COMP3420: Advanced Databases and Data Mining

Privacy-preserving data mining and data sharing

Lecture outline

- Privacy and confidentiality
  - Real-world scenarios
  - Re-identification
- Goals of privacy-preserving data mining
- Privacy-preserving data mining techniques
  - Data modifications and obfuscations
  - Summarisation
  - Data separation
  - Secure multi-party computations
- Privacy-preserving data sharing and linking
  - Blindfolded record linkage (Churches and Christen, 2004)

Privacy and confidentiality

- Privacy of individuals
  - Identifying information: Names, addresses, telephone numbers, dates-of-birth, driver licenses, racial/ethnic origin, family histories, political and religious beliefs, trade union memberships, health, sexual orientation, income, ...
  - Some of this information is publicly available, other is not
  - Individuals are happy to share some information with others (to various degrees)
- Confidentiality in organisations
  - Trade secrets, corporate plans, financial status, planned collaborations, ...
  - Collect and store information about many individuals (customers, patients, employees)
- Conflict between individual privacy and information collected by organisations
  - Personal details are valuable (phishing, identity fraud)
  - Privacy-preserving data mining and data sharing are mainly of importance when applied between organisations (businesses, government agencies)

Protect individual privacy

- Individual items (records) in a database must not be disclosed
  - Not only personal information
  - Confidential information about a corporation
  - For example, transaction records (bank account, credit card, phone call, etc.)
- Disclosing parts of a record might be possible
  - Like name or address only (but if data source is known even this can be problematic)
  - For example, a cancer register, HIV database, etc.
- Remove *identifier* so data cannot be traced to an individual
  - Otherwise data is not private anymore
  - But how can we make sure data can't be traced?

Real world scenarios

(based on slides by Chris Clifton, http://www.cs.purdue.edu/people/clifton)

- Multi-national corporation
  - Wants to mine its data from different countries to get global results
  - Some national laws may prevent sending some data to other countries
- Industry collaboration
  - Industry group wants to find best practices (some might be trade secrets)
  - A business might not be willing to participate out of fear it will be identified as conducting bad practice or not complying with rules compared to others
- Analysis of disease outbreaks
  - Government health departments want to analyse such topics
  - Relevant data (patient backgrounds, etc.) held by private health insurers and other organisations (can/should they release such data?)

More real world scenarios (data sharing)

- Data sharing between companies
  - Two pharmaceutical companies are interested in collaborating on the expensive development of new drugs
  - Companies wish to identify how much overlap of confidential research data there is in their databases (but without having to reveal any confidential data to each other)
  - Techniques are needed that allow sharing of large amounts of data in such a way that similar data items are found (and revealed to both companies) while all other data is kept confidential
- Geocoding cancer register addresses
  - Limited resources prohibit the register to invest in an in-house geocoding system
  - Alternative: The register has to send their addresses to an external geocoding service/company (but regulatory framework might prohibit this)
  - Complete trust needed in the capabilities of the external geocoding service to conduct accurate matching, and to properly destroy the register's address data afterwards

Re-identification

- *L. Sweeney* (Computational Disclosure Control, 2001)
  - Voter registration list for Cambridge (MA, USA) with 54,805 people: 69% were unique on postal code (5-digit ZIP code) and date of birth
  - 87% in whole of population of USA (216 of 248 million) were unique on: ZIP, date of birth and gender!
  - Having these three attributes allows linking with other data sets (quasi-identifying information)
- *R. Chaytor* (Privacy Advisor, SIGIR 2006)
  - A patient living in a celebrity's neighbourhood
  - Statistical data (e.g. from ABS – Australian Bureau of Statistics) says one male, between 30 and 40, has HIV in this neighbourhood (ABS mesh block: approx. 50 households)
  - A journalist offers money in exchange of some patients medical details
  - How much can the patient reveal without disclosing the identity of his/her neighbours?

Goals of privacy-preserving data mining

- Privacy and confidentiality issues normally do not prevent data mining
  - Aim is often summary results (clusters, classes, frequent rules, etc.)
  - Results often do not violate privacy constraints (they contain no identifying information)
  - But, certain rules or classification outcomes might compromise confidentiality
  - But: Certain techniques (e.g. outlier detection) aim to find specific records (fraudulent customers, potential terrorists, etc.)
  - Also, often detailed records are required by data mining algorithms
- The problem is: How to conduct data mining without accessing the identifying data
  - Legislation and regulations might prohibit access to data (especially between organisations or countries)
- Main aim is to develop algorithms to modify the original data in some way, so that private data and private knowledge remain private even after the mining process

Privacy-preserving data mining techniques (1)

- Many approaches to preserve privacy while doing data mining
  - Distributed data: Either *horizontally* (different records reside in different locations) or *vertically* (values for different attributes reside in different locations)
- Data modifications and obfuscation
  - Perturbation (changing attribute values, e.g. by specific new values -- mean, average - or randomly)
  - Blocking (replacement of values with for example a '?')
  - Aggregation (merging several values into a coarser category, similar to concept hierarchies)
  - Swapping (interchanging values of individual records)
  - Sampling (only using a portion of the original data for mining)
- Problems: Does this really protect privacy? Still good quality data mining results?

Privacy-preserving data mining techniques (2)

- Data summarisation
  - Only the needed facts are released at a level that prohibits identification of individuals
  - Provide overall data collection statistics
  - Limit functionality of queries to underlying databases (statistical queries)
  - Possible approach: *k*-anonymity (*L. Sweeney*, 2001): any combination of values appears at least *k* times (also has problems!)

- Problems
  - Can identifying details still be deducted from a series of such queries?
  - Is the information accessible sufficient to perform the desired data mining task?

Privacy-preserving data mining techniques (3)

- Data separation
  - Original data held by data creator or data owner
  - Private data is only given to a trusted third party
  - All communication is done using encryption
  - Only limited release of necessary data
  - Data analysis and mining done by trusted third party
- Problems
  - This approach secures the data sets, but not the potential results!
  - Mining results can still disclose identifying or confidential information
  - Can and will the trusted third party do the analysis?
  - If several parties involved, potential of collusion by two parties
- Privacy-preserving approaches for association rule mining, classification, clustering, etc. have been developed

Secure multi-party computation

- Aim: To calculate a function so that no party learns the values of the other parties, but all learn the final result
  - Assuming semi-honest behaviour: Parties follow the protocol, but they might keep intermediate results

- Example: Simple secure summation protocol (*Alan F. Karr*, 2005)
  - Consider $K > 2$ cooperating parties (businesses, hospitals, etc.)
  - Aim: to compute $v = \Sigma^{k}_{j=1} v_j$ so that no party learns other parties v_j
  - Step 1: Party 1 generates a large random number R, with $R >> v$
  - Step 2: Party 1 sends $(v_1 + R)$ to party 2
  - Step 3: Party 2 adds v_2 to $v_1 + R$ and sends $(v_1 + v_2 + R)$ to party 3 (and so on)
  - Step $K+1$: Party K sends $(v_1 + v_2 + \ldots + v_k + R)$ back to party 1
  - Last step: Party 1 subtracts R and gets final v, which it then sends to all other parties

Privacy-preserving data sharing and linking

- Traditionally, data linkage requires that identified data is being given to the person or institution doing the linkage
  - Privacy of individuals in databases is invaded
  - Consent of individuals involved is needed (impossible for very large data sets)
  - Alternatively, approval from ethics committees

- Invasion of privacy could be avoided (or mitigated) if some method were available to determine which records in two data sets match without revealing any identifying information.

Privacy-preserving record linkage

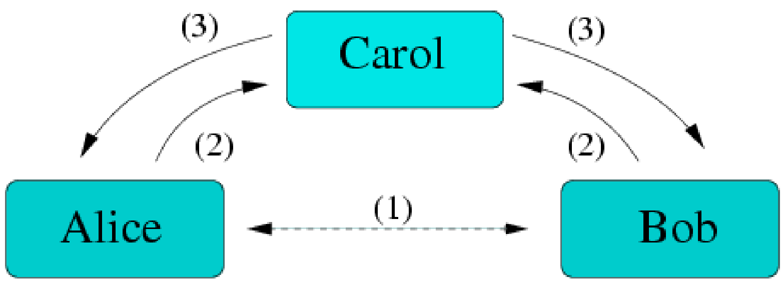
- *Alice* has database *A*
- *Bob* has database *B*
- *Alice* and *Bob* wish to determine whether any of the records in *A* match any of the values in *B*, without revealing the actual values in *A* and *B*
- Easy if only *exact matches* are considered
  - Use one-way message authentication digests (HMAC) based on secure one-way hashing like SHA or MD5
- More complicated if values contain errors or typographical variations
  - Even a single character difference between two strings will result in very different hash values
  - For example: 'peter' → 'd56&!2X2$#', while 'pete' → '8^5evK-)2Q'
  - Requires special encoding algorithms (ongoing research, including here at ANU)

Example Blindfolded record linkage

- A protocol is required which permits the *blind* calculation by a trusted third party (*Carol*) of a more general and robust measure of similarity between pairs of secret strings

- Proposed protocol is based on *q*-grams
  - For example (q=*2,* bigrams): *'peter' -> ['pe','et','te','er']*
  - Convert sub-lists of *q*-gram lists into hash codes

    peter: ['pe','et','te','er'] → 'd5&S!2TX2'
    ['pe','et','te'] → **'8^5eK-7)20Q'**
    ['et','te','er'] → …
    ….
    pete: ['pe','et',te'] → **'8^5eK-7)20Q'**
    ['pe','et'] → …
    ['et','te'] → ….



What now... things to do

- No more lectures this week – **work on assignment 2 Good luck!**

- Last lecture next Tuesday: Course review and exam prep

- Please complete SELT questionnaire and give us feedback, both positive and constructive critical comments – thanks!
  (SELT is completely anonymous)