Secure Health Data Linkage and Geocoding: Current Approaches and Research Directions

Peter Christen[1] and Tim Churches[2]

[1]Department of Computer Science, The Australian National University
[2]Centre for Epidemiology and Research, New South Wales Department of Health

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, the Australian Research Council (ARC) under Linkage Project 0453463.

Outline

- What is data linkage and geocoding?
  - Applications and challenges
- Health data linkage and geocoding scenarios
  - Illustrate privacy and confidentiality issues
- Current approaches to secure health data linkage
- Research directions
  - Ultimate aim: Automated secure linking and geocoding of very large data collections between organisations
- Australian health data linkage projects

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?

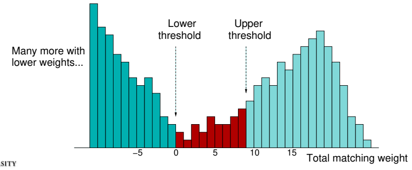
Dr Smith, Peter	*42 Miller Street 2602 O'Connor*
Pete Smith	*42 Miller St 2600 Canberra A.C.T.*
P. Smithers	*24 Mill Street 2600 Canberra ACT*

Traditional data linkage

- For each compared record pair a vector containing *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]

- Traditional approach sums all weights
  (then classifies record pairs into *matches*, *non-matches*, and *possible matches*) [Fellegi and Sunter 1969]



What is geocoding?

- The process of matching addresses to their geographic locations (longitude and latitude)
  - Large reference database of standardised addresses needed (in Australia: *G-NAF*) [Christen et al. 2006]
  - 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  - Addresses often contain typographical errors, are incomplete or out-of-date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
  (US Federal Geographic Data Committee)
- 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 / filtering techniques are used to reduce the large amount of comparisons

Challenge 2: Manual clerical review

- Traditionally, *possible matches* are manually looked at to decide their linkage status
- With larger data collections, the number of possible matches also increases
- Very time consuming and tedious, but also hard to make correct and consistent decisions
  (if only classifying one record pair at a time)
- 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 while keeping a high linkage quality

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 possible, so seek approval from ethics committees)

Health 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)

Health data linkage scenario 2

- 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, 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

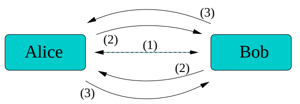
Health geocoding scenario

- 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 secure data linkage

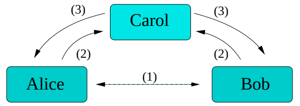
- Pioneered by French researchers in 1990s [Dusserre et al. 1995 and Quantin et al. 1998]
  - For situations where de-identified data needs to be centralised and linked for follow-up studies
  - Based on one-way hash-encoded values (For example: *'peter'* → *'51ddc7d3a611eeba6ca770'*)
  - Allow exact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

Cryptographic 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
- All communication is encrypted
- Two recent papers present approaches for secure approximate string comparisons [Atallah 2003 and Ravikumar 2004]

Cryptographic three-party protocols



- Data sources send their encoded data to a third party, which performs the linkage
  - No identifying information is revealed by the data sources
- Recent work in Australia
  - *Blindfolded record linkage* (allows approximate string matching) [Churches and Christen 2004]
  - *Privacy-preserving data linkage* (secure cohort extraction) [O'Keefe et al. 2004]

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 for comparing and evaluating secure matching techniques
- 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 [Winkler 2004] and k-anonymity [Sweeney 2002]

Australian data linkage projects

- Western Australian Data Linkage Unit
  - Currently re-designing their linkage system
  - Model for the NSW Centre for Health Record Linkage
- Health Data Integration (CSIRO / QLD Health)
  - Develop infrastructure and software for secure health data linkage
- *Febrl* (Freely Extensible Biomedical Record Linkage)
  - Australian National University / NSW Health
  - Aims to develop improved techniques for parallel large scale data linkage
  - Freely available open source software (for data cleaning, deduplication, linkage and geocoding)

Outlook

- Secure, automated and distributed data linkage for very large data collections is currently not feasible
- Four main technical research directions
  - Improved secure matching
  - Automated record pair classification
  - Scalability and computational issues
  - Preventing re-identification
- Public acceptance of data linkage is another major challenge
- For more information see our project Web site (publications, talks, *Febrl* software, Web links)
  http://datamining.anu.edu.au/linkage.html