Lay of the Land: Legal, Moral, and Historical reasons why Privacy Preserving Data Mining is Important

Chris Clifton

clifton@cs.purdue.edu

www.cs.purdue.edu/people/clifton

Privacy and Security Constraints

- Individual Privacy
  - Nobody should know more about any entity after the data mining than they did before
  - Approaches: Data Obfuscation, Value swapping
- Organization Privacy
  - Protect knowledge about a collection of entities
    - Individual entity values may be known to all parties
    - Which entities are at which site may be secret

Privacy constraints don't prevent data mining

- Goal of data mining is summary results
  - Association rules
  - Classifiers
  - Clusters
- The results alone need not violate privacy
  - Contain no individually identifiable values
  - Reflect overall results, not individual organizations

The problem is computing the results without access to the data!

Privacy-Preserving Data Mining: Who?

- Government / public agencies. Example:
  - The Centers for Disease Control want to identify disease outbreaks
  - Insurance companies have data on disease incidents, seriousness, patient background, etc.
  - But can/should they release this information?
- Industry Collaborations / Trade Groups. Example:
  - An industry trade group may want to identify best practices to help members
  - But some practices are trade secrets
  - How do we provide "commodity" results to all (Manufacturing using chemical supplies from supplier X have high failure rates), while still preserving secrets (manufacturing process Y gives low failure rates)?

Privacy-Preserving Data Mining: Who?

- Multinational Corporations
  - A company would like to mine its data for globally valid results
  - But national laws may prevent transborder data sharing
- Public use of private data
  - Data mining enables research studies of large populations
  - But these populations are reluctant to release personal information

Outline

- Privacy and Security Constraints
  - Types: Individual, collection, result limitation
  - Sources: Regulatory, Contractual, Secrecy
- Classes of solutions
  - Data obfuscation
  - Summarization
  - Data separation

Individual Privacy: Protect the "record"

- Individual item in database must not be disclosed
- Not necessarily a person
  - Information about a corporation
  - Transaction record
- Disclosure of parts of record may be allowed
  - Individually identifiable information

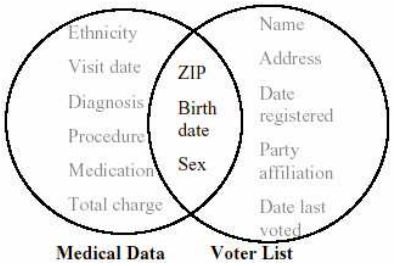
Individually Identifiable Information

- Data that can't be traced to an individual not viewed as private
  - Remove "identifiers"
- But can we ensure it can't be traced?
  - Candidate Key in non-identifier information
  - Unique values for some individuals

Data Mining enables such tracing!

Re-identifying "anonymous" data (Sweeney '01)

- 37 US states mandate collection of information
- She purchased the voter registration list for Cambridge Massachusetts
  - 54,805 people
- 69% unique on postal code and birth date
- 87% US-wide with all three



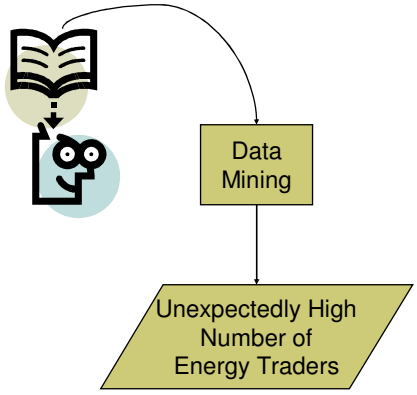
- Solution: *k*-anonymity
  - Any combination of values appears at least *k* times
- Developed systems that guarantee k-anonymity
  - Minimize distortion of results

Collection Privacy

- Disclosure of individual data may be okay
  - Telephone book
  - De-identified records
- Releasing the whole collection may cause problems
  - Trade secrets – corporate plans
  - Rules that reveal knowledge about the holder of data

Collection Privacy Example: Corporate Phone Book

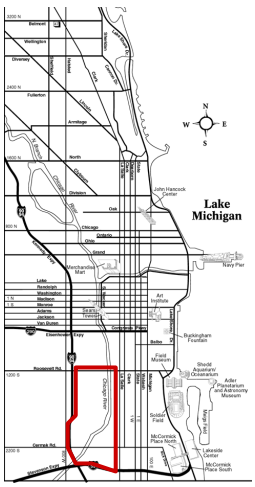
- Telephone Directory discloses how to contact an individual
  - *Intended use*
- Data Mining can find more
  - Relative sizes of departments
  - *Use to predict corporate plans?*
- Possible Solution: Obfuscation
  - Fake entries in phone book
  - *Doesn't prevent intended use*
- Key: Define Intended Use
  - *Not always easy!*



Restrictions on Results

- Use of Call Records for Fraud Detection vs. Marketing
  - FCC § 222(c)(1) restricted use of individually identifiable information
    Until overturned by US Appeals Court
  - 222(d)(2) allows use for fraud detection
- Mortgage Redlining
  - Racial discrimination in home loans prohibited in US
  - Banks drew lines around high risk neighborhoods!!!
  - These were often minority neighborhoods
  - Result: Discrimination (redlining outlawed)

What about data mining that "singles out" minorities?



Sources of Constraints

- Regulatory requirements
- Contractual constraints
  - Posted privacy policy
  - Corporate agreements
- Secrecy concerns
  - Secrets whose release could jeopardize plans
  - Public Relations – “bad press”

Regulatory Constraints: Privacy Rules

- Primarily national laws
  - European Union
  - US HIPAA rules (www.hipaadvisory.com)
  - Many others: (www.privacyexchange.org)
- Often control transborder use of data
- Focus on intent
  - Limited guidance on implementation

European Union Data Protection Directives

- Directive 94/46/EC
  - Passed European Parliament 24 October 1995
  - Goal is to ensure free flow of information
    - *Must preserve privacy needs of member states*
  - Effective October 1998
- Effect
  - Provides guidelines for member state legislation
    - Not directly enforceable
  - Forbids sharing data with states that don't protect privacy
    - Non-member state must provide adequate protection,
    - Sharing must be for "allowed use", or
    - Contracts ensure adequate protection
  - US "Safe Harbor" rules provide means of sharing (July 2000)
    - Adequate protection
    - But voluntary compliance
- Enforcement is happening
  - Microsoft under investigation for Passport (May 2002)
  - Already fined by Spanish Authorities (2001)

EU 95/46/EC: Meeting the Rules

- Personal data is any information that can be traced directly *or indirectly* to a specific person
- Use allowed if:
  - Unambiguous consent given
  - Required to perform contract with subject
  - Legally required
  - Necessary to protect vital interests of subject
  - In the public interest, or
  - Necessary for legitimate interests of processor and doesn't violate privacy
- Some uses specifically proscribed
  - Can't reveal racial/ethnic origin, political/religious beliefs, trade union membership, health/sex life
- Must make data available to subject
  - Allowed to object to such use
  - Must give advance notice / right to refuse direct marketing use
- Limits use for automated decisions
  - Onus on processor to show use is legitimate

europa.eu.int/comm/internal_market/en/dataprot/law

US Healthcare Information Portability and Accountability Act (HIPAA)

- Governs use of patient information
  - Goal is to protect the patient
  - Basic idea: Disclosure okay if anonymity preserved
- Regulations focus on outcome
  - A covered entity may not use or disclose protected health information, except as permitted or required…
    - To individual
    - For treatment (generally requires consent)
    - To public health / legal authorities
  - Use permitted where "there is no reasonable basis to believe that the information can be used to identify an individual"
- Safe Harbor Rules
  - Data presumed not identifiable if 19 identifiers removed (§ 164.514(b)(2)), e.g.:
    - Name, location smaller than 3 digit postal code, dates finer than year, identifying numbers
  - Shown not to be sufficient (Sweeney)
  - Also not necessary

Moral: Get Involved in the Regulatory Process!

Regulatory Constraints: Use of Results

- Patchwork of Regulations
  - US Telecom (Fraud, not marketing)
    - Federal Communications Commission rules
    - Rooted in antitrust law
  - US Mortgage "redlining"
    - Financial regulations
    - Comes from civil rights legislation
- Evaluate on a per-project basis
  - Domain experts should know the rules
  - You'll need the domain experts anyway – ask the right questions

Contractual Limitations

- Web site privacy policies
  - "Contract" between browser and web site
  - Groups support voluntary enforcement
    - TrustE – requires that web site DISCLOSE policy on collection and use of personal information
    - BBBonline
      - posting of an online privacy notice meeting rigorous privacy principles
      - completion of a comprehensive privacy assessment
      - monitoring and review by a trusted organization, and
      - participation in the programs consumer dispute resolution system
    - Unknown legal "teeth"
      - Example of customer information viewed as salable property in court!!!
  - P3P: Supports browser checking of user-specific requirements
    - Internet Explorer 6 – disallow cookies if non-matching privacy policy
    - PrivacyBird – Internet Explorer plug-in from AT&T Research
- Corporate agreements
  - Stronger teeth/enforceability
  - But rarely protect the individual

Secrecy

- Governmental sharing
  - Clear rules on sharing of classified information
  - Often err on the side of caution
    - Touching classified data "taints" everything
    - Prevents sharing that wouldn't disclose classified information
- Corporate secrets
  - Room for cost/benefit tradeoff
  - Authorization often a single office
    - Convince the right person that secrets aren't disclosed and work can proceed
  - Antitrust: Need to be able to show that secrets aren't shared!
- Bad Press
  - Lotus proposed "household marketplace" CD (1990)
    - Contained information on US households from public records
    - Public outcry forced withdrawal
  - Credit agencies maintain public and private information
    - Make money from using information for marketing purposes
  - Key difference? *Personal information isn't disclosed*
    - Credit agencies do the mining
    - "Purchasers" of information don't see public data

Antitrust Example: Airline Pricing

- Airlines share real-time price and availability with reservation systems
  - Eases consumer comparison shopping
  - Gives airlines access to each other's prices

  Ever noticed that all airlines offer the same price?
- Shouldn't this violated price-fixing laws?
  - *It did!*

Antitrust Example: Airline Pricing

- Airlines used to post "notice of proposed pricing"
  - If other airlines matched the change, the prices went up
  - If others kept prices low, proposal withdrawn
  - This violated the law
- Now posted prices effective immediately
  - If prices not matched, airlines return to old pricing
- Prices are still all the same
  - *Why is it legal?*

The Difference: *Need to Know*

- Airline prices easily available
  - Enables comparison shopping
- Airlines can change prices
  - Competition results in lower prices
- *These are needed to give desired consumer benefit*
  - "Notice of proposed pricing" wasn't

Classes of Solutions

- Data Obfuscation
  - Nobody sees the *real* data
- Summarization
  - Only the needed facts are exposed
- Data Separation
  - Data remains with trusted parties

Data Obfuscation

- Goal: Hide the protected information
- Approaches
  - Randomly modify data
  - Swap values between records
  - Controlled modification of data to hide secrets
- Problems
  - Does it really protect the data?
  - Can we learn from the results?

Example: US Census Bureau Public Use Microdata

- US Census Bureau summarizes by census block
  - Minimum 300 people
  - Ranges rather than values
- For research, "complete" data provided for sample populations
  - Identifying information removed
    - Limitation of detail: geographic distinction, continuous → interval
    - Top/bottom coding (eliminate sparse/sensitive values)
  - Swap data values among similar individuals (Moore '96)
    - Eliminates link between potential key and corresponding values
    - If individual determined, sensitive values likely incorrect

    Preserves the privacy of the individuals, as no entity in the data contains actual values for any real individual.
  - Careful swapping preserves multivariate statistics
    - Rank-based: swap similar values (randomly chosen within max distance)

    Preserves dependencies with (provably) high probability
  - Adversary can estimate sensitive values if individual identified

    But data mining results enable this anyway!

Summarization

- Goal: Make only innocuous summaries of data available
- Approaches:
  - Overall collection statistics
  - Limited query functionality
- Problems:
  - Can we deduce data from statistics?
  - Is the information sufficient?

Example: Statistical Queries

- User is allowed to query protected data
  - Queries must use statistical operators that summarize results
    - Example: Summation of total income for a group doesn't disclose individual income
  - Multiple queries can be a problem
    - Request total salary for all employees of a company
    - Request the total salary for all employees but the president
    - Now we know the president's salary
- Query restriction – Identify when a set of queries is safe (Denning '80)
  - *query set overlap control* (Dobkin, Jones, and Lipton '79)
    - Result generated from at least *k* items
    - Items used to generate result have at most *r* items in common with those used for previous queries
    - At least $1+(k-1)/r$ queries needed to compromise data
  - Data perturbation: introducing noise into the original data
  - Output perturbation: leaving the original data intact, but introducing noise into the results

Example: Statistical Queries

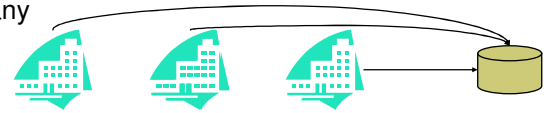
- Problem: Can approximate real values from multiple queries (Palley and Simonoff '87)
  - Create histograms for unprotected independent variables (e.g., job title)
  - Run statistical queries on the protected value (e.g., average salary)
  - Create a synthetic database capturing relationships between the unprotected and protected values
  - Data mining on the synthetic database approximate real values
- Problem with statistical queries is that the adversary creates the queries
  - Such manipulation likely to be obvious in a data mining situation
  - Problem: *Proving* that individual data not released

Data Separation

- Goal: Only trusted parties see the data
- Approaches:
  - Data held by owner/creator
  - Limited release to trusted third party
  - Operations/analysis performed by trusted party
- Problems:
  - Will the trusted party be willing to do the analysis?
  - Do the analysis results disclose private information?

Example: Patient Records

- My health records split among providers
  - Insurance company
  - Pharmacy
  - Doctor
  - Hospital



- Each agrees not to release the data without my consent
- Medical study wants correlations across providers
  - Rules relating complaints/procedures to "unrelated" drugs
- Does this need my consent?
  - *And that of every other patient!*
- It shouldn't!
  - Rules don't disclose my individual data

When do we address these concerns?

- Must articulate that
  - A problem exists
    - There will be problems if we don't worry about privacy
  - We need to know the issues
    - Domain-specific constraints
  - A technical solution is feasible
    - Results valid
    - Constraints (provably) met

What we need to know

- Constraints on release of data
  - Define in terms of Disclosure, not Privacy
  - What can be released, what mustn't
- Ownership/control of data
  - Nobody allowed access to "real" data
  - Data distributed across organizations
    - Horizontally partitioned: Each entity at a separate site
    - Vertically partitioned: Some attributes of each entity at each site
- Desired results: Rules? Classifier? Clusters?

Summary

- Privacy and Security Constraints can be impediments to data mining
  - Problems with access to data
  - Restrictions on sharing
  - Limitations on use of results
- Technical solutions possible
  - Randomizing / swapping data doesn't prevent learning good models
  - We don't need to share data to learn global results
  - When the secrets are in the results and we want to share the data
- *Let's Hear How!*