PURDUE
UNIVERSITY

Is Privacy Still an Issue for Data Mining?

Chris Clifton

11 October, 2007

CERIAS
Center for Education and Research in Information Assurance and Security

Indiana Center for Database Systems

Privacy-Preserving Data Mining: History

2000: First PPDM papers
- Srikant&Agrawal: Perturbation
- Lindell&Pinkas: Secure Multiparty Computation

Both assumed horizontal partitioning of data

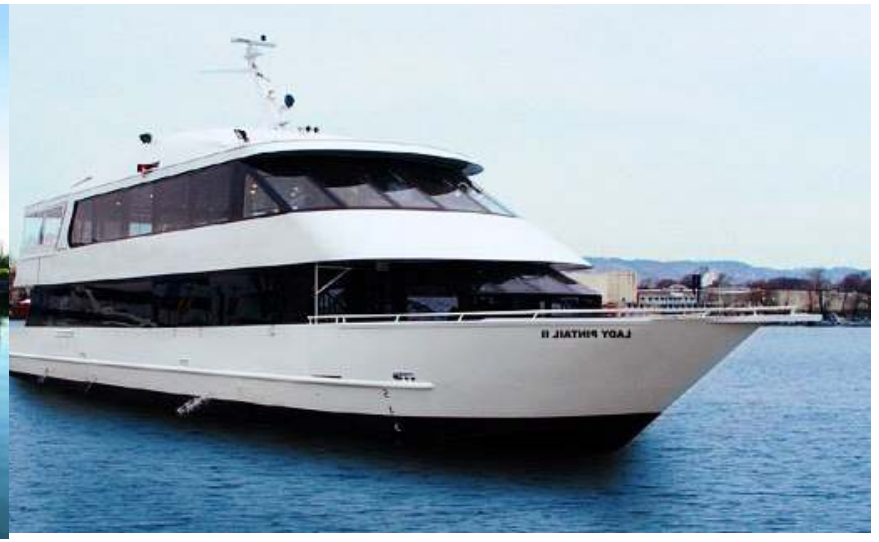
2002: Previous NGDM
- First solution for vertical partitioning
- First workshop on privacy-preserving data mining

2003:
- ICDM Best Paper showed issues with multivariate perturbation

Today: Many solutions
- SMC, rotation, perturbation
- Implementations
- No practice?

When Does this Hit the Mainstream?

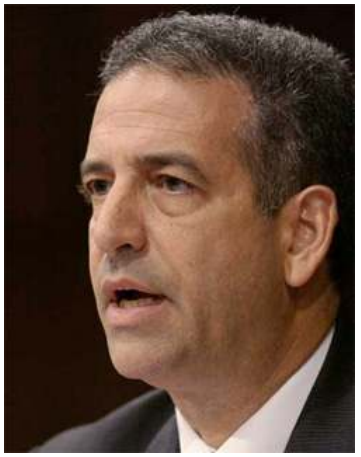
- Data Mining moves FAST
  - VLDB'94 – Fast association rule mining
  - Intelligent Miner for Data – 1995?
- Has PPDM missed the boat?



What does PPDM Need?

- Understand the problem
  - What is privacy?
  - What is the problem with data mining?

Did it go away when the Data Mining Moratorium Act of 2003 died?



- Find a market for the technology
  - Privacy is good
  - But confidentiality pays

What is Privacy?

- It's all about Individually Identifiable Data
  - Standard in nearly all privacy laws
  - But not yet clearly defined

Ongoing research in anonymity

- Data Mining developments exacerbate the problem
  - Text mining
  - Social networks
  - Multirelational data mining

New research challenges!

Alternate Privacy Notions

- Range / approximate value
- True encryption / multiple-message indistinguishability
  - Probing
- Plausible deniability
  - Libel legal standards
- Possible worlds scenarios
  - Probabilistic possible worlds

- Threat models
  - Identity theft
  - Blackmail / ruin political campaign
  - Embarrassment
  - Trust
  - Legal
  - Law enforcement / government

Kevin Du

- Understanding privacy
  - Different techniques use different ways to quantify privacy
  - No way to compare
  - What us unified notion of privacy?
- Threat models
  - Identity theft
  - Blackmail / ruin political campaign
  - Embarrassment
  - Trust
  - Legal
  - Law enforcement / government
- Annie Anton?

The Real Problem: Misuse

- Misuse doesn't require data mining
  - High profile cases from disclosure of raw data, not data mining
  - Is data mining a privacy "red herring"?
- Problem: Data Mining is *why* the data is there to be misused
  - Example: CardSystems saved data for analysis
  - Without data mining, no need for data

  Privacy-Preserving Data Mining can help!

Marketing PPDM as Misuse Protection

- Reduced Risk
  - No data warehouse to be protected
    - Cost savings to offset PPDM cost
  - Lowered risk of disclosure
    - Lower cost of handling disclosure
- Better data → better data mining results
  - Studies show people willing to give better data if privacy protected

Misunderstanding Data Mining can Lead to Misuse

- Data Mining Reporting Act of 2007:

  An assessment of the efficacy or likely efficacy of the data mining activity in providing accurate information consistent with and valuable to the stated goals and plans for the use or development of the data mining activity.

- Research Agenda
  - Confidence bounds
    - On particular prediction, not average
  - Limits on learning

Marketing PPDM as Collaboration Technology

- Work in Secure Supply Chain gaining traction
  - Optimize supply chain without losing competitive advantage
- Shared model development using confidential data
- True “Need to know”
  - Share knowledge, not data
  - Prove need without disclosing reasons

Challenges with Secure Multiparty Communication for Collaboration

- Most work under semi-honest model
  - If you trust your partners, why not just share data?
- Extension to malicious model expensive
  - And still not enough
- Other models
  - Incentive-compatible
  - Auditable

Next Generation of Privacy-Preserving Data Mining

- Understanding Privacy
  - Beyond “Individually Identifiable Data”
- Research supporting profitable use
  - Controlling disclosure risk/cost
  - Collaboration without Disclosure
  - Incentives
- Understanding data mining benefit
  - Limits on learning
  - Confidence in outcomes