

## Privacy-Preserving Data Mining and Data Sharing

- Privacy and confidentiality
  - Real world scenarios
  - Re-identification
- Goals of (privacy-preserving) data mining
- Privacy-preserving data mining techniques
  - Data modifications and obfuscation
  - Summarisation
  - Data separation
  - Secure multi-party computations
- Privacy-preserving data sharing and linking

Slide 1 of 18

MATH3346 – Data Mining

## 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, sex life, 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
  - Information about many individuals (customers, patients)
- Privacy-preserving data mining and data sharing mainly of importance when applied between organisations

Slide 2 of 18

MATH3346 – Data Mining

## Protect individual privacy

- Individual items (records) in a database must not be disclosed
  - Not only personal information
  - Confidential information about a corporation
  - Transaction record (bank account, 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)
- 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?

Slide 3 of 18

MATH3346 – Data Mining

## Real world scenarios <sup>1</sup>

- 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 collaborations
  - 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 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 (can/should they release such data?)

<sup>1</sup>Based on slides by Chris Clifton, <http://www.cs.purdue.edu/people/clifton>

Slide 4 of 18

MATH3346 – Data Mining

## More real world scenarios (data sharing)

- 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 (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
- 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

Slide 5 of 18

MATH3346 – Data Mining

## Re-identification

- L. Sweeney (Computational Disclosure Control, 2001)
  - Voter registration list for Cambridge (MA) 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) 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?

Slide 6 of 18

MATH3346 – Data Mining

## Goals of (privacy-preserving) data mining

- Privacy and confidentiality normally don't prevent data mining
  - Aim is often summary results (clusters, classes, frequent rules, etc.)
  - Results often don't violate privacy constraints (they contain no identifying information)
  - But: Certain techniques (e.g. outlier detection) aim to find specific records (fraudulent customers, potential terrorists, etc.)
- 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

Slide 7 of 18

MATH3346 – Data Mining

## 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 results?

Slide 8 of 18

MATH3346 – Data Mining

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

Slide 9 of 18

MATH3346 – Data Mining

## 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 rules, decision trees, clustering, etc. have been developed

Slide 10 of 18

MATH3346 – Data Mining

## Secure multi-party computations

- Aim: To calculate a function involving several parties, 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  $v = \sum_{j=1}^K v_j$  so that no party learns other parties  $v_j$
  - Step 1: Party 1 generates a large random number  $R$
  - 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  $(v_1 + v_2 + \dots + v_k + R)$  back to party 1
  - Last step: Party 1 subtracts  $R$  and gets final  $v$ , which it then sends to all other parties

Slide 11 of 18

MATH3346 – Data Mining

## 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 data sets 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.*

Slide 12 of 18

MATH3346 – Data Mining

## 'Blindfolded record linkage': Methods <sup>2</sup>

- Alice has database **A**, with attributes **A.a**, **A.b**, etc.
- Bob has database **B**, with attributes **B.a**, **B.b**, etc.
- Alice and Bob wish to determine whether any of the values in **A.a** match any of the values in **B.a**, without revealing the actual values in **A.a** and **B.a**
- 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)

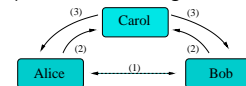
<sup>2</sup>Churches & Christen, PAKDD 2004; see: datamining.anu.edu.au

Slide 13 of 18

MATH3346 – Data Mining

## 'Blindfolded record linkage': Protocol (1)

- 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  $n$ -grams  
For example ( $n = 2$  bigrams): 'peter'  $\rightarrow$  ('pe', 'et', 'te', 'er')
- Protocol step 1
  - Alice and Bob agree on a secret random key
  - They also agree on a secure one-way message authentication algorithm (HMAC)
  - They also agree on a standard of preprocessing strings

Slide 14 of 18

MATH3346 – Data Mining

## 'Blindfolded record linkage': Protocol (2)

- Protocol step 2 (Alice)
  - Alice computes a sorted list of  $n$ -grams for each of her values in **A.a**
  - Next she calculates all possible sub-lists with length larger than 0 (power-set without empty set) For example: 'peter'  $\rightarrow$   
( 'er' ), ( 'et' ), ( 'pe' ), ( 'te' ),  
( 'er','et' ), ( 'er','pe' ), ( 'er','te' ), ( 'et','pe' ), ( 'et','te' ), ( 'pe','te' ),  
( 'er','et','pe' ), ( 'er','et','te' ), ( 'er','pe','te' ), ( 'et','pe','te' ),  
( 'er','et','pe','te' )
  - Then she transforms each sub-list into a secure hash digest (using shared secret key) and stores these in **A.a\_hash\_bigr\_comb** attribute

Slide 15 of 18

MATH3346 – Data Mining

## 'Blindfolded record linkage': Protocol (3)

- Protocol step 2 (Alice, continued)
  - Alice computes encrypted version of the record identifier and stores it in **A.a\_encrypt\_rec\_key**
  - Next she places the number of bigrams of each **A.a\_hash\_bigr\_comb** into **A.a\_hash\_bigr\_comb\_len**
  - She then places the length (total number of bigrams) of each original string into **A.a\_len**
  - Alice then sends the quadruplet [**A.a\_encrypt\_rec\_key**, **A.a\_hash\_bigr\_comb**, **A.a\_hash\_bigr\_comb\_len**, **A.a\_len**] to Carol
- Protocol step 2 (Bob)
  - Bob carries out the same as Alice in step 2 with his **B.a**

Slide 16 of 18

MATH3346 – Data Mining

## 'Blindfolded record linkage': Protocol (4)

- Protocol step 3
  - For each value of **a\_hash\_bigr\_comb** shared by **A** and **B**, for each unique pairing of [**A.a\_encrypt\_rec\_key**, **B.a\_encrypt\_rec\_key**], Carol calculates a *bigram score*

$$\mathbf{bigr\_score} = \frac{2 \cdot \mathbf{A.a\_hash\_bigr\_comb\_len}}{(\mathbf{A.a\_len} + \mathbf{B.a\_len})}$$

- Carol then selects the maximum **bigr\_score** for each pairing [**A.a\_encrypt\_rec\_key**, **B.a\_encrypt\_rec\_key**] and sends these results to Alice and Bob (highest score for each pair of strings from **A.a** and **B.a**)

## Further information

- Privacy Preserving Data Mining Bibliography:  
[http://www.cs.umbc.edu/~kunliu1/research/privacy\\_review.html](http://www.cs.umbc.edu/~kunliu1/research/privacy_review.html)
- Privacy, Security and Data Mining:  
[http://www.cs.ualberta.ca/~oliveira/psdm/psdm\\_index.html](http://www.cs.ualberta.ca/~oliveira/psdm/psdm_index.html)
- Cryptography / Privacy-Preserving Data Mining  
[http://www.adastral.ucl.ac.uk/~helger/crypto/link/data\\_mining/](http://www.adastral.ucl.ac.uk/~helger/crypto/link/data_mining/)