Geocode Matching and Privacy Preservation

Peter Christen

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

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

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

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

Outline

- Data and geocode matching
  - Applications and challenges
  - Geocode matching techniques
  - Reverse geocoding
- Privacy and confidentiality issues with matching
- Data and geocode matching scenarios
- Current privacy-preserving matching approaches
- Privacy-preserving geocode matching
- Research directions
- Conclusions

Data matching

- The process of matching and aggregating records that represent the same entity (such as a patient, a customer, a business, or an *address*)
  - Also called *record* or *data linkage*, *entity resolution*, *data scrubbing*, *object identification*, *merge-purge*, etc.
- Example applications
  - Health, biomedical and social sciences
  - Census, taxation, social security
  - Crime and fraud detection, national security
  - Deduplication of (business mailing) lists
  - Bibliographic databases and online libraries

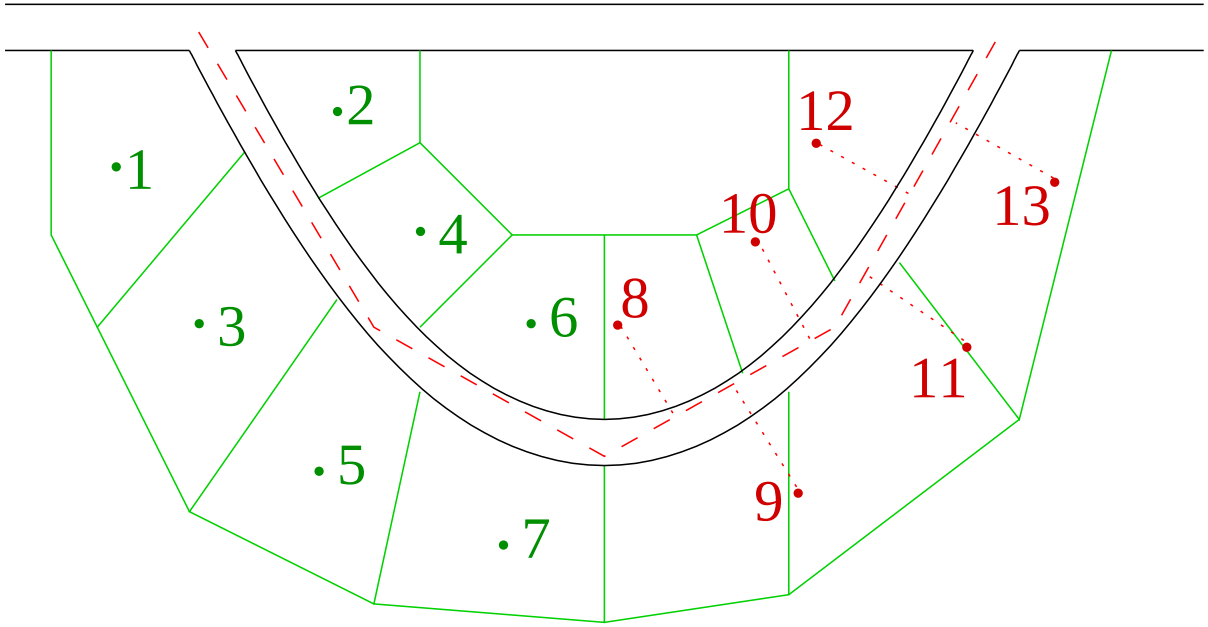
Data matching challenges

- Often no unique entity identifiers are available
- Real world data is dirty
  (typographical errors and variations, missing and out-of-date values, different coding schemes, etc.)
- Scalability
  - Naïve comparison of all record pairs is $\boldsymbol{O}(n \times m)$
  - Some form of blocking, indexing or filtering is required
- Privacy and confidentiality
  (because personal information, like names and addresses, are commonly required for matching)
- No training data in many matching applications
  - No record pairs with known true match status

Geocode matching

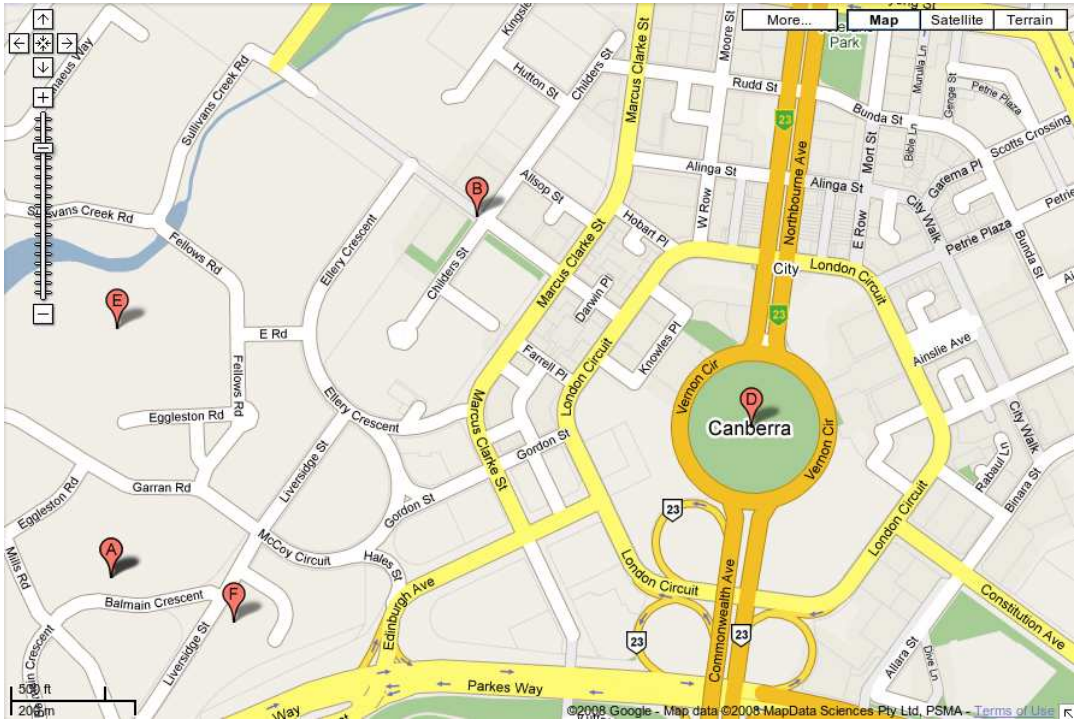
- The task of matching addresses or place names to geographic locations (latitude and longitude)
  - Large reference database of cleaned and standardised addresses and place names is needed
  - Accurate matching is important (but addresses often contain errors, are incomplete, or out-of-date)
- It is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

Geocode matching techniques



- Street centreline based (many commercial systems)
- Property centre based (requires accurate and complete database)
- A recent study found substantial differences (especially in rural areas) *[Cayo & Talbot, IJHG 2003]*

Reverse geocoding



- The process of matching a point on a map back to an address
  - Of great concern in the health sector when maps of disease cases are being published for research

Privacy and confidentiality issues

- 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 matching
  - Privacy of individuals in data sets is invaded
  - Consent of individuals needed (often not possible, so approval from ethics review boards required)

Data matching scenario

- A researcher is interested in analysing the effects of car accidents upon hospital admissions
  - *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 hospital data, data from car insurances, and from the police
  - All identifying data has to be given to the researcher, or alternatively a trusted data matching unit
- This might prevent an organisation from being able or willing to participate (car insurances or police)

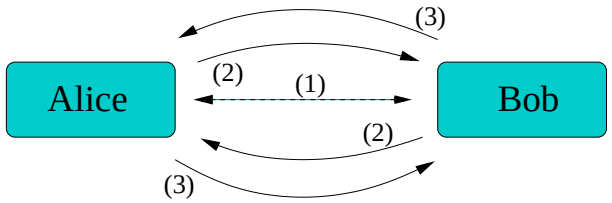
Geocode matching scenario 1

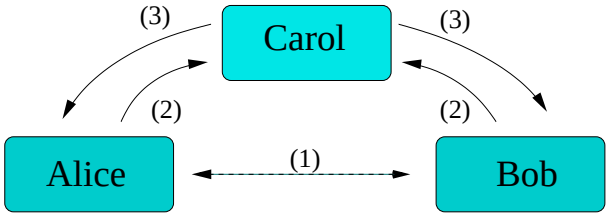
- A cancer register aims to geocode its data
  (to conduct spatial analysis of different types of cancer)
- Due to limited resources the register cannot invest in an in-house geocoding system
  (software and personnel)
- They are reliant on an external geocoding service
  (commercial geocoding company 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)

Geocode matching 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 matching approaches





- Based on cryptographic techniques
  (secure multi-party computations)
- Assume two data sources, and possibly a third (trusted) party to conduct the matching
- Objective: No party learns about the other parties' private data, only matched records are released
  - Various approaches with different assumptions about threats, what can be inferred by parties, and what is being released

Privacy-preserving matching techniques

- Pioneered by French researchers for exact matching *[Dusserre et al. 1995; Quantin et al. 1998]*
  - Using one-way hash-encoding (*'tim'* $\rightarrow$ *'51d3a6a70'*)
- Secure and private sequence comparisons (edit distance) *[Atallah et al. WPES'03]*
- Blindfolded record linkage (approximate string matching using encoded *q*-grams)
  [Churches and Christen, BioMed Central 2004]
- Secure protocol for computing string distance metrics (TF-IDF and Euclidean distance)
  [Ravikumar et al. PSDM'04]
- Privacy-preserving blocking *[Al-Lawati et al. IQIS'05]*

Challenges with privacy-preserving matching

- Many secure multi-party computations (SMC) are computationally very expensive
  - Some have large communication overheads
  - Scalability to very large databases currently not feasible
  - Recent approach combines SMC with sanitisation techniques (such as *k*-anonymisation) to reduce complexity *[Inan et al. ICDE'08]*
- 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)

Privacy-preserving geocode matching

- Privacy-preserving matching approaches need to be modified for privacy-preserving geocoding
- Geocode scenario 1 (cancer register): The aim is to enable geocoding of cancer records at an external geocode service, such that the service does not learn which addresses are matched
- Geocode scenario 2 (crime maps): Sanitisation and generalisation methods have to be applied (no exact addresses – pins on map – can be released, only averaged area data)
  - Still possibility of re-identification with external data (e.g. media information and population data) *[Chaytor et al. SIGIR workshop 2006]*

Research directions

- Develop privacy-preserving matching techniques specific for geocoding
  - Develop address specific approximate comparison functions (address values are often long strings, for example Australian suburb name ‘Woolloomooloo’)
- Improve computational and communication complexity (for scalable privacy-preserving matching)
- Automated record pair classification
  (no training data available in privacy-preserving settings, so unsupervised techniques required)
- Guarantee that re-identification is impossible
- Publicly available test-beds are also required

Conclusions

- Geocode matching (as data matching in general) can have some serious privacy implications
- A variety of privacy-preserving data matching approaches have been developed (however, not specific to geocode matching yet)
- More research and development needed to enable privacy-preserving matching of very large databases
- Also required are regulations and policies that allow and support privacy-preserving matching
  - And public support of data matching in general

Thank you very much!

Any questions?

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