Privacy Issues in the News

- ChoicePoint: untrustworthy customers
- Paris Hilton/T-Mobile: Poor security practices and/or design: Web site hacked; guessable backup authentication; reliance on callerID
- Bank of America: backup tape stolen from airline
- SAIC: physical theft of desktop computers from the office
- Think Computer: amazingly bad web site design

Lessons

- If data is collected, it can be stolen
- There's no single way to protect against theft
- Fair Information Principles and Practices:

  Security The integrity of the information and the system should be maintained to ensure against loss, destruction, unauthorized access, modification, unauthorized use, or disclosure.

  Accountability The organization collecting and using information can be held responsible for abiding by these principles through:

  – Enforcement and/or

  – Redress.

The Final Paper

- About 10 pages
- Think of it as a take-home exam where you get to pick the question
- *Don't* just summarize a single paper; that's an ordinary homework assignment
- A good choice would be an in-depth analysis of some topic we covered in class (or could have covered in class).
  Either a technical topic or a legal/policy topic is fine
- Follow usual academic practices for your bibliography
- You're *strongly* encouraged to discuss your topic with me before starting
- Note: with permission, I'd like to make these papers available on the web site.

Experimental Analysis of Privacy-Preserving Statistics Computation

- Possible to do *selective private function evaluation*
- Example: private calculation of the sum of a subset of database items
- Use *homomorphic encryption*: $E(a) \cdot E(B) = E(A + b)$
- If I_i are element index selectors (0 or 1) and x_i are the values, the client sends $E(I_1), E(I_2), \ldots, E(I_n)$; the server computes and sends back

$$\prod_{i=1}^{n} E(I_i)^{x_i} = E\left(\sum_{i=1}^{n} I_i x_i\right)$$

- If the server is insufficiently paranoid, there's a security flaw lurking here. . .

Performance

- Paillier encryption is expensive
- Note Fig. 2: computation time is measured in *minutes*
- Various optimizations possible: pipelining, precomputation of encrypted index selectors, distributed processing

Privacy and DRM

- Technologies: OS features, rights management language, encryption, digital signatures, fingerprinting
- General principle: limit data collection to the precise purposes necessary
- Some privacy threats independent of DRM
- Tieing content to user poses serious risks

Crypto Alone is Insufficient

- Academic models don't match commercial reality
- Complex systems are insecure and hard to use
- Must interoperate with legacy gear
- Mechanisms are too expensive
- Metcalfe's law: no critical mass of privacy-enhanced systems
- Usage data needed for compensation

Principles

- Follow Fair Information Practices
- Use identity-based authentication only as necessary
- Create separate databases for different roles, linked by pseudonyms
- Client-side aggregation
- Use technologies such as anonymizers, P3P, etc.

Privacy-Preserving Data Mining Using Multi-Group Randomized Response Techniques

- Let users randomize randomize their data: with probability θ, tell the truth; with probability $1 - \theta$, lie
- Randomization done by groups of data items
- Select group size to balance privacy and accuracy — if group size is very large, outside information on one item in the group can compromise the entire group. But the goal of datamining is to create linkages; you need correlated answers
- Another tradeoff: as θ approaches .5, accuracy decreases