\omega \wedge \omega') \cdot P(\omega \vee \omega')$

A discrete analog of convex functions

Monotonicity Criterion

- For $\Omega = \{0, 1\}^n$, an *up-set* is a “Yes” answer to a monotone query, a *down-set* is a “No” answer.
- Monotonicity criterion of privacy:
 - If the user’s knowledge is log-supermodular, it is safe to disclose a down-set B while protecting an up-set A (or vice versa).



Follows from *Four-Functions Theorem* [Ahlsweede, Daykin 1978]:

- Assume: $\forall \omega, \omega' \in \Omega: P_1(\omega) \cdot P_2(\omega') \leq P_3(\omega \wedge \omega') \cdot P_4(\omega \vee \omega')$
- Then we have: $\forall S, S' \subseteq \Omega: P_1[S] \cdot P_2[S'] \leq P_3[S \wedge S'] \cdot P_4[S \vee S']$
 - Notation: $S \wedge S' := \{\omega \wedge \omega' \mid \omega \in S, \omega' \in S'\}$, $S \vee S' := \{\omega \vee \omega' \mid \omega \in S, \omega' \in S'\}$
 - Take $P_{1..4} = P$, $S = \bar{A}$, $S' = B$, use $\bar{A} \wedge B = \bar{A} \cap B$ for down-sets \bar{A} and B

Polynomials

- For $\Omega = \{0, 1\}^n$, privacy over the product distributions \Leftrightarrow an inequality for polynomials over (p_1, \dots, p_n) :

$$\forall P \in \Pi: P[A | B] \leq P[A] \Leftrightarrow$$

$$\sum_{\substack{\omega \in A \cap B \\ \omega' \notin A \cup B}} \prod_{i=1}^n p_i^{\omega[i] + \omega'[i]} (1 - p_i)^{2 - \omega[i] - \omega'[i]} \leq \sum_{\substack{\omega \in A - B \\ \omega' \in B - A}} \prod_{i=1}^n p_i^{\omega[i] + \omega'[i]} (1 - p_i)^{2 - \omega[i] - \omega'[i]}$$

$(\forall i = 1 \dots n: 0 \leq p_i \leq 1)$

- All terms in both sums are non-negative

$\Pi = \{\text{product distributions}\}$:

$$P(\omega) = \prod_{i=1}^n p_i^{\omega[i]} \cdot (1 - p_i)^{1 - \omega[i]}$$

Cancellation Criterion

- A simple *cancellation criterion* (C.C.) of privacy, for $P \in \Pi$:
 - B is safe while protecting A if each term to the left of “ \leq ” is cancelled by an identical term to the right of “ \leq ”, in the expression

$$\sum_{\substack{\omega \in A \cap B \\ \omega' \notin A \cup B}} \prod_{i=1}^n p_i^{\omega[i] + \omega'[i]} (1 - p_i)^{2 - \omega[i] - \omega'[i]} \leq \sum_{\substack{\omega \in A - B \\ \omega' \in B - A}} \prod_{i=1}^n p_i^{\omega[i] + \omega'[i]} (1 - p_i)^{2 - \omega[i] - \omega'[i]}$$

$(\forall i = 1 \dots n: 0 \leq p_i \leq 1)$

$\Pi = \{\text{product distributions}\}$:

$$P(\omega) = \prod_{i=1}^n p_i^{\omega[i]} \cdot (1 - p_i)^{1 - \omega[i]}$$

- The C.C. is relatively strong:
 - If A is an up-set and B is a down-set, they satisfy the C.C.
 - If A and B do not share critical records, they satisfy the C.C.

(Miklau-Suciu criterion)

Beyond Cancellation

- While sufficient, cancellation criterion is *not necessary*:
 - Take $\Omega = \{0, 1\}^3$; for $\omega = (u, x, y)$ define $A \equiv ux \vee \bar{u}y$, $B \equiv uy \vee \bar{u}x$
 - Disclosing B is safe for A, over $P \in \Pi$
 - The terms do not cancel, unless we apply $2\alpha\beta \leq \alpha^2 + \beta^2$
- Can we find a stronger sufficient criterion?

Hilbert's 17th problem

- In 1900, Hilbert asked:
 - Can we represent every non-negative rational function as a sum of squares of rational functions?
 - Rational function is a fraction $f(x_1, \dots, x_n) / g(x_1, \dots, x_n)$ of two polynomials
- In 1927, Artin answered “Yes” (but the sum may be long & complex)
 - This suggests a sufficient non-negativity test for $P[A] - P[A | B]$ over the product distributions $P \sim (p_1, \dots, p_n) \in \Pi$: represent

$$P[A] - P[A | B] = R_1(\vec{x})^2 + \dots + R_m(\vec{x})^2 = \vec{r}(\vec{x})^T M \vec{r}(\vec{x})$$

$$p_i = x_i^2 / (x_i^2 + 1)$$

$$-\infty \leq x_i \leq +\infty$$

Pick a “library” of rational functions

Look for a positive semi-definite matrix

(Parrilo, Sturmfels [2001] implement a more practical version)

How Hard is Privacy Testing?

- For non-product distribution families, checking privacy can be a hard problem
 - We can show the NP-hardness of checking $\forall P \in \Pi': P[A | B] \leq P[A]$ for specific families Π' of distributions
 - (sets A and B are given by their element lists)

Efficiency Bound for Π

- Let $\Omega = \{0, 1\}^n$
- Sets $A, B \subseteq \Omega$ are given by element lists
 - Input size $N = |A| + |B| = O(2^n)$
- **Theorem:** Testing if the disclosure of B protects A, over the product distributions, can be done in time $N^{O(\log \log N)}$.
 - Almost polynomial in the input size!

Reduction to a result of Basu, Pollack and Roy [1996] on the complexity of quantifier elimination over \mathbb{R}

- Efficient in N because, for product distributions, the polynomial $P[A] \cdot P[B] - P[A \cap B]$ has only $n \approx \log N$ variables

Conclusion

- **The Challenge:** Privacy definitions that are mathematically sound, yet practically useful
- **Our Approach:** Protect against confidence gain in A, allow arbitrary confidence loss (*“semi-perfect secrecy”*)
- **Our Results:** Many more queries permitted, some with no assumptions or with relaxed assumptions on the user
- **Future Work:**
 - Efficient privacy tests for Select-Project-Join queries
 - Extend to proactive privacy enforcement
 - *Other flexible privacy definitions??*

Thank You!

Questions?

BACK-UP

Polynomials

- Privacy over the family Π of product distributions translates into an inequality for polynomials:

$$- \forall P \in \Pi: P[A \mid B] \leq P[A] \Leftrightarrow$$

$$- \forall P \in \Pi: P[A \cap B] \leq P[A] \cdot P[B] \Leftrightarrow$$

$$- \forall P \in \Pi: P[A \cap B] \cdot P[\Omega - (A \cup B)] \leq P[A - B] \cdot P[B - A] \Leftrightarrow$$

$$- \forall P \in \Pi: \sum_{\substack{\omega \in A \cap B \\ \omega' \notin A \cup B}} P(\omega) \cdot P(\omega') \leq \sum_{\substack{\omega \in A - B \\ \omega' \in B - A}} P(\omega) \cdot P(\omega') \Leftrightarrow$$

$$\sum_{\substack{\omega \in A \cap B \\ \omega' \notin A \cup B}} \prod_{i=1}^n p_i^{\omega[i] + \omega'[i]} (1 - p_i)^{2 - \omega[i] - \omega'[i]} \leq \sum_{\substack{\omega \in A - B \\ \omega' \in B - A}} \prod_{i=1}^n p_i^{\omega[i] + \omega'[i]} (1 - p_i)^{2 - \omega[i] - \omega'[i]}$$

$(\forall i = 1 \dots n: 0 \leq p_i \leq 1)$