Privacy-preserving data linkage

Part two of the AusDM'08 tutorial on

Privacy preserving data sharing and mining

Peter Christen

Department of Computer Science,
ANU College of Engineering and Computer Science,
The Australian National University,
Canberra ACT 0200, Australia

Contact: **peter.christen@anu.edu.au**

Outline

- Introduction to data linkage
  - Applications, challenges and techniques
  - The data linkage process
- Privacy and confidentiality issues with data linkage
- Data linkage scenarios
- Privacy-preserving matching approaches
  - *Blindfolded data linkage* in more details
- Challenges and research directions
  - Ultimate aim: Automated and secure linking of very large data collections between organisations

What is data linkage

- The process of matching and aggregating records that represent the same entity (such as a patient, a customer, a business, an address, an article, etc.)
  - Also called *data matching*, *entity resolution*, *data scrubbing*, *object identification*, *merge-purge*, etc.
- Challenging if no unique entity identifiers available

  For example, which of these three records refer to the same person?

Dr Smith, Peter	*42 Miller Street 2602 O'Connor*
Pete Smith	*42 Miller St, 2600 Canberra A.C.T.*
P. Smithers	*24 Mill Street; Canberra ACT 2600*

Applications of data linkage

- Health, biomedical and social sciences
  (for epidemiological or longitudinal studies)
- Census, taxation, immigration, and social security
  (for improved data processing and analysis)
- Deduplication of (business mailing) lists
  (to improve data quality and reduce costs)
- Bibliographic databases and online libraries
  (to measure impact - for example for *ERA*)
- Geocode matching ('geocoding') of addresses
  for spatial analysis
- Crime and fraud detection, national security

Data linkage challenges

- Real world data is dirty
  (typographical errors and variations, missing and out-of-date values, different coding schemes, etc.)
- Scalability
  - Comparison of all record pairs has quadratic complexity (however, the maximum number of matches is in the order of the number of records in the databases)
  - Some form of blocking, indexing or filtering required
- No training data in many matching applications
  - No record pairs with known true match status
  - Possible to manually prepare training data (but, how accurate will manual classification be?)

Data linkage techniques

- Deterministic linkage
  - Exact matching (if a *unique identifier* of high quality is available: precise, robust, stable over time)
    Examples: *Medicare*, *ABN* or *Tax file* number (?)
  - Rules based matching (complex to build and maintain)
- Probabilistic linkage
  - Use available (personal) information for matching (like *names*, *addresses*, *dates-of-birth*, etc.)
  - Can be wrong, missing, coded differently, or out-of-date
- Modern approaches
  (based on machine learning, AI, data mining, database, or information retrieval techniques)

Probabilistic data linkage

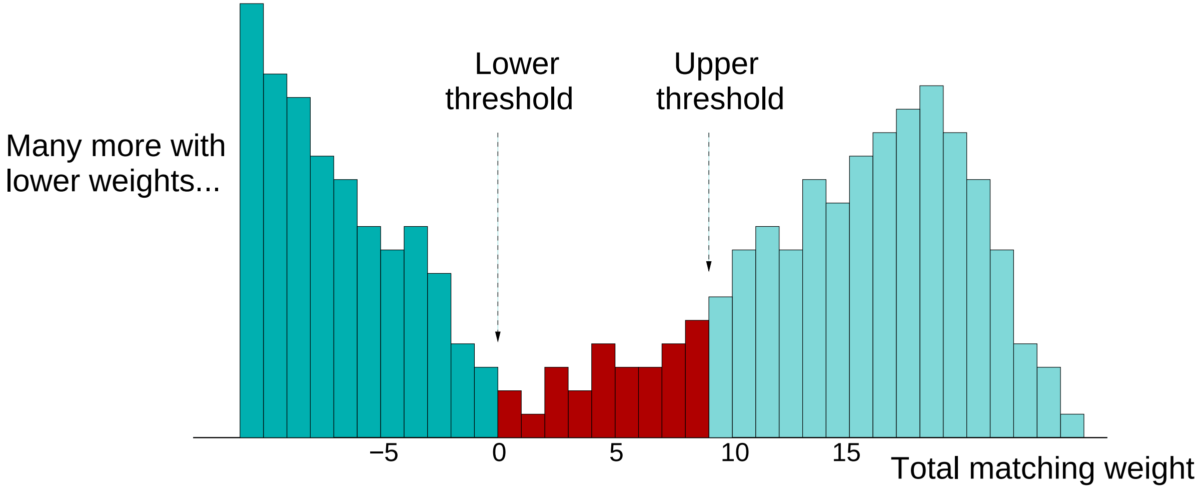
- Computer assisted data linkage goes back as far as the 1950s (based on ad-hoc heuristic methods)
- Basic ideas of probabilistic linkage were introduced by *Newcombe & Kennedy* (1962)
- Theoretical foundation by *Fellegi & Sunter* (1969)
  - Compare common record attributes (or fields)
  - Compute matching weights based on frequency ratios (global or value specific ratios) and error estimates
  - Sum of the matching weights is used to classify a pair of records as *match*, *non-match*, or *possible match*
  - Problems: Estimating errors and threshold values, assumption of independence, and *clerical review*

Fellegi and Sunter classification

- For each compared record pair a vector with *matching weights* is calculated

Record A:	`[‘dr’, ‘peter’, ‘paul’, ‘miller’]`
Record B:	`[‘mr’, ‘john’, ‘’, ‘miller’]`
Matching weights:	`[0.2, -3.2, 0.0, 2.4 ]`

- *Fellegi and Sunter* approach sums all weights (then uses two thresholds to classify record pairs as *matches*, *non-matches*, or *possible matches*)

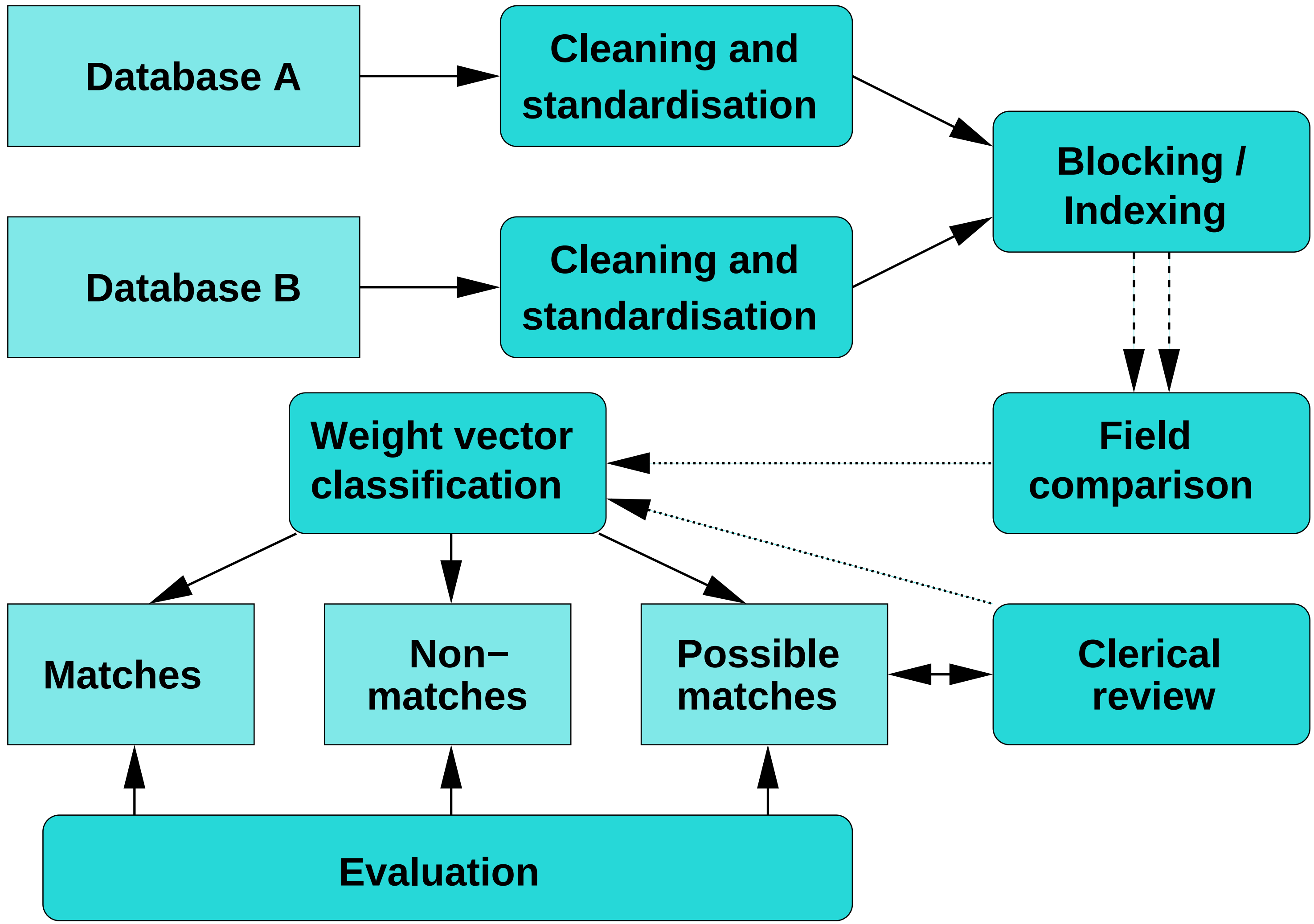


THE AUSTRALIAN NATIONAL UNIVERSITY

Modern linkage approaches

- Summing of weights results in loss of information (like *same name but different address*, or *different address but same name*)
- View record pair classification as a *multi-dimensional binary classification* problem (use weight vector to classify record pairs as *matches* or *non-matches*, but not *possible matches*)
- Many machine learning techniques can be used
  - Supervised: *Decision trees*, *neural networks*, *learnable string comparisons*, *active learning*, etc.
  - Un-supervised: Various *clustering* algorithms
- Major issue: Lack of training data

The data linkage process



Privacy and confidentiality issues

- The public is worried about their information being matched and shared between organisations
  - Good: health and social research; statistics, crime and fraud detection (taxation, social security, etc.)
  - Scary: intelligence, surveillance, commercial data mining (not much details known, no regulation)
  - Bad: identity fraud, re-identification
- Traditionally, *identified data* has to be given to the person or organisation performing the linkage
  - Privacy of individuals in data sets is invaded
  - Consent of individuals needed (often not possible, so approval from ethics review boards required)

Data linkage scenario 1

- A researcher is interested in analysing the effects of car accidents upon the health system
  - *Most common types of injuries?*
  - *Financial burden upon the public health system?*
  - *General health of people after they were involved in a serious car accident?*
- She needs access to data from hospitals, doctors, car insurances, and from the police
  - All identifying data has to be given to the researcher, or alternatively a trusted data linkage unit
- This might prevent an organisation from being able or willing to participate (car insurances or police)

Data linkage scenario 2

- Two pharmaceutical companies are interested in collaborating on the development of new drugs
- The companies wish to identify how much overlap of confidential data there is in their databases (without having to reveal any of that data to each other)
- Techniques are required that allow comparison of large amounts of data such that similar data items are found (while all other data is kept confidential)
- Involvement of a third party to undertake the linkage will be undesirable (due to the risk of collusion of the third party with either company, or potential security breaches at the third party)

Data linkage scenario 3

- A researcher has access to several linked data sets (which separately do not permit re-identification of individuals)
- He has access to a HIV database and a midwives data set (both contain postcodes, and year and month of birth – in the midwives data for both mothers and babies)
- Using birth notifications from a public Web site (news paper), the curious researcher is able to link records and identify births in rural areas by mothers who are in the HIV database
- Re-identification is a big issue due to the increase of data publicly available on the Internet

Geocoding scenario 1

- A cancer register aims to geocode its data
  (to conduct spatial analysis of different types of cancer)
- Due to limited resources the register cannot invest in an in-house geocoding system
  (software and personnel)
- They are reliant on an external geocoding service
  (commercial geocoding company or data matching unit)
- Regulations might not allow the cancer register to send their data to any external organisation
- Even if allowed, complete trust is required into the geocoding service (to conduct accurate matching, and to properly destroy the register's address data afterwards)

Geocoding scenario 2

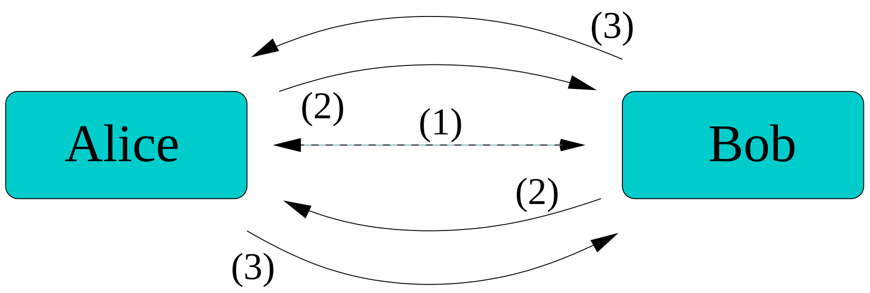
- A local police department publishes online maps with crime statistics
  - Such maps might result in businesses and residents leaving an area
  - Or attract burglars who see an area as a lucrative and easy target
- Serious and rare crimes might allow identification of the victim (reverse geocoding if exact location given)
  - Victims can be re-traumatised, or be seen as easy targets by criminals
  - Victims might therefore decide not to report a crime (such as sexual assault)

Privacy-preserving data linkage

- Pioneered by French researchers in 1990s [Dusserre et al. 1995; Quantin et al. 1998]
  - For situations where de-identified data needs to be centralised and linked for follow-up studies
  - Based on one-way hash-encoded values (SHA, MD5) (for example: *'peter'* $\rightarrow$ *'51ddc7d3a611eeba6ca770'*)
  - Allow exact matching only (improve using *Soundex* etc.)
- Best practice protocol [Kelman et al. 2002]
  - Physically separate identifying information from medical and other sensitive details
  - A variation of this approach is currently used by the *Western Australian Data Linkage Unit*

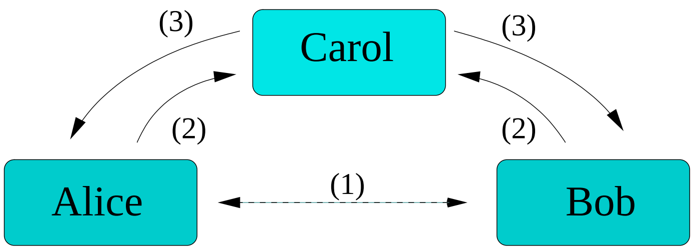
THE AUSTRALIAN NATIONAL UNIVERSITY

Two-party protocols



- Two data sources wish to link data (so that only information about the shared data is revealed to both)
- At any time, no party has the information needed to infer details about the other party's data
- Two recent approaches:
  - *Secure protocol for computing string distance metrics* (like TF-IDF and Euclidean) [Ravikumar et al. 2004]
  - *Secure and private sequence comparisons* (edit distance) [Atallah et al. 2003]

Three-party protocols



- Data sources send their encoded data to a third party, which performs the linkage
- Several recent approaches, including:
  - *Blindfolded data linkage* (more next)
  - *Privacy-preserving data linkage* (secure cohort extraction) [O'Keefe et al. 2004]
  - *Privacy-preserving blocking* [Al-Lawati et al. 2005]
  - *Hybrid approach combining anonymisation with secure-multi-party computation* [Inan et al. 2008]

THE AUSTRALIAN NATIONAL UNIVERSITY

Blindfolded data linkage

- Based on approximate string matching using q-grams [Churches and Christen, 2004]
- Assuming a three-party protocol
  - 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
- More complicated if values contain errors or variations (a single character difference between two strings will result in very different hash codes)

Protocol – Step 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 q-grams
  For example ($q = 2$, 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

Protocol – Step 2

- Protocol step 2
  - Alice computes a sorted list of *q*-grams for each of her values in **A.a**
  - Next she calculates all possible not empty sorted sub-lists (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 and stores these in **A.a_hash_bigr_comb**

Protocol – Steps 2 and 3

- Protocol step 2 (continued)
  - Alice computes an 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 3
  - Bob carries out the same as in step 2 with his **B.a**

THE AUSTRALIAN NATIONAL UNIVERSITY

Protocol – Step 4

- Protocol step 4
  - 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*:

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

  - 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
(or she only send the number of matches with a **bigr_score** above a certain similarity threshold)

Example

- Alice: *'peter'* $\rightarrow$ *[('er'), ... ('et','pe','te'), ...]*

  For bigram sub-list *('et','pe','te')*:

  – **A.a_hash_bigr_comb** = *'W5gO1@'*

  – **A.a_hash_bigr_comb_len** = *3*

  – **A.a_len** = *4*

  Alice sends to Carol: *['A-7D4W', 'W5gO1@', 3, 4]*

- Bob: *'pete'* $\rightarrow$ *[('er'), ... ('et','pe','te')]*

  For bigram sub-list *('et','pe','te')*:

  – **B.a_hash_bigr_comb** = *'W5gO1@'*

  – **B.a_hash_bigr_comb_len** = *3*

  – **B.a_len** = *3*

  Bob sends to Carol: *['B-T5YS', 'W5gO1@', 3, 3]*

- Carol calculates: **bigr_score** = $\frac{2 \times 3}{(4+3)} = \frac{6}{7} = 0.857$

Full blindfolded data linkage

- Several attributes **a**, **b**, **c**, etc. can be compared independently (by different Carols)
- Different Carols send their results to another party (David), who forms a (sparse) matrix by joining the results
- The final *matching weight* for a record pair is calculated using individual **bigr_score**s
- David arrives at a set of *blindly linked records*
  (pairs of [**A.a_encrypt_rec_key**, **B.a_encrypt_rec_key**])
- Neither Carol nor David learn what records and values have been matched

Challenges with privacy-preserving matching

- Many secure multi-party computations are computationally very expensive
  - Some have large communication overheads
  - Scalability to very large databases currently not feasible
- Not integrated with accurate classification techniques (because only encoded values are available, unsupervised learning is required)
- Assessment of matching quality problematic (not easy to verify if matched records correspond to true matches, and how many true matches were missed)
- Re-identification can still be a problem (if released records allow matching with external data)

Research directions (1)

- Secure matching
  - New and improved secure matching techniques (e.g. *Jaro-Winkler* comparator)
  - Reduce computational complexity and communication overheads of current cryptographic approaches
  - Frameworks and test-beds for comparing and evaluating secure data linkage techniques are needed
- Automated record pair classification
  - In secure three-party protocols, the linkage party only sees encoded data (no manual clerical review possible)
  - How to modify unsupervised classification techniques so they can work on encoded data?

Research directions (2)

- Scalability / Computational issues
  - Techniques for distributed (between organisations) linkage of very large data collections are needed
  - Combine secure matching and automated classification with distributed and high-performance computing
  - Also to be addressed: access protocols, fault tolerance, data distribution, charging policies, user interfaces, etc.
- Preventing re-identification
  - Make sure de-identified data linked with other (public) data does not allow re-identification
  - Possible approaches like *micro-data confidentiality* and *k-anonymity* [previous part of this tutorial]

Conclusions

- Scalable, automated and privacy-preserving data linkage is currently not feasible
- Four main research directions
  1. Improved secure matching
  2. Automated record pair classification
  3. Scalability and computational issues
  4. Preventing re-identification
- Public acceptance of data linkage is another major challenge
- For more information see project Web site

  (publications, talks, *Febrl* data linkage software)

  http://datamining.anu.edu.au/linkage.html

Thank you very much!

Any questions?

http://datamining.anu.edu.au/linkage.html

Contact: **peter.christen@anu.edu.au**

References (1)

- Al-Lawati A, Lee D and McDaniel P: *Blocking-aware private record linkage.* IQIS, Baltimore, 2005.
- Atallah MJ, Kerschbaum F and Du W: *Secure and private sequence comparisons.* WPES, Washington DC, pp. 39–44, 2003.
- Blakely T, Woodward A and Salmond C: *Anonymous linkage of New Zealand mortality and census data.* ANZ Journal of Public Health, 24(1), 2000.
- Chaytor R, Brown E and Wareham T: *Privacy advisors for personal information management.* SIGIR workshop on Personal Information Management, Seattle, pp. 28–31, 2006.
- Christen P and Churches T: *Secure health data linkage and geocoding: Current approaches and research directions.* ehPASS, Brisbane, 2006.
- Christen P: *Privacy-preserving data linkage and geocoding: Current approaches and research directions.* PADM workshop, held at IEEE ICDM, Hong Kong, pp. 497–501, 2006.
- Churches T: *A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers.* BMC Medical Research Methodology, 3(1), 2003.

THE AUSTRALIAN NATIONAL UNIVERSITY

References (2)

- Churches T and Christen P: *Some methods for blindfolded record linkage.* BMC Medical Informatics and Decision Making, 4(9), 2004.
- Clifton C, Kantarcioglu M, Doan A, Schadow G, Vaidya J, Elmagarmid AK and Suciu D: *Privacy-preserving data integration and sharing.* SIGMOD workshop on Research Issues in Data Mining and Knowledge Discovery, Paris, 2004.
- Dusserre L, Quantin C and Bouzelat H: *A one way public key cryptosystem for the linkage of nominal files in epidemiological studies.* Medinfo, 8:644-7, 1995.
- Elmagarmid AK, Ipeirotis PG and Verykios VS: *Duplicate record detection: A survey.* IEEE TKDE 19(1), pp. 1–16, 2007.
- Fienberg SE: *Privacy and confidentiality in an e-Commerce World: Data mining, data warehousing, matching and disclosure limitation.* Statistical Science, IMS Institute of Mathematical Statistics, 21(2), pp. 143–154, 2006.
- Hansen DP, Pang C and Maeder AJ: *HDI: Integrated services for health data.* ICMLC, Guangzhou, China, pp. 5554–5559, 2005.
- Inan A, Kantarcioglu M, Bertino E and Scannapieco M: *A hybrid approach to private record linkage.* IEEE ICDE, Cancun, Mexico, pp. 496–505, 2008.

References (3)

- Jonas J and Harper J: *Effective counterterrorism and the limited role of predictive data mining.* Policy Analysis, 584, 2006.
- Kelman CW, Bass AJ and Holman CDJ: *Research use of linked health data – A best practice protocol.* ANZ Journal of Public Health, 26(3), pp. 251–255, 2002.
- Li Y, Tygar JD and Hellerstein JM: *Private matching.* Computer Security in the 21st Century, Lee DT, Shieh SP and Tygar JD (editors), Springer, 2005.
- Malin B, Airoldi E, Edoho-Eket S and Li Y: *Configurable security protocols for multi-party data analysis with malicious participants.* IEEE ICDE, Tokyo, pp. 533–544, 2005.
- Malin B and Sweeney L: *A secure protocol to distribute unlinkable health data.* American Medical Informatics Association 2005 Annual Symposium, Washington DC, pp. 485–489, 2005.
- O'Keefe CM, Yung M, Gu L and Baxter R: *Privacy-preserving data linkage protocols.* WPES, Washington DC, pp. 94–102, 2004.
- Quantin C, Bouzelat H and Dusserre L: *Irreversible encryption method by generation of polynomials.* Medical Informatics and The Internet in Medicine, Informa Healthcare, 21(2), pp. 113–121, 1996.

References (4)

- Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J and Dusserre L: *How to ensure data quality of an epidemiological follow-up: Quality assessment of an anonymous record linkage procedure.* International Journal of Medical Informatics, 49, pp. 117–122, 1998.
- Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J and Dusserre L: *Automatic record hash coding and linkage for epidemiological follow-up data confidentiality.* Methods of Information in Medicine, Schattauer, 37(3), pp. 271–277, 1998.
- Ravikumar P, Cohen WW and Fienberg SE: *A secure protocol for computing string distance metrics.* PSDM held at IEEE ICDM, Brighton, UK, 2004.
- Schadow G, Grannis SJ and McDonald CJ: *Discussion paper: Privacy-preserving distributed queries for a clinical case research network.* CRPIT'14: Proceedings of the IEEE international Conference on Privacy, Security and Data Mining, Maebashi City, Japan, pp. 55–65, 2002.
- Sweeney L: *K-anonymity: A model for protecting privacy.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, World Scientific Publishing Co., Inc., 10(5), pp. 557–570, 2002.

References (5)

- Sweeney L: *Privacy-enhanced linking.* ACM SIGKDD Explorations, 7(2), pp. 72–75, 2005.
- Verykios VS, Bertino E, Fovino IN, Provenza LP, Saygin Y and Theodoridis Y: *State-of-the-art in privacy preserving data mining.* ACM SIGMOD Rec., 33(1), pp. 50–57, 2004.
- Wartell J and McEwen T: *Privacy in the information age: A Guide for sharing crime maps and spatial data.* Institute for Law and Justice, National Institute of Justice, 188739, 2001.
- Winkler WE: *Masking and re-identification methods for public-use microdata: Overview and research problems.* Privacy in Statistical Databases, Barcelona, Springer LNCS 3050, pp. 216–230, 2004.
- Winkler WE: *Overview of record linkage and current research directions.* RR 2006/02, US Census Bureau, 2006.
- Zhang Q and Hansen D: *Approximate processing for medical record linking and multidatabase analysis.* International Journal of Healthcare Information Systems and Informatics, 2(4), pp. 59–72, 2007.