Last time

- Database Security
  - Data Inference
  - Statistical Inference
  - Controls against Inference
- Multilevel Security Databases
  - Separation
  - Integrity Locks
  - Designs of MLS Databases

This time

- Data Mining
  - Integrity and Availability
  - Privacy and Data Mining
  - Privacy-Preserving Data Mining

Data Mining

- Multilevel databases weren't a commercial success
  - Mainly military clients, finding all possible inferences is NP-complete
- However, the combination of (sensitive) information, stored in multiple (maybe huge) databases, as done for data mining, raises similar concerns and has gotten lots of attention recently
- So far, a single entity has been in control of some data
  - Knows what kind of data is available
  - Who has accessed it (ignoring side channels)
- No longer the case in data mining, data miners actively gather additional data from third parties

Data Mining (cont.)

- Data mining tries to automatically find interesting patterns in data using a plethora of technologies
  - Statistics, machine learning, pattern recognition,…
  - Still need human to judge whether pattern makes sense (causality vs. coincidence)
- Data mining can be useful for security purposes
  - Learning information about an intrusion from logs

Security Problems of Data Mining

- Confidentiality
  - Derivation of sensitive information
- Integrity
  - Mistakes in data
- Availability
  - (In)compatibility of different databases

Confidentiality

- Data mining can reveal sensitive information about humans (see later) and companies
- In 2000, the U.S. National Highway Traffic Safety Administration combined data about Ford vehicles with data about Firestone tires and become aware of a problem with the Ford Explorer and its Firestone tires
  - Problem started to occur in 1995, and each company individually had some evidence of the problem
  - However, data about product quality is sensitive, which makes sharing it with other companies difficult
- Supermarket can use loyalty cards to learn who buys what kind of products and sell this data, maybe to manufacturers' competitors

Data Correctness and Integrity

- Data in a database might be wrong
  - E.g., input or translations errors
- Mistakes in data can lead to wrong conclusions by data miners, which can negatively impact individuals
  - From receiving irrelevant mail to being denied to fly
- Privacy calls for the right of individuals to correct mistakes in stored data about them
  - However, this is difficult if data is shared widely or if there is no formal procedure for making corrections
- In addition to false positives, there can also be false negatives: don't blindly trust data mining applications

Availability

- Mined databases are often created by different organizations
  - Different primary keys, different attribute semantics,…
    - Is attribute “name” last name, first name, or both?
    - US or Canadian dollars?
- Makes combination of databases difficult
- Must distinguish between inability to combine data and inability to find correlation

Privacy and Data Mining

- Data mining might reveal sensitive information about individuals, based on the aggregation and inference techniques discussed earlier
- Avoiding these privacy violations is active research
- Data collection and mining is done by private companies
  - Privacy laws (e.g., Canada's PIPEDA or U.S.' HIPAA) control collection, use, and disclosure of this data
  - Together with PETs
- But also by governments
  - Programs tend to be secretive, no clear procedures
  - Phone tapping in U.S., no-fly lists in U.S. and Canada

Privacy-Preserving Data Mining

- Anonymize data records before making them available
  - E.g., strip names, addresses, phone numbers
  - Unfortunately, such simple anonymization might not be sufficient
- August 6, 2006: AOL released 20 million search queries from 658,000 users
- To protect users' anonymity, AOL assigned a random number to each user
  - 4417749 "numb fingers"
  - 4417749 "landscapers in Lilburn, Ga"
  - 17556639 "how to kill your wife"
- August 9: New York Times article re-identified user 4417749
  - Thelma Arnold, 62-year old widow from Lilburn, GA

Another Example (by L. Sweeney)

- 87% of U.S. population can be uniquely identified based on person's ZIP code, gender, and date of birth
- Massachusetts' Group Insurance Commission released anonymized health records
- Records left away individuals' name, but gave their ZIP code, gender, and date of birth (and health information, of course)
- Massachusetts's voter registration lists contain these three items plus individuals' names and are publicly available
- Enables re-identification by linking

k-Anonymity

- Ensure that for each released record, there are at least *k-1* other released records from which record cannot be distinguished
- For health-records example, release a record only if there are *k-1* other records that have same ZIP code, gender, and date of birth
  - Assumption: there is only one record for each individual
- Because of the 87% number, this won't return many records, need some pre-processing of records
  - Strip one of { ZIP code, gender, date of birth } from all records
  - Reduce granularity of ZIP code or date of birth

Discussion

- In health-records example, the attributes ZIP code, gender, and date of birth form a “quasi-identifier”
- Determining which attributes are part of the quasi-identifier can be difficult
  - Should health information be part of it?
  - Some diseases are rare and could be used for re-identification
  - However, including them is bad for precision
- Quasi-identifier should be chosen such that released records do not allow any re-identification based on any additional data that attacker might have
  - Clearly we don’t know all this data

Value Swapping

- Data perturbation based on swapping values of some (not all!) data fields for a subset of the records
  - E.g., swap addresses in subset of records
- Any linking done on the released records can no longer considered to be necessarily true
- Trade off between privacy and accuracy
- Statistically speaking, value swapping will make strong correlations less strong and weak correlations might go away entirely

Adding Noise

- Data perturbation based on adding small positive or negative error to each value
- Given distribution of data after perturbation and the distribution of added errors, distribution of underlying data can be determined
  - But not its actual values
- Protects privacy without sacrificing accuracy

Recap

- Data Mining
  - Integrity and Availability
  - Privacy and Data Mining
  - Privacy-Preserving Data Mining

Next time

- Administering Security
  - Security planning
  - Risk Analysis