Privacy-Protecting Statistics Computation: Theory and Practice

Rebecca Wright
Stevens Institute of Technology

27 March, 2003

Erosion of Privacy

ì You have zero privacy. Get over it.î

- Scott McNealy, 1999

¡ Changes in technology are making privacy harder.

ñ reduced cost for data storage

ñ increased ability to process large amounts of data

¡ Especially critical now (given increased need for security-related surveillance and data mining)

Overview

- Announcements
- Introduction
- Privacy-preserving statistics computation
- Selective private function evaluation

Announcements

- DIMACS working group on secure efficient extraction of data from multiple datasets. Initial workshop to be scheduled for Fall 2003.
- DIMACS crypto and security tutorials to kick off Special Focus on Communication Security and Information Privacy: August 4-7, 2003.
- NJITES Cybersecurity Symposium, Stevens Institute of Technology, April 28, 2003.

What is Privacy?

- May mean different things to different people
  - seclusion: the desire to be left alone
  - property: the desire to be paid for oneís data
  - autonomy: the ability to act freely
- Generally: the ability to control the dissemination and use of oneís personal information.

Different Types of Data

- Transaction data
  - created by interaction between stakeholder and enterprise
  - current privacy-oriented solutions useful
- Authored data
  - created by stakeholder
  - digital rights management (DRM) useful
- Sensor data
  - stakeholders not clear at time of creation
  - growing rapidly

Sensor Data Examples

- surveillance cameras (especially with face recognition software)
- desktop monitoring software (e.g. for intrusion or misbehavior detection)
- GPS transmitters, RFID tags
- wireless sensors (e.g. for location-based PDA services)

Sensor Data

- Can be difficult to identify stakeholders and even data collectors
- Cross boundary between ìreal worldî and cyberspace
- Boundary between transaction data and sensor data can be blurry (e.g. Web browsing data)
- Presents a real and growing privacy threat

Product Design as Policy Decision

- product decisions by large companies or public organizations become de facto policy decisions
- often such decisions are made without conscious thought to privacy impacts, and without public discussion
- this is particularly true in the United States, where there is not much relevant legislation

Example: Metro Cards

Washington, DC

- no record kept of per card transactions
- damaged card can be replaced if printed value still visible

New York City

- transactions recorded by card ID
- damaged card can be replaced if card ID still readable
- have helped find suspects, corroborate alibis

Transactions without Disclosure

→ Doní t disclose information in first place!

¡ Anonymous digital cash [Chaum et al]

¡ Limited-use credit cards [Sha01, RW01]

¡ Anonymous web browsing [Crowds, Anonymizer]

¡ Secure multiparty computation and other cryptographic protocols

ñ perceived (often correctly) as too cumbersome or inefficient to use

ñ but, same advances in computing change this

Privacy-Preserving Data Mining

Allow multiple data holders to collaborate to compute important (e.g., security-related) information while protecting the privacy of other information.

Particularly relevant now, with increasing focus on security even at the expense of some privacy.

Advantages of privacy protection

- protection of personal information
- protection of proprietary or sensitive information
- fosters collaboration between different data owners (since they may be more willing to collaborate if they need not reveal their information)

Privacy Tradeoffs?

- ï Privacy vs. security: maybe, but doesnít mean giving up one gets the other (who is this person? is this a dangerous person?)

- ï Privacy vs. usability: reasonable defaults, easy and extensive customizations, visualization tools

Tradeoffs are to cost or power, rather than inherent conflict with privacy.

Privacy/Security Tradeoff?

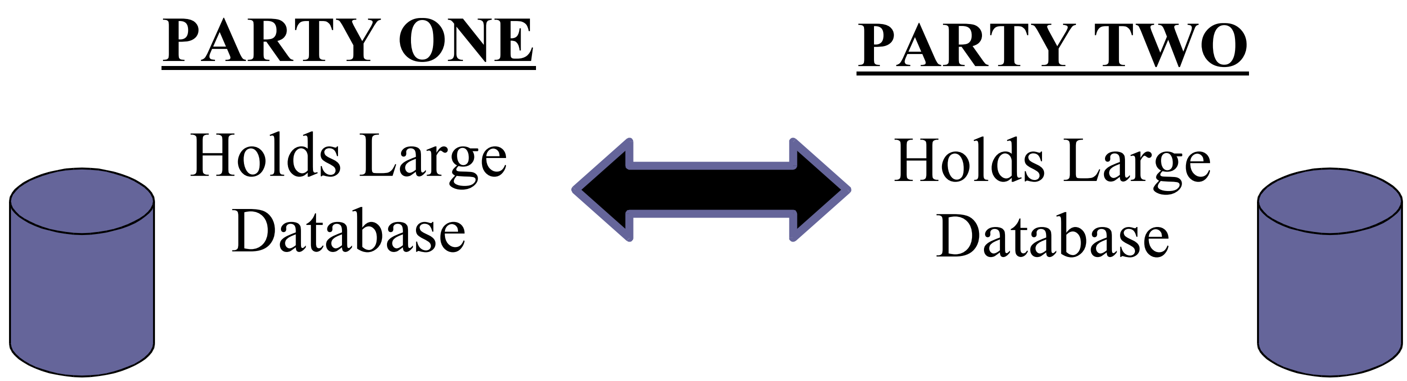
- Claim: No inherent tradeoff between security and privacy, though the cost of having both may be significant.

- Experimentally evaluate the practical feasibility of strong (cryptographic) privacy-preserving solutions.

Examples

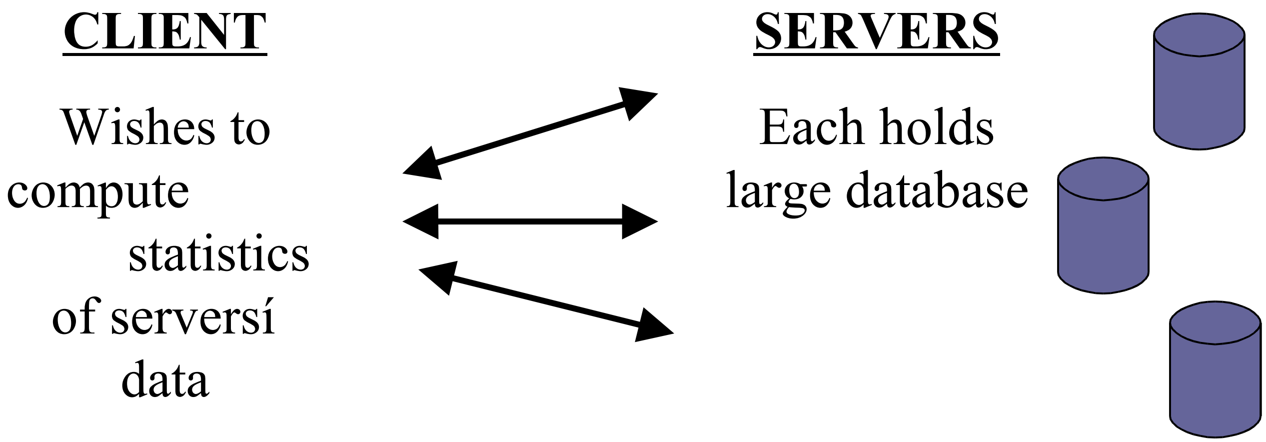
- Privacy-preserving computation of decision trees [LP00]
- Secure computation of approximate Hamming distance of two large data sets [FIMNSW01]
- Privacy-protecting statistical analysis [CIKRRW01]
- Selective private function evaluation [CIKRRW01]

Similarity of Two Data Sets



- Parties can efficiently and privately determine whether their data sets are similar
- Current measure of similarity is approximate Hamming distance [FIMNSW01]
- Securing other measures is topic for future research

Privacy-Protecting Statistics [CIKRRW01]



¡ Parties communicate using cryptographic protocols designed so that:

ñ Client learns desired statistics, but learns nothing else about data (including individual values or partial computations for each database)

ñ Servers do not learn which fields are queried, or any information about other serversí data

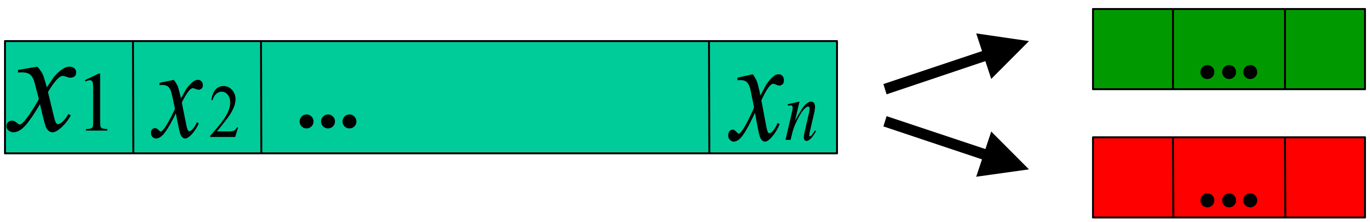
ñ Computation and communication are very efficient

Privacy Concerns

- Protect clients from revealing type of sample population, type of specific data used
- Protect database owners from revealing unnecessary information or providing a higher quality of service than paid for
- Protect individuals from large-scale dispersal of their personal information

Privacy-Protecting Statistics (single DB)

- Database contains public information (e.g. zip code) and private information (e.g. income):



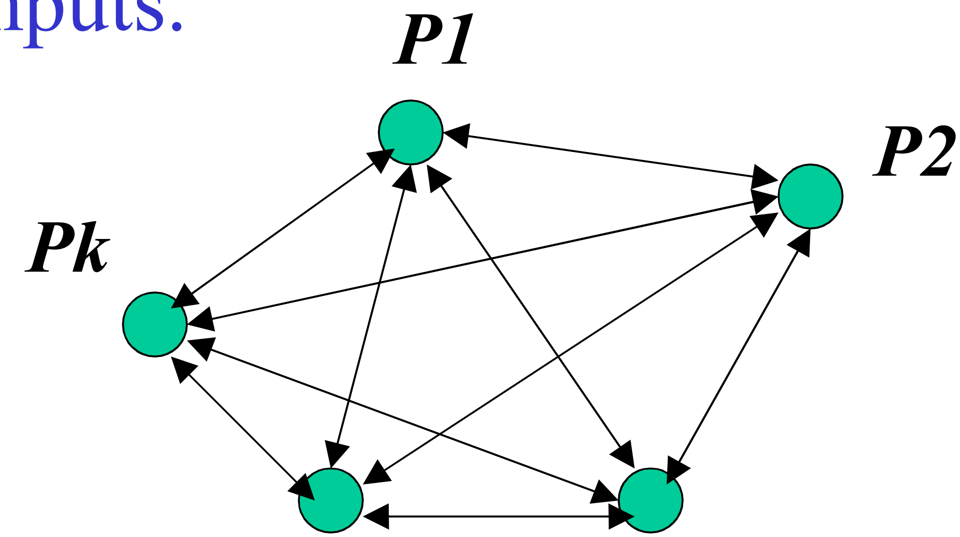
- Client wants to compute statistics on private data, of subset selected by public data. Doesnít want to reveal selection criteria or private values used.
- Database wants to reveal only outcome, not personal data.

Non-Private and Inefficient Solutions

- Database sends client entire database (violates database privacy)
- For sample size *m*, use SPIR to learn *m* values (violates database privacy)
- Client sends selections to database, database does computation (violates client privacy, doesnít work for multiple databases)
- general secure multiparty computation (not efficient for large databases)

Secure Multiparty Computation

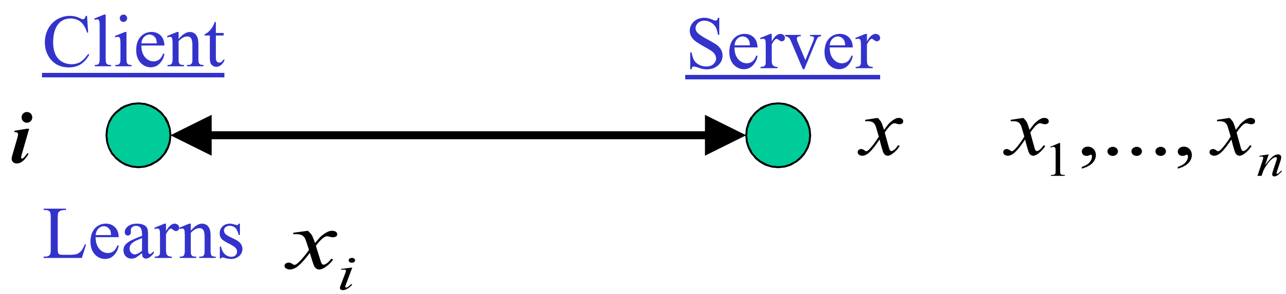
- Allows k players to privately compute a function f of their inputs.



- Overhead is polynomial in size of inputs and complexity of f [Yao, GMW, BGW, CCD, ...]

Symmetric Private Information Retrieval

- Allows client with input i to interact with database server with input x to learn (only) x_i



- Overhead is polylogarithmic in size of database x [KO,CMS,GIKM]

Homomorphic Encryption

- Certain computations on encrypted messages correspond to other computations on the cleartext messages.

- For additive homomorphic encryption,

  - $\boldsymbol{E(m_1)}$ **ï** $\boldsymbol{E(m_2) = E\,(m_1 + m_2)}$

  - also implies $\boldsymbol{E(m)^x = E(mx)}$

Privacy-Protecting Statistics Protocol

- To learn mean and variance: enough to learn sum and sum of squares.

- Server stores:

x_1	x_2	…	x_n
z_1	z_2	…	z_n

$(z_i \quad x_i^2)$

and responds to queries from both

- efficient protocol for sum $\rightarrow$ efficient protocol for mean and variance

Weighted Sum

Client wants to compute selected linear combination of m items: ${}_{j\ \ 1}^{m} w_j x_{i_j}$

<u>Client</u> <u>Server</u>

Homomorphic encryption E, D

$$\begin{array}{lll} & w_j & \text{if } i \quad i_j \\ {}_i & 0 & \text{o/w} \end{array}$$

$E(\ _1), \ldots, E(\ _n)$ $\longrightarrow$

computes

$$v \qquad {}_{i\ \ 1}^{n} (E(\ _i)^{x_i})$$

$$E(\ {}_{i\ \ 1}^{n} \ \ {}_i x_i)$$

$\xleftarrow{\quad v \quad}$

decrypts to obtain

$${}_{i\ \ 1}^{n} \ \ {}_i x_i \qquad\qquad {}_{j\ \ 1}^{m} w_j x_{i_j}$$

Efficiency

- Linear communication and computation (feasible in many cases)
- If n is large and m is small, would like to do better

Selective Private Function Evaluation

- Allows client to privately compute a function f over m inputs $x_{i_1}, \quad , x_{i_m}$
- client learns only $f(x_{i_1}, \quad , x_{i_m})$
- server does not learn $i_1, \dots, i_m$

Unlike general secure multiparty computation, we want communication complexity to depend on m, not n. (More accurately, polynomial in m, polylogarithmic in n).

Security Properties

- Correctness: If client and server follow the protocol, clientís output is correct.
- Client privacy: malicious server does not learn clientís input selection.
- Database privacy:
  - weak: malicious client learns no more than output of some m-input function g
  - strong: malicious client learns no more than output of specified function f

Solutions based on MPC

- Input selection phase:
  - server obtains blinded version of each x_{i_j}
- Function evaluation phase
  - client and server use MPC to compute f on the m blinded items

Input selection phase

Client

Server

Homomorphic encryption ***D,E***

Computes encrypted database

$E(x_1)$	...	$E(x_n)$

Retrieves $E(x_{i_1}),\ldots,E(x_{i_m})$ using SPIR

SPIR(m,n), ***E***

Picks random $c_1,\ldots,c_m$

computes $E(x_{i_j} \quad c_j)$

$E(x_{i_j} \quad c_j)$

Decrypts received values:

$s_j \quad x_{i_j} \quad c_j$

Function Evaluation Phase

- Client has $c \quad c_1,\ldots,c_m$
- Server has $s \quad s_1,\ldots,s_m \qquad s_j \quad x_{i_j} \quad c_j$

Use MPC to compute:

$$g(c,s) \quad f(s \quad c) \quad f(x_1,\ldots,x_m)$$

- Total communication cost polylogarithmic in n, polynomial in m, $|f|$

Distributed Databases

- Same approach works to compute function over distributed databases.

  - Input selection phase done in parallel with each database server

  - Function evaluation phase done as single MPC

  - only final outcome is revealed to client.

Performance

	Complexity	*Security*
1	m SPIR(n,1,k) + O(k|f|)	Strong
2	m SPIR(n,1,1) + MPC(m,|f|)	Weak
3	SPIR(n,m,log n) + MPC(m,|f|) + km^2	Weak
4	SPIR(n,m,k) + MPC(m,|f|)	Honest client only

→ Current experimentation to understand whether these methods are efficient in real-world settings.

Conclusions

- Privacy is in danger, but some important progress has been made.
- Important challenges ahead:
  - Usable privacy solutions
  - Sensor data
  - better use of hybrid approach: decide what can safely be disclosed, use cryptographic protocols to protect critical information, weaker and more efficient solutions for the rest
- Technology, policy, and education must work together.