Overview and taxonomy of techniques for privacy-preserving record linkage

Peter Christen

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

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

Based on work done by and with Dinusha Vatsalan (ANU) and
Vassilios Verykios (Hellenic Open University, Greece)

THE AUSTRALIAN NATIONAL UNIVERSITY

Motivation

- Large amounts of data are being collected both by organisations in the private and public sectors, as well as by individuals
- Much of these data are about people, or they are generated by people
  - Financial, shopping, and travel transactions
  - Electronic health and financial records
  - Emails, tweets, SMSs, blog posts, etc.
  - Tax, social security, and census records
- Analysing such data can provide significant benefits to governments and businesses

Motivation (continued)

- Often data from different sources need to be integrated and linked
  - Improve data quality
  - Enrich data
  - Allow data analyses that are impossible on individual databases
- Lack of unique entity identifiers means that linking is often based on personal information
- When databases are linked across organisations, maintaining privacy and confidentiality is vital
- This is where privacy-preserving record linkage (PPRL) can help

Outline

- What is record linkage?
- Applications, history, and challenges
- The record linkage and PPRL processes
- A definition of PPRL
- A taxonomy for PPRL
- Summary of the state of PPRL
- Challenges and future work

What is record linkage?

- The process of linking records that represent the same entity in one or more databases
  (patient, customer, business name, etc.)
- Also known as *data matching*, *entity resolution*, *data linkage*, *object identification*, *identity uncertainty*, *merge-purge*, etc.
- Major challenge is that unique entity identifiers are often not available in the databases to be linked (or if available, they are not consistent)

  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 Rd 2600 Canberra ACT*

Applications of record linkage

- Applications of record linkage
  - Remove duplicates in a data set (de-duplication)
  - Merge new records into a larger master data set
  - Compile data for longitudinal (over time) studies
  - Clean and enrich data sets for data mining projects
  - Geocode matching (with reference address data)
- Example application areas
  - Immigration, taxation, social security, census
  - Fraud, crime, and terrorism intelligence
  - Business mailing lists, exchange of customer data
  - Social, health, and biomedical research

A short history of record linkage (1)

- Computer assisted record 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) and error estimates
  - Sum of the matching weights is used to classify a pair of records as a *match*, *non-match*, or *potential match*
  - Problems: Estimating errors and thresholds, assumption of independence, and *clerical review*

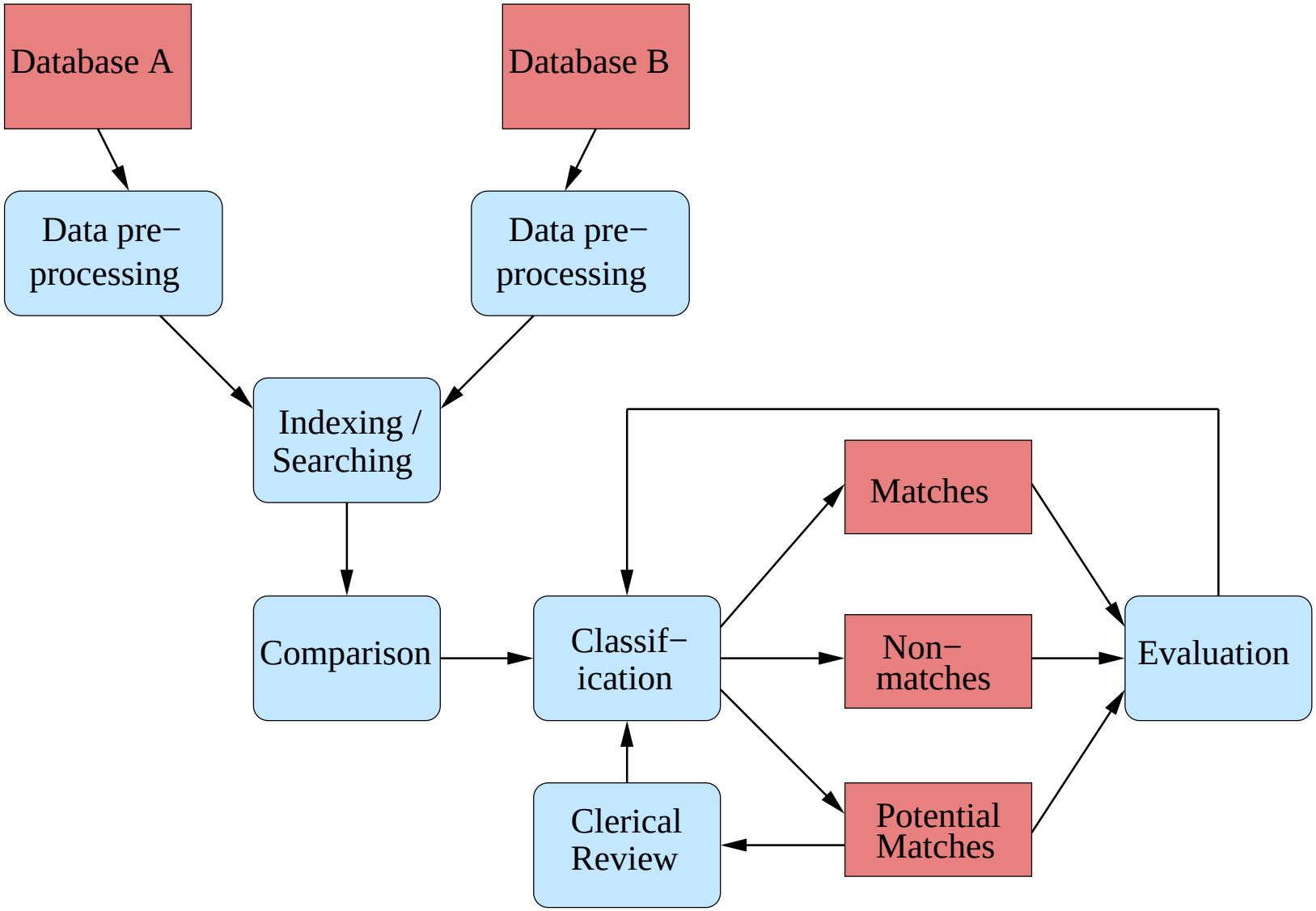
A short history of record linkage (2)

- Strong interest in the last decade from computer science (from data mining, AI, knowledge engineering, information retrieval, databases, digital libraries, etc.)
- Many different techniques have been developed
- Major focus is on scalability to large databases, and linkage quality
  - Various indexing/blocking techniques to efficiently and effectively generate candidate record pairs
  - Various machine learning-based classification techniques, both supervised and unsupervised, as well as active learning based

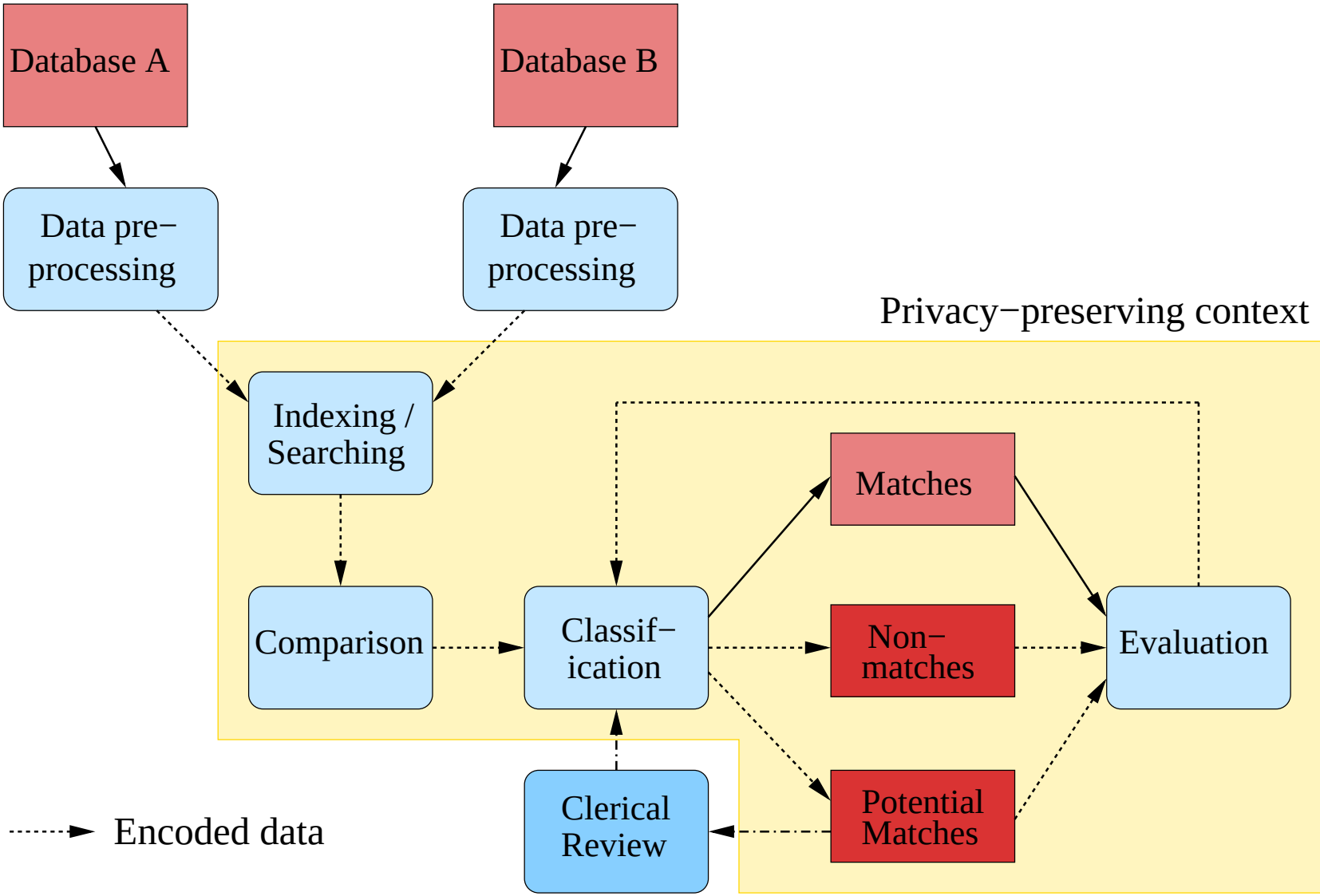
Record linkage challenges

- No unique entity identifiers 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 quadratic
  - Remove likely no-matches as efficiently as possible
- No training data in many linkage applications
  - No record pairs with known true match status
- Privacy and confidentiality
  (because personal information, like names and addresses, are commonly required for linking)

The record linkage process



The PPRL process

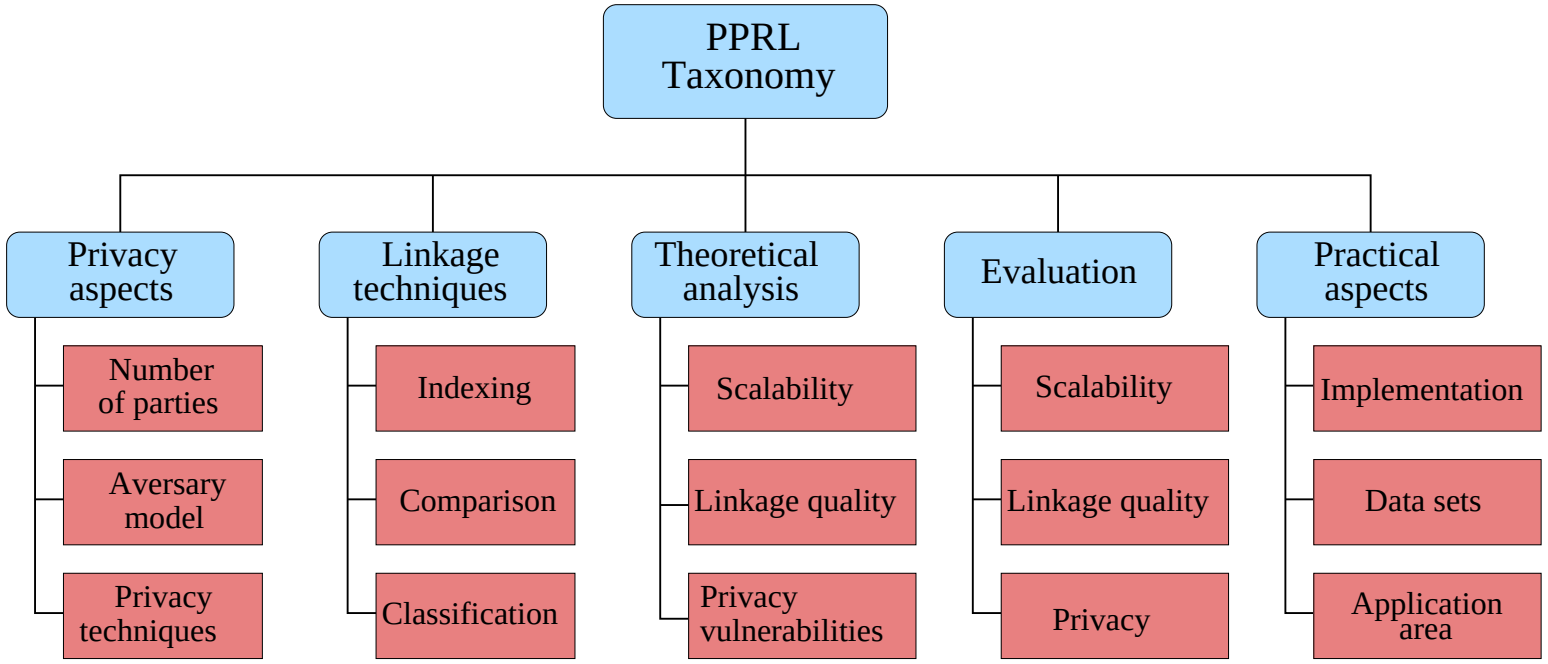


A definition of PPRL

- Assume $O_1 \cdots O_d$ are the d owners of their respective databases $D_1 \cdots D_d$
- They wish to determine which of their records $r_1^i \in D_1$, $r_2^j \in D_2$, $\cdots$, and $r_d^k \in D_d$, match according to a decision model $C(r_1^i, r_2^j, \cdots, r_d^k)$ that classifies pairs (or groups) of records into one of the two classes M of matches, and U of non-matches
- $O_1 \cdots O_d$ do not wish to reveal their actual records $r_1^i \cdots r_d^k$ with any other party

  (they are however prepared to disclose to each other (or to an external party) the actual values of some attributes of the record pairs that are in class M to allow further analysis)

A taxonomy for PPRL



- Characterise PPRL techniques with the aim to
  - Get a clearer picture of current approaches to PPRL
  - Specify gaps between record linkage and PPRL
  - Identify directions for future research in PPRL

Taxonomy: Privacy aspects

- Number of parties involved in a protocol
  - **Two-party protocol**: Two *database owners* only
  - **Three-party protocol**: Require a (trusted) third party
- Adversary model
  - Based on models used in cryptography:
    Honest-but-curious or **malicious** behaviour
- Privacy technologies – many different approaches
  - One-way hash encoding, generalisation, secure multi-party computation, differential privacy, Bloom filters, public reference values, phonetic encoding, dummy extra values, and various others

Taxonomy: Linkage techniques

- Indexing / blocking
  - Indexing aims to identify candidate record pairs that likely correspond to matches
  - Different techniques used: blocking, sampling, generalisation, clustering, hashing, binning, etc.
- Comparison
  - **Exact** or **approximate** (consider partial similarities, like *"vest"* and *"west"*, or *"peter"* and *"pedro"*)
- Classification
  - Based on the similarities calculated between records
  - Various techniques, including similarity threshold, rules, ranking, probabilistic, or machine learning techniques

Taxonomy: Theoretical analysis

- Scalability (of computation and communication, usually done using 'big ***O***' notation – ***O(n)***, ***O(n^2)***, etc.)
- Linkage quality
  - Fault (error) tolerance
  - Field- or record based (matching)
  - Data types (strings, numerical, age, dates, etc.)
- Privacy vulnerabilities
  - Different types of attack (frequency, dictionary, and crypt-analysis)
  - Collusion between parties

Taxonomy: Evaluation

- Scalability
  - We can measure **run-time** and **memory usage**
  - Implementation independent measures are based on the number of candidate record pairs generated
- Linkage quality
  - Classifying record pairs as **matches** or **non-matches** is a binary classification problem, so we can use traditional accuracy measures
- Privacy
  - Least ‘standardised’ area of evaluation, with various measures used (including information gain, simulation proofs, probability of re-identification, etc.)

Taxonomy: Practical aspects

- Implementation
  - Programming language used (if implemented), or only theoretical proof-of-concept
  - Sometimes no details are published
- Data sets
  - Real-world data sets or synthetic data sets
  - Public data (from repositories) or confidential data
- Targeted application areas
  - Include health care, census, business, finance, etc.

Summary of the state of PPRL

- Significant advances to achieving the goal of PPRL have been developed in recent years
  - Various approaches based on different techniques
  - Can link records securely, approximately, and in a (somewhat) scalable fashion
- So far, most PPRL techniques concentrated on approximate matching techniques, and on making PPRL more scalable to large databases
- However, no large-scale comparative evaluations of PPRL techniques have been published
- Only limited investigation of classification and linking assessment in PPRL

Challenges and future work (1)

- Improved classification for PPRL
  - Mostly simple threshold based classification is used
  - No investigation into linkage advanced methods, such as collective entity resolution techniques
  - Supervised classification is difficult – no training data in most situations
- Assessing linkage quality and completeness
  - How to assess linkage quality?
    – How many classified matches are true matches?
    – How many true matches have we found?
  - Evaluating actual record values is not possible (as this would reveal sensitive information)

Challenges and future work (2)

- A framework for PPRL is needed
  - To facilitate comparative experimental evaluation of PPRL techniques
  - Needs to allow researchers to plug-in their techniques
  - Benchmark data sets are required (biggest challenge, as such data are sensitive!)
- PPRL on multiple databases
  - Most work so far is limited to linking two databases (in reality often databases from several organisations)
  - Pair-wise linking does not scale up
  - Preventing collusion between (sub-groups of) parties becomes more difficult

For more on this topic...

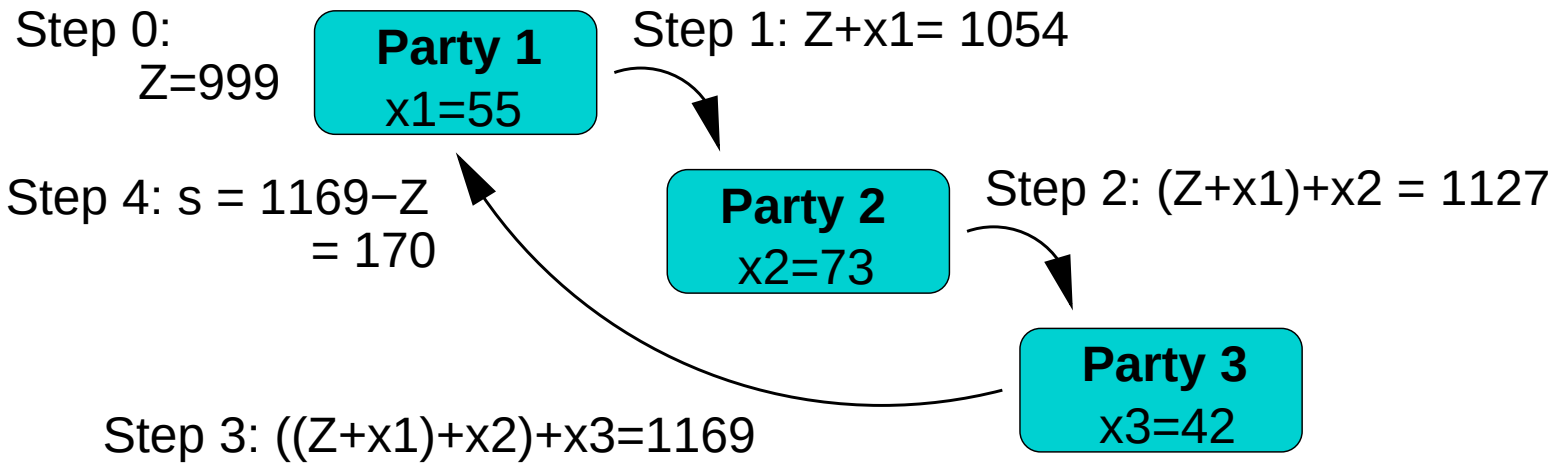
- *A Taxonomy of Privacy-Preserving Record Linkage Techniques*
  D Vatsalan, P Christen, and V Verykios
  Elsevier Information Systems, 2012.
  http://dx.doi.org/10.1016/j.is.2012.11.005

- *Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*
  P Christen
  Springer Data-Centric Systems and Applications, July 2012.
  http://cs.anu.edu.au/people/Peter.Christen/data-matching-book-2012.html

Extra slides follow...

Secure multi-party computation

- Compute a function across several parties, such that no party learns the information from the other parties, but all receive the final results
  [Yao 1982; Goldreich 1998/2002]
- Simple example: Secure summation $s = \sum_i x_i$.



Example scenario (1): Public health research

- A research group 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?*
- They need access to data from hospitals, doctors, car and health insurers, and from the police
  - All identifying data have to be given to the researchers, or alternatively a trusted record linkage unit
- This might prevent an organisation from being able or willing to participate (insurers or police)

Example scenario (2): Business collaboration

- Collaboration benefits businesses (for example in improving efficiency and reducing the costs of their supply chains)
- They are not willing to share confidential data such as strategies and competitive knowledge
- Identifying which supplies and/or customers two businesses have in common must be done without revealing any other confidential knowledge
- Involvement of a third party to undertake the linking will be undesirable
  (due to the risk of collusion of the third party with either company, or potential security breaches at the third party)

Example scenario (3): Crime investigation

- A national crime investigation unit is tasked with fighting against crimes that are of national significance (such as organised crime syndicates)
- This unit will likely manage various national databases which draw from different sources (including law enforcement and tax agencies, Internet service providers, and financial institutions)
- These data are highly sensitive; and storage, retrieval, analysis and sharing must be tightly regulated (collecting such data in one place makes them vulnerable to outsider attacks and internal adversaries)
- Ideally, only linked records (such as those of suspicious individuals) are available to the unit

Contact

Peter Christen

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

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

THE AUSTRALIAN NATIONAL UNIVERSITY