INTERNATIONAL
Population Data Linkage
NETWORK

November 2020

LINKING SENSITIVE DATA
APPLICATIONS, TECHNIQUES, AND CHALLENGES

Prof Peter Christen, The Australian National University

Dr Thilina Ranbaduge, The Australian National University

Prof Rainer Schnell, University Duisburg-Essen

IPDLN2020 is proudly sponsored by

GOLD SPONSORS

Australian Government
Australian Institute of Health and Welfare

PHRN Population Health Research Network

SA·NT DataLink

SILVER SPONSORS

IC/ES Data Discovery Better Health

UNIVERSITY OF CALGARY
O'Brien Institute for Public Health

BRONZE SPONSORS

University of South Australia

THE UNIVERSITY OF WESTERN AUSTRALIA

Linking Sensitive Data – Applications, Techniques, and Challenges

Peter Christen[1,*], Thilina Ranbaduge[1], and Rainer Schnell[2]

[1] The Australian National University, Australia
[2] University Duisburg-Essen, Germany

* Presenter

Parts of this work were funded by the Australian Research Council under DP130101801 and DP160101934.

Linking Sensitive Data

- Increasingly applications require records from different databases (owned by different organisations) to be linked
- Due to privacy and confidentiality concerns, organisations are often not willing or allowed to share or reveal their sensitive data
- Privacy-preserving record linkage (PPRL) methods aim to perform linkage of different databases using encoded or encrypted values
- The outcomes of a PPRL project are only the set of matched record pairs, while no information about sensitive data can be learned by any party involved in the linkage, nor any external party

Example case studies

- **Linking data on newborns**
  - Measurements on newborns (birth weight, survival of the first week, and so on) are considered as important indicators for the quality of a health system
  - Requires the linking of highly sensitive maternity records potentially across hundreds of hospitals
- **Linking census data over time**
  - Many countries conduct censuses on a regular basis
  - To be able to create longitudinal data about a population, census data need to be linked over time (challenging due to changes in personal details such as names and addresses)
  - The public is generally concerned about governments storing personal census data over time (only encrypted data can potentially be kept)

Privacy-preserving record linkage

- PPRL techniques are based on encrypted or encoded quasi-identifying values
  - Must allow approximate matching, scalability to linking large databases, and provide privacy protection
- PPRL techniques can be categorised into secure multiparty computation (SMC) and perturbation based techniques
  - SMC based techniques are provably secure but generally have higher computation and communication requirements
  - Perturbation based techniques are more efficient, allow for approximate matching, but might be vulnerable to privacy attacks

Perturbation based PPRL

- Techniques such as Bloom filter encoding have shown to be popular for practical PPRL applications
  - They are efficient, easy to implement, and allow comparison of textual and numerical values, and even hierarchical codes (such as occupation or disease codes)
  - Research has however shown that Bloom filter encoding can be vulnerable to certain cryptanalysis attacks (that exploit patterns in sets of Bloom filters)

Evaluating PPRL techniques

- Traditional evaluation of linkage techniques only considers linkage quality and scalability
  - Quality measures such as precision, recall, sensitivity, positive predictive value, etc.
  - Scalability measures such as reduction ratio, runtime, memory usage, etc.
- Privacy evaluation is more challenging
  - No single measure for privacy
  - Measures from statistical disclosure control or information gain have been adapted
  - Recent work is looking at vulnerability assessments

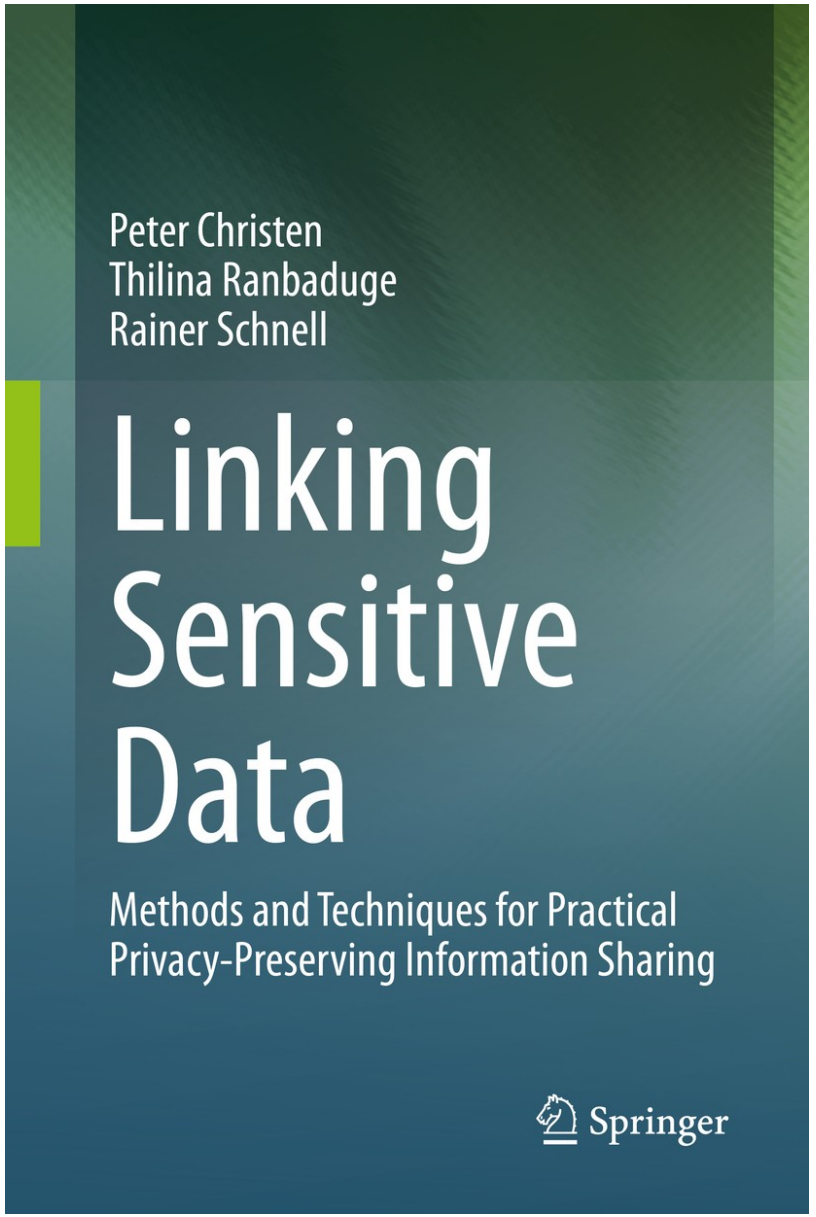
Practical aspects of PPRL

- From a practical perspective, other aspects of PPRL will also be of importance
  - Formal legal constraints and their organisational implementation
  - Dealing with dirty and missing data, as well as temporal and dynamic data
  - Dealing with bias and uncertainty in linked data
  - Costs of false and missed matches
  - Lack of ground truth data, and how to evaluate linkage quality in a PPRL context
  - Suitability of certain PPRL techniques for a given linkage (for example many communication steps required for some SMC techniques)
  - The actual linkage scenario (including threat scenario)
  - Technical knowledge available in an organisation
  - Availability of software or ease of implementation of a technique

Discussion and Conclusion

- PPRL techniques are now becoming mature enough to be used in practical applications
  - Providing high linkage quality, scalability, and privacy guarantees
- However, various practical aspects are still hindering the use of PPRL
  - Formal legal constraints and their organisational implementation
  - Lack of standard privacy evaluation measures
  - Lack of evaluation frameworks (like benchmark data sets)
  - Lack of available high quality software
  - Lack of people with required expertise (both in linkage methods as well as encoding and encryption techniques)

Linking Sensitive Data – the book



Springer, November 2020
approx. 490 pages

The book describes how linkage methods work and how to evaluate their performance. It covers all the major concepts and methods and also discusses practical matters such as computational efficiency, which are critical if the methods are to be used in practice - and it does all this in a highly accessible way!

Prof David J. Hand, OBE
Imperial College London