

A Cryptographic Approach to Privacy:

Privacy Preserving Learning of Decision Trees



Benny Pinkas, HP Labs

Joint work with Yehuda Lindell
(done at the Weizmann Institute)

Cryptographic Protocols for Privacy Preserving Computation



Input:

x

y

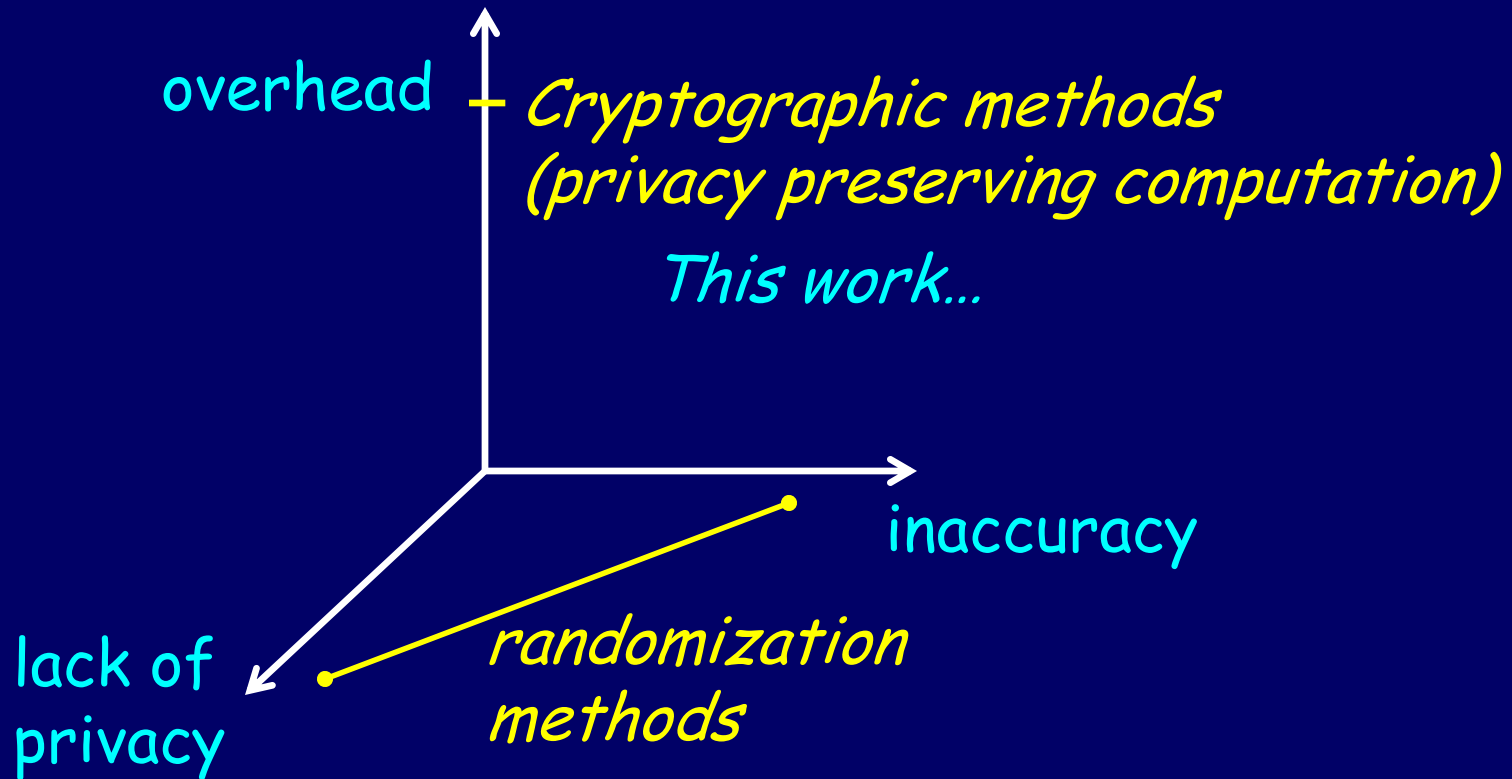
Output:

$F(x,y)$ and nothing else

Examples (with reasonable solutions):

- Is $X = Y$? Is $X > Y$?
- What items appear in both X and Y ?
- *Auctions (negotiations)*. Many parties, private bids. Compute the winning bidder and the sale price, but nothing else.
- Add privacy to existing data mining algorithms.

Cryptographic methods vs. randomization methods



A story



We're experiencing
a lot of fraud patterns

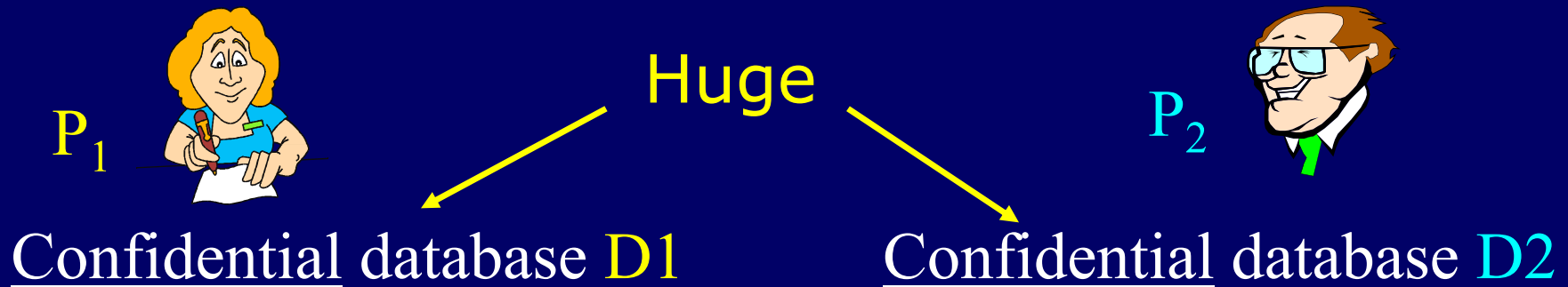
to recognize fraud in advance. But, what about
Maybe we should share information. Here too.. Neither can I..

- Patients' privacy
- Business secrets

Have you heard of "Secure
function evaluation" ?

This is all "theory".
It can't be efficient.

Privacy preserving data mining



Wish to “mine” $D1 \cup D2$ without revealing more info

Examples:

- Medical databases protected by law
- Competing businesses
- Government agencies (privacy, “need to know”)

Secure Function Evaluation [Yao '86]

- $F(x,y)$ - A public function.
- Represented as a Boolean circuit $C(x,y)$.



x



y

Input:

Output:

$C(x,y)$ and nothing else

Implementation:

- Two passes
- $O(|X|)$ "oblivious transfers". $O(|C|)$ communication.
- **Pretty efficient for small circuits!**

Our Contribution

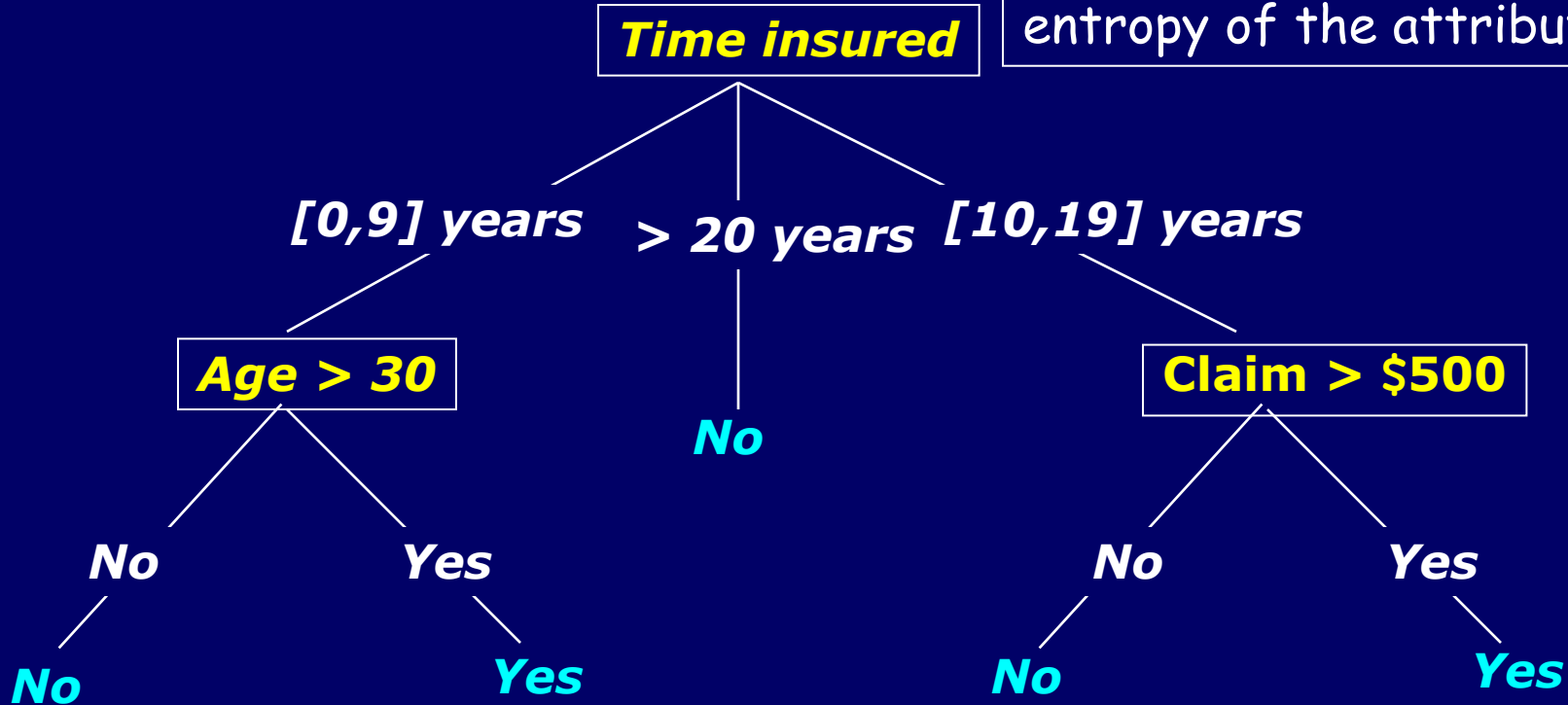
- An efficient sub-linear protocol for secure computation of a complex well-known data-mining alg (ID3).
- Comparison to the randomization approach (e.g. [AS'00])
 - Neither privacy nor accuracy are compromised
 - Higher overhead (but much lower than one could expect)

The classification problem

	Age > 30	Sex	History	Claim > \$500	Did fraud occur?
C1	Yes	M	$t \in [0,9]$ years	No	No
C2	No	F	$t \in [10,19]$ years	Yes	Yes
...
Cn	Yes	F	$t \in [20,29]$ years	No	No

Classification using Decision Trees

ID3: Choose attribute A that minimizes the conditional entropy of the attribute class



Privacy Preserving ID3

- A circuit encoding of ID3 is huge...
- **Core of the problem:** Comparing entropies while preserving privacy. (entropy = $\sum x \log x$)
- **Privacy:** for each party, all intermediate values are random.
- **Efficiency:** most computation done independently by parties.

- **Basic task:** compute $x \log x$.

x = e.g. # of patients with (age > 30) and (fraud = yes)

Privacy Preserving ID3

- *Computing $x \log x$:*
 - $x = x_1 + x_2$ known to P_1 and P_2 respectively (independently computed from databases).
 - Might as well compute ~~$x \ln x$~~ $\ln x$.
 - Should design a protocol to compute random shares, $y_1 + y_2 = \ln x$
- $\ln x$ is *Real*. Crypto works over finite fields. Must do numerical analysis.

Cryptographic Tools

- Secure Function Evaluation [Yao]
- Oblivious Polynomial Evaluation [NP]



Input:

x

$Q(\cdot)$

Output:

$Q(x)$ and nothing else

nothing

Efficient implementation:

Two passes, $O(\text{degree})$ (or $O(\log|F|)$) exponentiations.

Computing random shares of $\ln x = \ln(x_1 + x_2)$

Use Taylor approximation for $\ln x$

- $x = x_1 + x_2 = 2^n (1 + \varepsilon) \quad -\frac{1}{2} < \varepsilon < \frac{1}{2}$

- $\ln x = \ln(2^n (1 + \varepsilon)) = \ln 2^n + \ln(1 + \varepsilon)$

$$\approx \ln 2^n + \sum_{i=1..k} (-1)^{i-1} \varepsilon^i / i$$

$$= \ln 2^n + T(\varepsilon)$$

- $T(\varepsilon)$ is a polynomial of degree k . Error is exponentially small in k .

$\ln(x_1+x_2)$ Protocol (Cont.)

- Step 1 of the protocol - Find n, ε
 - Apply Yao's protocol to the following **small** circuit
 - Input: x_1 and x_2
 - Output (random shares):
 - random a_1 and a_2 s.t. $a_1 + a_2 = x - 2^n = \varepsilon \cdot 2^n$
 - random b_1 and b_2 s.t. $b_1 + b_2 = \ln 2^n$
 - Operation:
 - Find 2^n closest to x_1+x_2
 - Compute $\varepsilon 2^n = x_1+x_2 - 2^n$.

$\ln(x_1+x_2)$ Protocol (Cont)

Step 2 of the protocol

- Compute random shares of $T(\varepsilon)$ (Taylor approx.)
- P_1 chooses a random $w_1 \in F$ and defines a polynomial $Q(x)$, s.t. $w_1 + Q(a_2) = T(\varepsilon)$
- Namely, $Q(x) = T((a_1+x)/2^n) - w_1$.
- P_2 runs an oblivious polynomial evaluation to compute $w_2 = Q(a_2) = T(\varepsilon) - w_1$.
- Now the parties have random w_1 and w_2 s.t.
 - $w_1 + w_2 = T(\varepsilon) \approx \ln(1+\varepsilon)$
 - $(b_1 + w_1) + (b_2 + w_2) \approx \ln 2^n + \ln(1+\varepsilon) = \ln x$

The rest of the work..

- Compute $x \ln x$ (given x and $\ln x$).
- Each party computes a share of the entropy by summing shares of $x \ln x$
- A small circuit finds the attribute giving the minimal conditional entropy
- The attribute is assigned to the node
- The databases are divided according to the value of this attribute

Efficiency

- Inx protocol:
 - secure computation of a small circuit
 - one oblivious polynomial evaluation
- ID3 for a database with:
 - 1,000,000 transactions
 - 15 attributes
 - 10 values per attribute
 - 4 class values
 - Communication per node takes seconds (T1)
 - Computation per node takes minutes (P3)

Issues

- Only two participants
- "Curious but honest" participants
- Approximating $\ln x$ gives an approximation of ID3
- The participants learn the decision tree, which reveals some information

Contributions

- A cryptographic protocol where the bulk of the operations is done independently. Therefore useful for large databases.
- **Data mining**
 - Rigorous model for privacy in data-mining.
 - Efficient, secure protocol for ID3.
- **Cryptography**
 - **Sub-linear** complexity.
 - An efficient protocol for a **complex** known algorithm.
 - Secure computation of **logarithms** (real function - numerical analysis).

An announcement

- *Workshop on Privacy Preserving Computation*
- DIMACS, Rutgers University, NJ.
- Fall 2003.
- Organizers: Benny Pinkas, Rebecca Wright.