Privacy-Preserving Data Linkage and Geocoding: Current Approaches and Research Directions

Peter Christen

Department of Computer Science,
Faculty of Engineering and Information Technology,
ANU College of Engineering and Computer Science,
The Australian National University

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

Project Web site: **http://datamining.anu.edu.au/linkage.html**

Funded by the Australian National University, the NSW Department of Health, and the Australian Research Council (ARC) under Linkage Project 0453463.

- What is data linkage and geocoding?
  - Applications and challenges
- Some data linkage and geocoding scenarios
  - Illustrate privacy and confidentiality issues
- Current privacy-preserving data linkage approaches
- Research directions
  - Ultimate aim: Automated secure linking and geocoding of very large data collections between organisations
- Outlook

What is data (or record) linkage?

- The process of linking and aggregating records from one or more data sources representing the same entity (patient, customer, business name, etc.)
  - Also called *data matching*, *data integration*, *data scrubbing*, *ETL (extraction, transformation and loading)*, *object identification*, *merge-purge*, etc.
- Challenging if no unique entity identifiers available
  E.g., which of these records represent 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 2600 Canberra ACT*

What is geocoding?

- The process of matching addresses to their geographic locations (longitude and latitude)
  - Large reference database of cleaned and standardised addresses is needed
  - Accurate matching is important
  - Addresses often contain typographical errors, are incomplete or out-of-date
- It is estimated that 80% to 90% of governmental and business data contain address information
  [Federal geographic data committee, US Pub Health, 2003]
- Useful in many application areas
  - Visualisation, spatial data analysis and mining

Challenge 1: Larger data collections

- Data collections with tens or even hundreds of millions of records are not uncommon
- Number of possible record pairs to compare equals the product of the sizes of the two data sets
  (linking two data sets with *1,000,000* records each will result in $10^6 \times 10^6 = 10^{12}$ record pairs)
- Performance bottleneck in a data linkage system is usually the (expensive) comparison of attribute (field) values between record pairs
- Blocking, indexing, clustering and filtering techniques are used to reduce the large number of comparisons

Challenge 2: Manual clerical review

- Traditional data linkage classifies record pairs into *matches*, *non-matches*, and *possible matches*
  - *Possible matches* are manually clerically reviewed to decide their linkage status
  - Very time consuming and tedious, but also hard to make correct and consistent decisions
- With larger data collections, the number of possible matches also increases
- Long durations for linkage projects not uncommon
  (days or even weeks, involving several linkage experts)
- Decision models are needed that will reduce or even eliminate the manual clerical review step

Challenge 3: Privacy and confidentiality

- General public is worried about their information being linked 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 information from businesses, 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 involved is needed (often not

Data linkage scenario 1

- A researcher is interested in analysing the effects of car accidents upon hospital admissions
  (for example what types of injuries are most common, the resulting financial burden upon the public health system, and the general health of people that were involved in serious car accidents)
- She needs access to hospital data, as well as detailed data from car insurers and possibly even access to a police database
  (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 insurers or police)

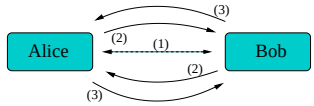
- Two pharmaceutical companies are interested in collaborating on the development of new drugs
- The companies wish to identify how much overlap of confidential data there is in their databases (without having to reveal any of that data to each other)
- Techniques are required that allow comparison of 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)

- A cancer register aims to geocode its data (to conduct a 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 linkage unit)
- Regulations might not allow the cancer register to send their data to an 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)

Current approaches to privacy-preserving data linkage

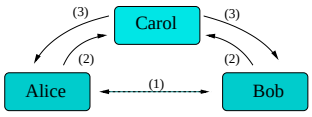
- Pioneered by French researchers in 1990s [Dusserre et al. 1995; Quantin et al. 1998]
  - For situations where de-identified data needs to be centralised and linked for follow-up studies
  - Based on one-way hash-encoded values (For example: *'peter'* → *'51ddc7d3a611eeba6ca770'*)
  - Allow exact matching only
- 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*

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* (TF-IDF and Euclidean distance) [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 publications, including:
  - *Blindfolded record linkage* (approximate string matching using *q*-grams) [Churches and Christen 2004]
  - *Privacy-preserving data linkage* (secure cohort extraction) [O'Keefe et al. 2004]
  - *Privacy-preserving blocking* [Al-Lawati et al. 2005]

Research directions (1)

- Secure matching
  - New and improved secure matching techniques (e.g. *Jaro-Winkler* comparator)
  - Many cryptographic approaches have computational overheads (impractical for very large data collections)
  - Frameworks and test-beds for comparing and evaluating secure matching techniques are needed
- Automated record pair classification
  - In secure three-party protocols, the linkage party only sees encoded data (no manual clerical review possible)
  - Unsupervised classification techniques are needed

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

Outlook

- Secure, automated and distributed data linkage for very large data collections 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 our project Web site (publications, talks, *Febrl* data linkage software)