School of Public Health

Implementing Privacy-Preserving Record Linkage in a Cloud Computing Environment

Adrian Paul Brown
0000-0002-7514-8646

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

April 2021

Declaration of Authorship

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

The research presented and reported in this thesis was conducted in accordance with the National Health and Medical Research Council National Statement on Ethical Conduct in Human Research (2007) – updated March 2014. The research study received human research ethics approval from the Curtin University Human Research Ethics Committee (EC00262), Approval Number HRE2016-0195.

Signed:

_____

Date:

_____

Adrian Paul Brown

Abstract

The increase in demand for record linkage of administrative data across multiple sectors presents a considerable challenge for linkage units and their existing systems. These systems struggle to scale with the increased size and frequency of data. Privacy risks compound these challenges, as many datasets are unable to participate in traditional linkage processes. New linkage methods that can accurately link records without the need for personally identifying information are emerging as a potential solution to this privacy problem.

The primary objective of this research was to develop an operational cloud model for privacy-preserving record linkage (PPRL) utilising scalable computing infrastructure. Techniques for maximising the accuracy, privacy and performance of PPRL were developed and evaluated separately using both synthetic and real administrative data. A cloud model for record linkage is presented and evaluated, incorporating these PPRL techniques for improving accuracy, privacy and performance.

Research undertaken as part of this PhD study found that techniques that help maximise linkage accuracy of PPRL include probabilistic parameter estimation (using the expectation maximisation (EM) algorithm), partial field agreements and grouping methods that can leverage the quality of existing data. Combining these techniques was found to produce the best quality linkage. Other techniques for improving privacy and performance (such as multibit trees, homomorphic encryption and dynamic match keys) were evaluated, and while they all showed promise for use in some situations, there were trade-offs on either privacy, accuracy or performance.

Several cloud models were developed and presented as part of this thesis, all relying on the accuracy provided by PPRL. Evaluation of a hybrid cloud model that distributes linkage processing using managed containers and existing linkage software showed operational feasibility for scalable linkage compute at a reasonable cost. The additional analytical service capability of the cloud provides greater opportunity for advanced analysis, rich analytics, machine learning and automation.

Acknowledgements

After almost twenty years of working in the software development industry, I began working with the team at the Centre for Data Linkage (CDL) at Curtin University on developing some new linkage software. Apart from the interesting challenges of programming record linkage algorithms, this project exposed me to the world of health research. When offered an opportunity to extend my time with the CDL on research into record linkage methodologies, I jumped at the chance to again work with such a fantastic group of people. Thank you, Professor James Boyd, Associate Professor Anna Ferrante, Margo Gillies and Dr Sean Randall for making the decision such an easy one. I will always look fondly on those years at the CDL.

I want to give a special thanks to Professor James Boyd and Associate Professor Anna Ferrante, who have consistently guided my journey, providing invaluable advice, guidance and encouragement to move forward. Suffice it to say, without this I would not have embarked on this journey.

Thank you to Professor James Semmens, who also provided great encouragement and guidance in getting started. Your enthusiasm and belief in me were always very inspiring, even if your time estimate for me completing my thesis was somewhat ambitious.

Thank you to Professor Chris Reid, who graciously stepped in as supervisor upon Professor Semmens' retirement. I have appreciated your perspective and pragmatic advice.

I also have to acknowledge my amazing family, who have kept me grounded and are a constant reminder of what is most important in life. Mandy, thank you for your understanding over the many times I've worked into the night and on weekends. I promise I won't be writing another thesis anytime soon. Clara, my wonderful daughter, you always brighten up my day.

Research output from this thesis

This thesis contains the following peer-reviewed scientific publications. These papers have been published in quality medical and public health journals. This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

Published manuscripts:

1. **Brown AP**, Ferrante, AM, Randall, SM, Boyd, JH, Semmens, JB (2017). *Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage.* Frontiers in Public Health, 5(March), 34. https://doi.org/10.3389/fpubh.2017.00034

2. **Brown AP**, Randall SM, Ferrante AM, Semmens JB, Boyd JH (2017). *Estimating parameters for probabilistic linkage of privacy-preserved datasets* BMC Medical Research Methodology, 17(1), 95. https://doi.org/10.1186/s12874-017-0370-0

3. **Brown AP**, Randall, SM, Boyd, JH, Ferrante, AM (2019). *Evaluation of approximate comparison methods on Bloom filters for probabilistic linkage.* International Journal of Population Data Science, 4(1). https://doi.org/10.23889/ijpds.v4i1.1095

4. **Brown AP**, Borgs C, Randall SM, Schnell R (2017). *Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets.* BMC Medical Informatics and Decision Making, 17(1), 83. https://doi.org/10.1186/s12911-017-04785

5. **Brown AP**, Randall SM (2020). *Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model.* JMIR Medical Informatics, 8(9), e18920. https://doi.org/10.2196/18920

6. Randall SM, **Brown AP**, Ferrante AM, Boyd JH (2019). *Privacy preserving linkage using multiple match-keys* (2019) International Journal of Population Data Science, 4(1). https://doi.org/10.23889/ijpds.v4i1.1094

7. Randall SM, Ferrante AM, Boyd JH, **Brown AP**, Semmens JB (2016). *Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581?* Health Information Management Journal. https://doi.org/10.1177/1833358316647587

Conference proceedings:

8. Randall SM, Ferrante AM, Boyd JH, **Brown AP**, Semmens JB (2015). *Grouping methods for ongoing record linkage* (2015) Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference.

9. Randall SM, **Brown AP**, Boyd JH, Ferrante AM, Semmens JB (2015). *Privacy preserving record linkage using homomorphic encryption* (2015) Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference.

Published letters:

10. Boyd JH, Ferrante AM, Irvine K, Smith M, Moore E, **Brown AP**, Randall SM (2016). *Understanding the Origins of record linkage error and how they affect research outcomes* Australia and New Zealand Journal of Public Health. https://doi.org/10.1111/1753-6405.12597.

International conference presentations:

11. **Brown AP**, Borgs C, Randall SM, Schnell R (2016). *High quality linkage using Multibit Trees for privacy-preserving blocking.* Conference: International Population Data Linkage Conference, Swansea, Wales, August 2016.

12. Boyd JH, Ferrante AM, Randall SM, **Brown AP**, Semmens JB (2016). *Implementing privacy-preserving record linkage: welcome to the real world.* Conference: International Population Data Linkage Conference, Swansea, Wales, August 2016.

13. **Brown AP**, Randall SM, Ferrante AM, Boyd JH (2018). *Public Cloud: The Future of Record Linkage?* Conference: International Population Data Linkage Conference, Banff, Alberta, Canada, September 2018.

Competitive grants:

14. Boyd JH, Ferrante AM, **Brown AP**, Randall SM, Semmens JB Australia-Germany Joint Research Co-operation Scheme. DAAD 2016 (Funding commencing in 2017). Australia Universities. ($12,220)

15. Boyd JH, Ferrante AM, **Brown AP**, Semmens JB. National Collaborative Research Infrastructure Strategy (NCRIS 2016). Population Health Research Network. Department of Education and Training. ($485,465)

I warrant that I have obtained, where necessary, permission from the copyright owners to use any third party copyright material reproduced in the thesis (e.g. questionnaires, artwork, unpublished letters), or to use any of my own published work (e.g. journal articles) in which the copyright is held by another party (e.g. publisher, co-author).

Contents

List of Figures

List of Abbreviations

ACT	Australian Capital Territory
ASD	Australian Signals Directorate
AWS	Amazon Web Services
CDL	Centre for Data Linkage
CHeReL	Centre for Health Record Linkage
CLK	Cryptographic Longterm Key
CUPLE	CUstomisable Probabilistic Linkage Engine
DoHWA	Department of Health, Western Australia
DLU	Data Linkage Unit
ED	Emergency Department
EDDC	Emergency Department Data Collection
EM	Expectation Maximisation
FP	False Positive
FPR	False Positive Rate
FN	False Negative
GP	General Practice
HREC	Human Research Ethics Committee
HMAC	Hash Message Authentication Code
HMDS	Hospital Morbidity Data System
IaaS	Infrastructure as a Service
IPDLN	International Population Data Linkage Network
LSH	Locality Sensitive Hashing
MBS	Medicare Benefits Schedule
MLK	Master Linkage Key
NCRIS	National Collaborative Research Infrastructure Strategy
NSW	New South Wales
NT	Northern Territory
ONS	Office for National Statistics UK
PaaS	Platform as a Service
PHRN	Population Health Research Network
PopData	Population Data British Columbia
PPH	Potentially Preventable Hospitalisation
PPRL	Privacy Preserving Record Linkage
QLD	Queensland
RBF	Row-level Bloom Filter

RMSE	Root-Mean-Square Error
SA	South Australia
SAIL	Secure Anonymised Information Linkage Databank
SLK	Statistical Linkage Key
SMC	Secure Multi-party Computation
TP	True Positive
TTP	Trusted Third-Party
TN	True Negative
UK	United Kingdom
UWA	University of Western Australia
WA	Western Australia
WADLB	Western Australian Data Linkage Branch

Glossary

ad hoc linkage The linkage of two or more datasets for a specific purpose (typically a one-off project). Ad hoc data linkage does not involve the maintenance of a master linkage file or master linkage keys.

administrative data Information collected for the purpose of, or in the process of, service delivery; such as providing health care, responding to the legal requirements of registering particular events (e.g. births and deaths registration data) or providing a particular service.

agreement weight A weight applied to a field comparison if both field values are the same (or in some instances similar).

approximate comparison A comparison between values that uses a function to determine how similar they are. Common approximate comparisons include Jaro-Winkler, Sorensen-Dice, Jaccard and Hamming.

bigram An adjacent sequence of two characters from a string of tokens. Bigrams may be extracted from field values in record linkage to compare the similarity of values.

blocking A technique to reduce the number of comparisons required for matching by only comparing record pairs with one or more fields in common. Records from each dataset are placed into blocks based on specific fields and only records within each block are compared.

Bloom filter A probabilistic data structure that was originally created for efficient storage and checking of set membership. It can also be used to approximate the equality of two sets.

cartesian product The cartesian product of two sets A and B, denoted by A x B, is the set of all ordered pairs (a, b) where a is in A and b is in B.

cleartext Field values stored in cleartext (or plaintext) have not be encrypted or encoded in any way. In the context of record linkage, cleartext is used to denote values that have not been privacy-preserved.

clerical review	A manual review of record pairs whose link status cannot be automatically determined from linkage. Clerical review can also be used to obtain a quality assessment of a linkage.
cloud computing	A network of remote servers hosted on the Internet to store, manage and process data. Addition services and levels of abstraction are often provided to remove the management overhead of infrastructure.
comparison space	The comparison space, when linking two datasets A and B, are the set of pairs (a, b) that are compared during the matching process. It is impractical to compare every pair in the cartesian product of the two datasets, so indexing techniques are used to reduce the comparison space for record linkage.
conditional independence	In record linkage, conditional independence refers to the probabilities of field matches for a record being independent. In other words, the probability of a match on one field value has no bearing on the probability of a match on a different field value.
container	A container is a lightweight, standalone package of software containing all dependencies and requirements for execution.
cryptanalysis	The study of analysing information for the purposes of identifying weaknesses in cryptographic systems and algorithms.
cryptographic hash function	A one-way hash function suitable for use in cryptography, mapping data of any size into a fixed-sized value.
cryptography	The study of securing data against unauthorised access, including the techniques, methods and protocols used.
data cleaning	The process of editing data to remove errors such as illogical and out-of-scope values, and data entry errors, such as typographical errors and transposed values.
data custodian	The authority, body or person responsible for the safe custody, transport and storage of data, and implementation of business rules regarding use of the data. Data custodians may either have collected the data themselves or they may have legal and administrative custody of it on behalf of the owner or collector of the data.
data linkage unit	An organisation responsible for the linkage of data. A data linkage unit (DLU) often has close associations with data custodians and typically operate as a trusted third party, providing linkage services to researchers.

data standardisation	The process of making different datasets comparable and compatible, conforming to the same quality rules in terms of structure of dataset, scope, completeness, coding, structure and spelling of variable names, and range and format of data values.
deduplication	A technique for finding and removing duplicate copies of data. In record linkage, this is often used to describe the process for finding records belonging to the same individual within a single dataset.
de-identified information	Data that does not contain personal information, or from which the identity of the individual to whom it pertains cannot be reasonably ascertained.
deterministic linkage	Deterministic linkage ranges from the simple joining of two or more datasets by a reliable and stable key to a series of sophisticated rules for determining matches.
disagreement weight	A weight applied to a field comparison if the field values are not the same.
entity groups	Collections of records from across one or more datasets that represent the same entity or person.
expectation maximisation	An iterative method used to find the maximum likelihood estimates in statistical models. The expectation maximisation (EM) algorithm can be used to estimate likelihood estimates for fields in record linkage.
false matches	The record pairs found during a linkage that are incorrectly classified as matches.
false negative	A pair of records belonging to the same individual or entity that is incorrectly assigned as non-matches or as not belonging to the same individual or entity.
false positive	A pair of records belonging to two different individuals or entities that are incorrectly assigned as links.
false positive rate	The proportion of all record pairs belonging to two different individuals or entities that are incorrectly assigned as links.
F-measure	The harmonic mean between precision and recall.
grouping	The process of bringing record pairs, found during linkage, together into groups representing individuals or separate entities.
Hamming distance	An approximate comparison that identifies the number of positions within a field value that are different.
homomorphic encryption	A form of encryption that allows operations (addition and/or multiplication) on encrypted values (ciphertexts). The result of the operation must be decrypted to obtain the answer.

indexing	A technique used in record linkage to reduce the comparison space for matching. Indexing techniques aim to reduce the number of comparisons while maintaining maximum pairs-completeness (coverage of all true matches).
Jaccard index	An approximate comparison that returns the similarity of two sets by dividing the intersection of the sets by their union.
Jaro-Winkler comparison	An approximate comparison that returns the edit distance between two string values.
Levenshtein distance	An approximate comparison that returns the similarity of two strings by determining the number of edits (insertions, deletions or substitutions) required to change one value to the other.
linkage key	A code created and stored by a data linkage unit that is used to identify a group of records that refer to the same person or entity.
linkage map	A collection of linkage keys for one or more datasets.
linkage quality	A measure of the accuracy of the linkage in terms of true matches, false matches and missed matches. Precision, recall and f-measure are metrics for determining quality.
linkage strategy	The methods and parameters used for linking two or more datasets. This includes the fields that are used, the likelihood estimates for fields if probabilisitic, and the indexing technique used.
load balancing	A method for distributing computational load over two or more nodes.
MapReduce	A programming model specifically designed for processing large datasets in parallel on a compute cluster.
matching	The process of comparing record pairs and classifying them as matches or non-matches.
missed matches	The pairs of records that are represent the same individual but are missed during linkage.
m-probability	For a specific field, this is the likelihood that two records that represent the same person have the same field value.
multibit tree	A data structure used to store binary arrays that allows for efficient searching using a similarity function.
pairs-completeness	Used to measure the portion of true matches that are included by a particular linkage indexing technique.
partial weight	A value somewhere between the disagreement weight and the agreement weight that represents the similarity between two fields.

personal identifiers The field values often used in record linkage that can be used in whole (or in part) to identify a person.

population spine Refers to one or more linked datasets that have a high coverage of a population. It is often used in record linkage as a reliable, high quality dataset to link other datasets of unknown quality.

precision The proportion of all found matches that are true matches, as opposed to false matches.

privacy-preserving A method to protect the privacy of data by encoding it into a format that is unreadable by a person, yet usable by a machine.

probabilistic linkage A method of record linkage that uses the probabilities of agreement and disagreement between a range of linkage variables.

recall The proportion of true matches found out of all possible true matches.

record linkage The process of bringing together two or more sets of information belonging to the same person, event or place, into a single record of information.

record pair Any pair of records being compared to determine whether or not they belong to the same person or entity.

re-identifiable data Data that does not contain personal information, however, it is possible to re-identify the individual by linking the data to other datasets or by inferring information from the available data.

statistical linkage key A code used in data linkage that replaces a person's identifiable data to protect the person's identity. It is generated from elements of an individual's personal demographic data and attached to de-identified data relating to the services received by that individual.

Sorensen-Dice coefficient An approximate comparison that returns the similarity of two sets, by dividing the number of items in common by the sum of the number of items in each.

synthetic data Data that has been generated, often to statistically represent real data, to be used for testing of systems and algorithms where real data is unavailable.

threshold value A numerical value that determines whether some item is to be included or excluded. In probabilistic linkage, this often refers to a value that determines whether record pair comparisons are classified as matches or non-matches.

true matches The record pairs found during a linkage that are correctly classified as matches.

true positive	Two records that truly do correspond to the same person or entity.
trusted third party	In record linkage, a trusted third party (TTP) is an entity that facilitates linkage on behalf of two or more parties (data custodians).
truth set	A set of data with known values or answers. In linkage, this often refers to records with linkage keys known to be true.
u-probability	For a specific field, this is the likelihood that two records that do not represent the same person have the same field value.

Exegesis

Explanatory Overview

Background

Record linkage is the process of finding data that refers to the same entity within one or more datasets. Widely used in the health sector, record linkage of administrative collections has become a strategic research priority internationally. While record linkage has traditionally focused on health, the integration of administrative data across all government sectors has been recognised as an essential requirement for investigations of health and social outcomes, the effectiveness of service delivery, and policy development.

As the demand for record linkage increases, an important challenge is to ensure systems are scalable. Record linkage is computationally expensive, with a potential comparison space equivalent to the Cartesian product of the record sets being linked, making linkage of large datasets a considerable challenge. Optimising systems, removing manual operations and increasing the level of automation in such processes is essential for the process to be sustainable and scalable.

Current best practices in record linkage carry some privacy risk. These risks derive from the need to release personally identifying information to trusted third parties (specialised record linkage units). Legal and administrative constraints can also prevent trusted third parties from being able to link particular datasets, often due to the sensitivity of the data. New record linkage techniques, collectively referred to as privacy-preserving record linkage (PPRL), significantly reduce privacy risks associated with record linkage as they operate on de-identified information and do not require the release of personal identifiers.

Record linkage units typically manage security and privacy risks by hosting local linkage solutions on dedicated hardware. This approach leaves the record linkage units with expensive, dedicated equipment and computing resources that require managing, maintaining, upgrading and replacing regularly. Alternative solutions that can scale with the data *and* reduce privacy risk are required to meet the demands of record linkage moving forward.

Cloud computing infrastructure offers a solution for addressing the increased size of data for linkage. This infrastructure can be utilised with PPRL techniques to protect the privacy of individuals in the data, keeping personal identifiers local while 'pushing' the linkage of privacy-preserved datasets to cloud infrastructure. High levels of privacy on these privacy-preserved datasets are essential for consideration by operational linkage units. While cloud infrastructure may provide the elastic scalability required to support the increase in data, maintaining sufficient levels of privacy and linkage accuracy remain a challenge.

Study Aims

The purpose of this research is to develop a working cloud model of PPRL that utilises scalable computing infrastructure and demonstrates high-quality linkage without the need for named identifiers.

The research will evaluate emerging privacy-preserving linkage techniques and associated methods for maximising linkage quality. Many techniques used with probabilistic linkage methods remain untested with privacy-preserved data. These techniques will be evaluated and extended to support the differences and limitations of privacy-preserved datasets. Alternative and additional techniques for improving privacy and performance while maintaining linkage accuracy will also be assessed and presented.

The quality and privacy techniques developed will be incorporated into scalable matching algorithms that will form part of a larger cloud model for PPRL. Alternative cloud models will be explored to support the requirements of different data linkage scenarios.

The research will develop working solutions that utilise these new privacy-preserving techniques. It will evaluate these using real-world data (i.e. tested at linkage units in Australia, Canada and the UK), and demonstrate the effectiveness of these techniques on real-world projects.

Results

This thesis presents several cloud models for record linkage that leverage cloud computing infrastructure and utilise privacy-preserving techniques that maintain the privacy of the individuals in the data. These cloud models rely on the linkage accuracy provided by PPRL, so techniques that can reliably produce high-quality linkage have been evaluated, adapted and enhanced. Evaluation of distributed linkage using managed containers and existing linkage software showed this to be a viable first step for scaling linkage on cloud infrastructure. However, further optimisation of algorithms for distributing and scaling load on demand are required to use the cloud infrastructure to its full potential.

Techniques that help maximise the linkage accuracy of PPRL include probabilistic parameter estimation (using the expectation maximisation (EM) algorithm), partial field agreements and flexible grouping methods that can leverage the quality of existing linked data. The EM algorithm can be used to estimate probabilities on large PPRL datasets using a 10 percent sample and produces linkage results comparable to actual probabilities. An extension to the EM algorithm, presented in this thesis, provides an estimate of single threshold cut-off value to determine matches from non-matches. Mapping partial Bloom field comparisons to agreement weights using weight curves are generally consistent across datasets with different error rates but can vary slightly per field. Linkage quality produced using these weight curves on Bloom filters produced linkage results comparable to Jaro-Winkler string comparisons on clear-text data. The Weighted Best Link grouping method presented in this thesis can further improve linkage accuracy in situations where datasets are linked to an existing population spine of known good quality. Combining these techniques produced exceptional quality linkage.

A PPRL technique using CLKs (Cryptographic Long-term Keys) and multibit trees is presented to provide a method for private indexing and matching using a single composite Bloom filter. CLKs produced the highest accuracy when fields without missing values and those that often change (such as address) are excluded. At its best, CLKs provided improved privacy over

single-field Bloom filters, but this was at the expense of accuracy. The use of homomorphic encryption on Bloom filters produced the highest levels of privacy with accuracy equal to that of regular Bloom filter linkage; however, this was at the expense of performance. An alternative method to standard probabilistic linkage was also presented, called match keys, that traded additional privacy for reduced accuracy. Each method examined exposed a trade-off between privacy, quality and performance.

Case studies that utilise the quality techniques described in this thesis have shown F-measure quality metrics between 0.88 and 0.99. Data known to be of high-quality produced results at the high end of this, while the lower values appear to be a result of either poor-quality data or from a poor choice of matching parameters. Real-world projects that have used these quality techniques showed great success in a number of evaluation projects. The privacy of PPRL has enabled the linkage of data that would not have otherwise been possible. The quality techniques have ensured linkage accuracy of this data.

Conclusion

PPRL techniques and cloud computing infrastructure provide a solution to address the increasing size and complexity of data requiring linkage. High-level linkage accuracy must be obtained and maintained for this solution to be viable. This thesis demonstrates that there are sufficient techniques available to provide this. Estimation of parameters for probabilistic linkage, combined with partial agreement weights and targeted grouping methods ensures the best possible accuracy is achieved from the linkage.

Methods to further improve the privacy of PPRL are available, such as homomorphic encryption. While homomorphic Bloom filters significantly increased the privacy of fields and produced the same level of linkage accuracy as standard Bloom filters, there was a significant negative impact on performance. Improvements to performance appear to be possible. However, it seems the compute benefits of cloud infrastructure are currently nullified by the performance costs of operating on homomorphically encrypted data.

Real-world use of PPRL techniques described in this thesis has shown much success in achieving high linkage accuracy in Australia and internationally. In particular, these privacy and quality techniques have enabled large projects in several Australian states to link primary health care data to secondary care data. This level of vertical integration of health care data in Australia has not been possible until now.

The results of this work will lay the foundation for a new model of record linkage that does not require the use of named identifiers and can cope with the rapidly increasing size and diversity of administrative and clinical datasets.

Aims and Objectives

Privacy-preserving record linkage is an emerging technique which is being used to link datasets that have traditionally been difficult to obtain. However, balancing the privacy, quality and scalability aspects of PPRL remains a challenge. This research focuses on improving the quality and scalability aspects of PPRL while maintaining good privacy, examining the challenges associated with linkage quality and managing the ever-increasing size of linked data over time. The primary goal of the research is to establish a model for record linkage that uses PPRL techniques and cloud computing capabilities to enable population-level research into areas that have, up to now, been difficult to study owing to the challenge of accessing and linking 'hard to get' datasets.

The specific aims and objectives of this thesis are listed below.

Aim 1 - Understand the challenges associated with linking datasets in an industry where privacy is paramount, and datasets are growing in size and number.

> **Objective 1:** Review and compare developments in record linkage that address privacy and scalability to understand how these might be applied within an operational context.
>
> **Objective 2:** Review current linkage applications and frameworks that attempt to address privacy and scalability issues, identifying the linkage challenges they address and any gaps that may be present.

Aim 2 – Identify methods for maximising the quality of privacy-preserving record linkage that do not rely on manual clerical review.

> **Objective 3:** Review the factors affecting linkage quality during the different phases of the record linkage lifecycle to understand where improvements can be made.
>
> **Objective 4:** Refine and evaluate methods for improving linkage quality within a privacy-preserving context.

Aim 3 – Identify methods for improving privacy and performance and privacy-preserving record linkage.

> **Objective 5:** Evaluate privacy-preserving indexing techniques for improved privacy and performance.
>
> **Objective 6:** Prototype and evaluate the use of matching algorithms on encryption-based privacy-preserving techniques.

Aim 4 – Develop a model for record linkage that retains the privacy of data and utilises the scalability of cloud computing

> **Objective 7:** Review possible cloud models for record linkage, identifying advantages and disadvantages of each.

Objective 8: Establish algorithms for distributed privacy-preserving record linkage that can be utilised within a cloud computing environment.

Objective 9: Propose, prototype and evaluate a cloud-enabled model for record linkage suitable for operational use.

Aim 5 – Validate privacy-preserving techniques in real-world projects

Objective 10: Demonstrate the value of privacy-preserving record linkage through real-world case application of the technology.

Thesis Overview

This thesis is presented as a cohesive body of research, comprising peer-reviewed publications grouped into chapters to address the specified aims and objectives.

All algorithms were tested on synthetic data and real-world datasets. In the case of real-world data, evaluations were conducted within a secure research environment (ensuring compliance with the requirements of data custodians) to confirm that the methods were feasible for application in real-world settings. Access to real data required ethics and data custodian approvals, which were obtained as part of the project.

Chapter 1 – Introduction and literature review

Chapter 1 introduces the concepts of record linkage, covering standard linkage methods used and providing an overview of privacy-preserving record linkage techniques. Adapting and extending privacy-preserving techniques for scale and quality is a fundamental part of the research. Addressing the first aim of the thesis, the Chapter considers challenges faced by data linkage units in managing datasets that are continually increasing in size. Obtaining access to sensitive datasets and maintaining high-quality linkage at scale are two major challenges which are addressed through the research. A summary of existing privacy-preserving solutions is also presented and used as a foundation for the research to build upon.

The research in this chapter is supported by the following peer-reviewed scientific publication(s):

1. **Brown AP**, Ferrante, AM, Randall, SM, Boyd, JH, Semmens, JB (2017). *Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage.* Frontiers in Public Health, 5 (March), 1–6. https://doi.org/10.3389/fpubh.2017.00034

7. Randall SM, Ferrante AM, Boyd JH, **Brown AP**, Semmens JB (2016). *Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581?* Health Information Management Journal. https://doi.org/10.1177/1833358316647587

Chapter 2 – Maximising quality

Traditional record linkage processes include a clerical review step following the matching of datasets. This process identifies entity groups for review based on a series of rules that may indicate a grouping error. These candidate groups (and the records within them) are examined manually to determine if and how the records in these groups should be regrouped. However, this type of post-linkage quality process is not available for privacy-preserved datasets as the data are in an encrypted state. Moreover, these processes are becoming less practical for all linkages as data size increases and manual processes do not scale well.

Addressing the second aim, Chapter 2 looks at quality techniques that could be applied during other parts of the record linkage pipeline to reduce the need for post-linkage quality processes

significantly. Parameter estimation for probabilistic linkage occurs before matching, and approximate comparison methods are used during matching. Both techniques are aimed at improving the accuracy of the resulting record-pair matches, reducing the need for clerical review. This research project adapts these concepts for PPRL and evaluates them on real-world data. Techniques for combining these record-pairs into entity groups are also examined, as alternative methods to the traditional approach of transitive closure, to merge connected record-pairs, may be more appropriate.

The research in this chapter is supported by the following peer-reviewed scientific publication(s):

2. **Brown AP**, Randall SM, Ferrante AM, Semmens JB, Boyd JH (2017). *Estimating parameters for probabilistic linkage of privacy-preserved datasets* BMC Medical Research Methodology, 17(1), 95. https://doi.org/10.1186/s12874-017-0370-0

3. **Brown AP**, Randall, SM, Boyd, JH, Ferrante, AM (2019). *Evaluation of approximate comparison methods on Bloom filters for probabilistic linkage.* International Journal of Population Data Science, 4(1).

8. Randall SM, Ferrante AM, Boyd JH, **Brown AP**, Semmens JB (2015). *Grouping methods for ongoing record linkage* (2015) Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference.

10. Boyd JH, Ferrante AM, Irvine K, Smith M, Moore E, **Brown AP**, Randall SM (2016). *Understanding the Origins of record linkage error and how they affect research outcomes* (2016) Australia and New Zealand Journal of Public Health. https://doi.org/10.1111/1753-6405.12597.

Chapter 3 – Privacy and performance

Privacy-preserving techniques are typically slower than traditional methods as they require considerably more computation. Chapter 3 addresses the third aim by developing and assessing new methods that improve both security and performance of PPRL.

Efficient indexing techniques that reduce the comparison space while maintaining high pairs-completeness are critical for record linkage of large datasets. The research presented in this Chapter evaluates the performance of indexing a single composite Bloom filter (cryptographic long-term key) using multibit trees.

This chapter also evaluates some new techniques for privacy-preserving record linkage that improve the security of privacy-preserved data while attempting to maintain high linkage quality. Homomorphic encryption is applied to Bloom filters, and a new technique (multiple dynamic match keys) is presented and evaluated.

The research in this chapter is supported by the following peer-reviewed scientific publication(s):

4. **Brown AP**, Borgs C, Randall SM, Schnell R (2017). *Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets.* BMC Medical Informatics and Decision Making, 17(1), 83. https://doi.org/10.1186/s12911-017-04785

9. Randall SM, **Brown AP**, Boyd JH, Ferrante AM, Semmens JB (2015). *Privacy preserving record linkage using homomorphic encryption* (2015) Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference.

6. Randall SM, **Brown AP**, Ferrante AM, Boyd JH (2019). *Privacy preserving linkage using multiple match-keys* (2019) International Journal of Population Data Science, 4(1). https://doi.org/10.23889/ijpds.v4i1.1094

Chapter 4 – Cloud models

Use of cloud services for record linkage has been rejected historically, with data custodians considering the risk of storing personal identifiers on cloud infrastructure as unacceptable. As a result, record linkage solutions have developed into on-premise systems that rely on increasing resources (CPUs and memory) to cater for large datasets. However, recent moves by governments across Australia to embrace cloud infrastructure has shown that there is no longer such a high level of perceived risk with these environments. Chapter 4 addresses the fourth aim by examining how cloud services can be used to address the 'big data' issue within record linkage without the additional risks associated with release of raw identifiers to cloud infrastructure. Different cloud models for record linkage are put forward, each ensuring that personal identifiers are kept local (on-premises) while utilising the advances of scalable cloud infrastructure for linkage and computation.

Research in this chapter is supported by the following peer-reviewed scientific publication(s):

5. **Brown AP**, Randall, SM (2020). *Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model.* JMIR Medical Informatics, 8(9), e18920. https://doi.org/10.2196/18920

Chapter 5 – Implementation/translation

Chapter 5 addresses the fifth aim, describing the adoption and application of privacy-preserving linkage in real-world settings. One translation linkage was used for a research study to evaluate the influence of patterns of primary care contact on emergency department (ED) visits and potentially preventable hospitalisations (PPHs). This project linked state-level ED and hospital data with Commonwealth Medicare Benefits Schedule (MBS) data. Without PPRL by a trusted third-party the study would not have been possible.

A large repository of de-identified information on individuals who have come into contact with key government agencies has been created in Western Australia to improve access to linked data for both researchers and policy analysts. Due to legal and privacy concerns, a number

of datasets were not available for linkage using raw personal identifiers. Incorporating PPRL into this project has enabled the linked data repository to include all of the key government agencies necessary to achieve the project's aims.

Access to primary health care data for linkage to secondary care data by state linkage units has been almost impossible to achieve in Australia. However, a privacy-preserving linkage of primary and secondary care data can help to derive valuable insights and enable better patient outcomes. The organisations involved in two large projects have embraced PPRL (and the techniques developed in this thesis) to help track patient journeys through the primary and secondary health care systems. NPS MedicineWise is building capability to provide a linked data warehouse with de-identified data from as many as 730 general practices. The Lumos project being run by NSW Ministry of Health aims to link data from as many as 500 GP practices to secondary care data. Both of these projects are well under way, and demonstrate the growing need for vertical integration of health data. Cloud models for PPRL that provide sufficient privacy will necessarily scale with the increased demands of this integration moving forward.

Chapter 1

Privacy-preserving record linkage

Included Manuscript(s):

1. **Brown AP**, Ferrante, AM, Randall, SM, Boyd, JH, Semmens, JB (2017). *Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage.* Frontiers in Public Health, 5 (March), 1–6. https://doi.org/10.3389/fpubh.2017.00034

7. Randall SM, Ferrante AM, Boyd JH, **Brown AP**, Semmens JB (2016). *Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581?* Health Information Management Journal. https://doi.org/10.1177/1833358316647587

Record linkage is a process which enables the collection and integration of person-based records for the same individual [205]. In places where a unique personal identifier exists (such as in Denmark, where a national person identifier is an accepted part of modern life [284]) the process of integration is trivial, with records simply joined through this unique identifier. Where unique personal identifiers are not available, personally identifying information such as name, date of birth and address are often used to determine if records belong to the same person. However, the ability to join records is limited by error, omission or change within these fields [96].

Record linkage includes a number of techniques for creating links between "pieces of information that are thought to relate to the same person, family, place or event" [62]. These techniques include basic sort and match algorithms, deterministic rule-based methods, and probabilistic approaches [27, 59, 84, 141, 295, 292]. Linking records across administrative data collections enables researchers to construct individual chronological histories from birth to death and undertake studies that deliver significant public benefit [42, 49]. Record linkage of these administrative collections has become a strategic research priority within Australia and internationally [38], and is widely used in the health research sector to gain event-based longitudinal information for entire populations [41].

With the advent of 'big data' analytics, record linkage is becoming an essential and well-used research tool for epidemiologists [193]. By linking together different administrative data collections (for instance, hospital admissions, emergency presentations, disease registries, and birth and death records), a detailed picture emerges of an individual's lifelong health. As administrative collections typically capture an entire population, this data allows researchers to answer numerous important health questions at low cost [144].

1.1 Stages within the record linkage process

Irrespective of which linkage technique is employed, the fundamental stages of record linkage are the same [62] and are shown in Figure 1.1. Typically, datasets require a level of preprocessing as a first step, cleaning and standardising the data to ensure consistency with the formatting of fields. The second step reduces the number of record-level comparisons required (often referred to as the comparison space), by indexing data in a way that removes comparisons that are unlikely to be true matches. This typically involves grouping the data into overlapping blocks or clusters based on sets of field values and can provide up to 99% reduction in the comparison space. Record pair comparisons occur next; the comparisons are carried out within the blocks or clusters determined during the indexing step. Classification of record pairs into matches, non-matches and potential matches result in groups of entities (or individuals) based on match results. Non-matches are discarded, matches are used to group records into entities, and potential matches will often be assessed manually or through special tooling to determine whether they should be classified as matches or non-matches. A common approach to grouping matches is to merge all records that link together into a single group; however, different approaches can be used to reduce linkage error [160]. Analysis of the entity groups

is the last step, where candidate groups are clerically reviewed to determine if and how the records should be regrouped.

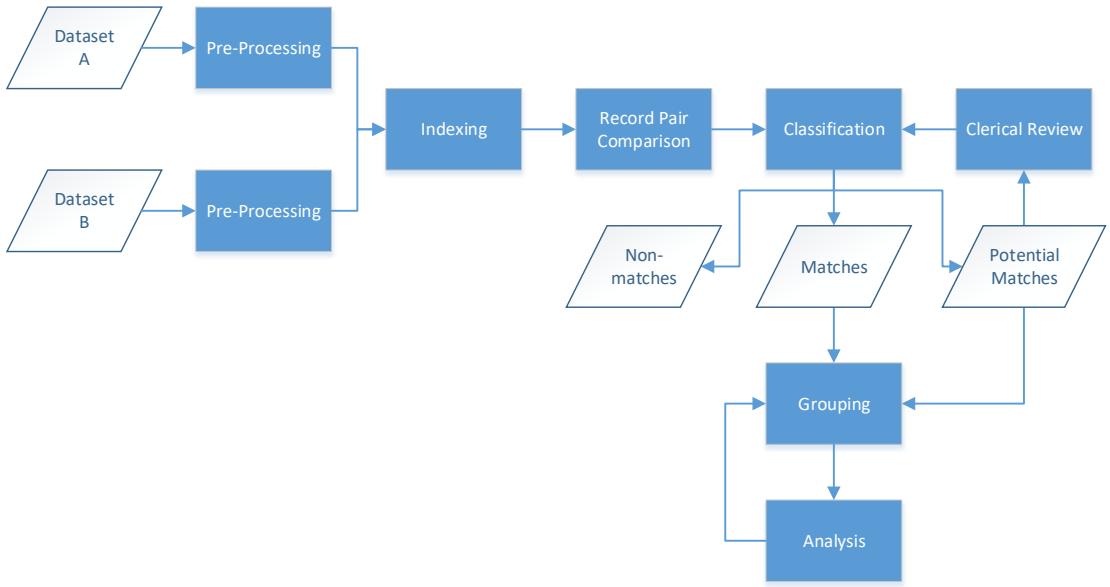


FIGURE 1.1: Record linkage process

Within established record linkage operations, there are two main linkage scenarios - project-based and on-going linkage [41]. A project-based linkage occurs when one or more datasets are linked together for a single research project. On-going linkage refers to a model where data are added over time, and a master 'linkage map' is continually updated. This may result in updates to some records (including deletion) and requires the data linkage process to cater for changes to existing groups based on new information. The quality of the linkage map often improves through on-going linkage as clerical review and analysis processes are established and refined over time. The resultant linkage map can then be used for multiple research projects [129, 179].

1.2 Deterministic record linkage

A traditional deterministic approach to record linkage involves the matching of records based on agreement between a collection of identifiers from each record [115]. Some advanced methods have sophisticated step-wise logic that allows for variation in the fields being compared [182] and even learning rules based on existing links [138].

Deterministic or rule-based matching are commonly used linkage methods [33, 214]. This is despite studies showing that probabilistic techniques result in better linkage quality [69, 111, 213], particularly as the quality of the data decreases. Linkage results from deterministic or rule-based matching often have a high rate of true matches as the data being compared is

exactly the same. Typographical errors, spelling mistakes and other variations in field values inevitably result in a high rate of missed matches. This outcome can be addressed through the use of multiple deterministic schemes with different collections of identifiers. However, in some scenarios, the high rate of missed matches may be acceptable [126].

1.3 Probabilistic record linkage

Probabilistic linkage is a widely used and robust technique for record linkage that uses conditional probabilities to determine the likelihood that particular records belong to the same individual. These methods were first developed and used by Newcombe [203], and mathematically formalised by Fellegi & Sunter [96].

Within probabilistic record linkage, individual records are compared on a pairwise basis. Comparing all records in one dataset with all records in another dataset would result in an excessive number of comparisons for all but small datasets. Therefore, records are only compared if they have certain pieces of information in common. That is, they have the same value in a particular field or set of fields. This type of indexing is known as blocking. Blocked records are compared and assessed through comparison of the values in all the individual fields (e.g. first name, surname, address, etc.). Each field comparison results in a field score, the value of which depends on whether these fields agree or disagree. These agreement and disagreement scores (weights) are computed separately for each field. All field scores are summed to determine a final score representing how likely two records being compared belong to the same person. Classification of these record pair scores is typically done through two predefined threshold values. If the field score is less than the lower threshold value, the record pair is designated as a non-match and discarded. If the score is higher than the higher threshold value, the record pair is designated as a match. All record pairs within the two threshold values are then considered potential matches and require some kind of clerical review to determine whether they should be classified as matches or non-matches.

Figure 1.2 presents a simplified example of a single record-pair comparison; both first name and date of birth match exactly, resulting in field-level agreement scores of 5 and 12. The surname, address and suburb all disagree, leading to disagreement weights of -4, -3 and -3. These scores are summed, resulting in a total score of 7. As this score is higher than a set threshold of 6, this record-pair is classified as a match.

Threshold	6

Record 1	ADRIAN	BROWN	22 FOURMILE AVENUE	BURNS BEACH	23-Jun-83
Record 2	ADRIAN	BROUN	4 SILVERWOOD STREET	MORLEY	23-Jun-83
Agreement Weight	**5**	10	16	3	**12**
Disagreement Weight	-2	**-4**	**-3**	**-3**	-4

Total
7

FIGURE 1.2: Probabilistic linkage

The agreement and disagreement scores used in field comparisons are based on the calculation of two specific probabilities, called the m-probability and u-probability [205].

The m-probability is the likelihood of two fields matching if the records belong to the same individual. For instance, fields such as gender are likely to be the same on two records from the same person, assuming a small recording error this probability could be as high as 0.999. Fields such as address are more likely to vary, and as a result, the m-probability will be lower.

The u-probability is the likelihood of two fields matching if the records do not belong to the same individual. For example, the probability of the gender field matching on two records belonging to different individuals is 0.5 (assuming our dataset contains equal numbers of males and females). On the other hand, the probability of the address field matching on two records belonging to different individuals is extremely low.

These two probabilities are converted into the agreement and disagreement weights found in Figure 1.2 as follows:

$$AgreementWeight = log(\frac{m}{u}) \tag{1.1}$$

$$DisagreementWeight = log(\frac{1-m}{1-u}) \tag{1.2}$$

Missing values are typically treated in one of three ways: a comparison involving a missing value is either assigned the disagreement, a zero weight, or a separate weight accounted for explicitly [244]. The last option extends the conditional independence assumption to include probabilities for missingness, changing the calculations for weights. Other solutions to the missing data problem involve imputing values from other records [110], removing the field from the matching or even removing the entire record [62].

1.3.1 Parameter estimation

Probabilistic linkage, while providing very high-quality linkage results, requires some input parameters. These include accurate weight estimates (derived from m-probability and u-probability for all fields) and accurate threshold values.

The calculation of appropriate weights requires the estimation of appropriate conditional probabilities, which are at least somewhat intuitive and understandable. Several methods have been developed to estimate m and u probabilities [140, 291]. A simple and highly accurate method for approximating u probabilities was proposed by Jaro [140]. The number of true record-pairs is typically dwarfed by the number of incorrect record-pairs, (in the administrative data used later in this paper, around 0.001% of possible record-pairs belong to the same individual). As such, the conditional probability for calculating the u-probability (where we are interested in only the incorrect pairs) can be ignored, and we can instead calculate the probability of any two fields having the same value, which is straightforward.

There are no equivalent techniques for estimating m-probabilities; in practice, most methods are based on investigations around data quality and previous experience, such as the iterative refinement procedure [205]. Automated methods for deriving m-probabilities, such as through EM estimation have been devised [19, 291], and while their results appear strong, their evaluation and use so far is limited.

There are some observed challenges with the EM algorithm that may affect its use in practice. Winkler [291] notes that if the Conditional Independence Assumption is not valid, linkage based on the m and u probabilities may not be optimal. Also, if the proportion of matches falls below 0.05, the EM algorithm will also likely fail to identify accurate m and u probabilities [299]. This could be because the EM algorithm is run over a large file without blocking, or the blocks selected are too wide [140, 244].

Despite these challenges, the EM algorithm has the potential to provide accurate estimates for m probabilities, in some cases outperforming the probabilities obtained via the iterative refinement procedure [296]. Even if the EM algorithm produces accurate m probabilities, blocking techniques will impact the accuracy of u probabilities. The use of blocking is important to help estimate accurate m probabilities, but greatly reduces the number of non-matches observed, contributing to biased u-probability estimates [140]. As such it is recommended that Jaro's method for calculating u probabilities on unblocked data, described above, is always used.

Other estimation methods do exist [96, 133] but are more sensitive to initial parameters and require adjustment functions to keep values within bounds [140].

Determination of the appropriate threshold setting above which to accept record-pairs as valid matches typically occurs through manual, clerical review of record-pairs at a range of threshold scores. The focus is usually on the area of high overlap between the weight distributions of true matches and true non-matches [62]. Note, however, that threshold scores provide a relative and not an absolute measure of the likelihood of the record-pair belonging to the same individual.

Apart from a manual examination, or choosing a score based on prior experience, few methods currently exist for determining appropriate threshold settings. There are some software packages that estimate m and u probabilities using the EM method; however, these packages provide no method for determining threshold settings.

1.3.2 Approximate comparisons

In a simplified model for record linkage, every field level comparison is a binary comparison. That is, the fields match precisely, or they don't at all. In reality, numerous comparisons can be made depending on the type of field that is being compared. Many of these comparisons will calculate the similarity between two values, allowing for typographical errors and misspellings. The Jaro-Winkler, Levenshtein, Jaccard and Hamming distance measures are all examples of comparisons used in record linkage for the comparison of string values like name

and address [62]. Other, more specialised comparisons may be used on fields like date of birth, allowing for transposition of day and month values or even allowing a small tolerance amount for the year of birth [63].

Extensions to the Fellegi-Sunter model have been developed for approximate comparisons, allowing the assignment of a partial weight somewhere between agreement and disagreement [89]. While there are many types of approximate comparisons for various types of data, most deal with the distance between two strings [62]. The distance is converted into a partial weight to fit the approximate comparisons into a probabilistic model [296].

1.4 Mixed linkage techniques

Many data linkage units utilise both deterministic and probabilistic record linkage techniques, leveraging the advantages of both [3, 4, 134, 143, 157, 247, 259]. Fast, deterministic passes are first used to match the bulk of the data, with rules that are able to capture duplicate and near-duplicate records. Probabilistic linkage techniques are then used on the remaining data to maximise match rates between datasets.

1.5 Privacy-preserving techniques

The primary issue with record linkage in any environment is privacy and confidentiality of data. The National Statement on Ethical Conduct in Human Research (2007) defines data in three categories:

- individually identifiable data (the identity of a particular individual can reasonably be ascertained),
- re-identifiable data (identifiers have been removed and replaced by a code, but an individual can be re-identified using the code or linking to other datasets), and
- non-identifiable data (personal identifiers have been permanently removed).

Data custodians, researchers and data linkage units have worked together to develop data access and usage models that comply with information privacy laws and provide necessary guards to privacy (e.g. Australian Government High Level Principles for Data Integration [9]). Data linkage units typically work with individually identifiable data and have implemented an array of best practice data governance policies to minimise the risk to privacy posed by their operations [38, 99, 121, 179, 239, 272].

The paper, *Ensuring Privacy When integrating Patient-Based Datasets: New Methods and Developments in record Linkage*, included as part of this thesis, discusses some of the challenges of record linkage with respect to privacy, risk and data sharing, and how problems with current record linkage practices impact the ability to provide linkage of particular datasets. Legal, administrative and technical issues can prevent linkage from occurring, the risks being deemed too high to release these datasets to linkage units. In recent years, privacy-preserving record linkage

(PPRL) methods have emerged that reduce the risk of identity disclosure by operating on information that has been cryptographically hashed, encrypted or transformed in some way. These methods do not require the release of personally identifying information by data custodians; rather, data custodians use specialised encoding processes to transform personally identifying information into a permanently non-identifiable state (an irreversible "privacy-preserved" state). Under a trusted third party linkage model, this operation occurs before the release of any data to data linkage units. Thus, personally identifying information is not disclosed by the data custodian. These PPRL methods can be used within existing record linkage frameworks, and are subject to some of the same challenges.

A growing number of novel PPRL techniques have emerged in the literature, with a recent review summarising nearly thirty variations [281]. These protocols differ in their method of preserving privacy, scalability, error tolerance and security. However, very few of these have been practically evaluated for use in operational record linkage settings [238]. For PPRL techniques to be considered a viable option in an operational context, they must be not only secure but also highly accurate and efficient. In practice, many PPRL techniques are vulnerable to frequency attacks [206, 252, 281], whereby the frequency of occurrences of encrypted values reveals information. Where stronger security guarantees are required, improved or hardened PPRL algorithms are needed. The most prominent techniques currently in use are the SLK-581, hashing, Bloom filters and secure multi-party computation (SMC).

1.5.1 SLK-581

In Australia, statistical linkage key (SLK) protocols have proved to be a popular method of providing privacy protection in linkage operations. An SLK is a derived variable generated from components of an individual's personal demographic data [12]. The most well-used SLKs is the SLK-581 – a key originally developed by the Australian Institute of Health and Welfare (AIHW) and used to enable linkage of records from the Home and Community Care (HACC) data collection [242]. The SLK-581 was developed to enable linkage of a person's records within and between datasets without explicitly identifying individuals. A number of variations were assessed during the development of the key; however, the SLK-581 proved to be the most effective for the linkage of HACC data, given the particular properties of that dataset [242]. The SLK-581 consists of the second and third letters of the individual's first name, the second, third and fifth letters of the surname, the individual's full date of birth and their sex. These are amalgamated to form a single field.

Matching rules involving SLKs typically only identify matches where a large proportion of attributes are identical. As a result, low rates of false-positive links have been observed [242]. However, a number of studies have found high rates of missed links [18, 62, 153], that can increase substantially over time [267]. While the SLK has been used for linkage in numerous research projects, the ability of the key achieving high linkage quality has been an issue of continuing concern. In practice, intricate linkage techniques [154], along with additional identifying information that does not form part of the SLK-581, such as postcode, language spoken

at home and country of birth [153], have been used to improve the linkage quality above levels achievable with the standard SLK-581.

While a number of studies have investigated the linkage quality of the SLK-581, much less has been written about the privacy risks associated with use of the key. Although the SLK-581 was developed to reduce the risk of identification of individuals, it consists entirely of unencrypted identifying information specifically selected to be unique for each individual. Although the SLK-581 can be encrypted (or cryptographically hashed) to improve privacy protection, concerns over the resulting linkage quality have meant that this method was not recommended for use by its original developers [12]. At the time of its development, the SLK-581 offered a reasonable option regarding safeguarding privacy, while still allowing linkage to occur.

The paper, *Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581?*, evaluates the SLK-581 and the encrypted SLK-581 for both privacy and for linkage quality, comparing them to a newer approach using Bloom filters. The results showed that the standard SLK-581 has no mechanisms that can prevent a specific individual from being located within a dataset, and both the normal and encrypted version of the SLK-581 are unable to cope particularly well with differences in data. Changes to names, even small typographical errors, can prevent matches from being found. Emerging methods for PPRL provide better privacy and return higher linkage accuracy results.

1.5.2 Hashing

In the context of record linkage, the term 'hashing' is associated with a privacy-preserving record linkage technique that uses a one-way encoding function to transform identifiers into a privacy-preserving state [62]. The resultant hashes can then be used in place of identifiers during the record linkage process. However, as even small variations in strings will result in completely different hashes, only exact comparisons can occur between fields.

Early approaches to hash-encoding for privacy used SHA and MD5 hash functions [88]. For example, an MD5 hash of the string "Jonathan" always returns the value "c1f80eddea77f14650a2062dda3eb15c". The idea behind the hash-encoding was that names and addresses could not be identified through human inspection but could be matched through linkage. However, this approach is susceptible to dictionary attacks [227]; a simple brute-force approach can quickly identify most names. Mitigation of these attacks requires additional mechanisms such as the addition of keys to the strings before hash encoding [62, 228]. The development of hashed message authentication codes (HMACs) has helped in this regard. An HMAC extends an existing hash function (such as SHA256) to provide a keyed hash [169]. A secret key is used to derive two new keys, which are used in separate passes during the HMAC computation. This provides better protection against traditional dictionary attacks as well as extension attacks [20].

Hash-encoded identifiers work well in deterministic record linkage, where only exact comparisons are made. In this scenario, it is also advantageous to combine the collections of identifiers used for deterministic linkage into single hash values, thereby improving on the privacy of the

method as well as the speed of the matching process itself [22]. Probabilistic methods also work well with hash-encoded identifiers. However, each field requires a separate hash-encoding to satisfy the independence assumption of the model.

1.5.3 Bloom filters

A Bloom filter is a probabilistic data structure that was developed to check set membership but can also be used to approximate the equality of two sets. The protocol for privacy-preserving record linkage protocol using Bloom filters on q-grams of identifiers was proposed by Schnell et al. [252].

Construction of a Bloom filter begins with an array of a set length, with all elements set to zero. Each field (e.g. first name) is broken down into overlapping sets of letters (q-grams). Padding is often used to distinguish the first and last letters. Each of these q-grams is passed through a series of cryptographic hash functions. A hash function is an algorithm which produces a fixed-length output with several important properties. Firstly, given the same input, it will always produce the same output (i.e. the same q-gram will produce the same hash value). Secondly, the hash function is one-way, meaning it is not possible to determine the encoded q-gram from any given hash value (i.e. it is irreversible). HMACs are typically used in the construction of Bloom filters for linkage (similarly to standard hashing) so that the output is derived from a secret key.

The modulus of these hashes is then computed on the length of the Bloom filter. This process allows each q-gram to be mapped to one or more positions in the Bloom filter. These positions are then set to 1 (see Figure 1.3).

A Bloom filter can be constructed from any number of identifiers. The Cryptographic Long-term Key (CLK), for example, is a single Bloom filter in which multiple identifiers are stored to produce an anonymous linkage code [251]. Each identifier uses different hash functions, the number of hashes and passwords [253]. The Composite Bloom filter, as proposed by Durham [86], is a record-level Bloom filter (RBF) made up of all fields within a row. Each field has apportioned a slice of the Bloom filter based on the discriminatory power of that field. A mechanism for mitigating cryptographic frequency attacks prevents infrequently set bits in each field from being used in the final RBF.

While Bloom filters comprised of multiple fields suggest improved privacy [86, 206], single field Bloom filters are simpler to incorporate into existing probabilistic linkage frameworks. Field-level probabilities for matches and non-matches remain the same, and it is more likely that data linkage units can make this small step into incorporating privacy-preserving linkages into their existing linkage models.

Comparitors for Bloom filters include the Sørensen–Dice coefficient [252], Jaccard (or Tanimoto) index [16] and Hamming distance [163]. The Sørensen–Dice coefficient and Jaccard index measure similarity between values. Scores range from 0 to 1, where higher values represent greater similarity, and a score of 1 represents identical values. Hamming measures the number of bits

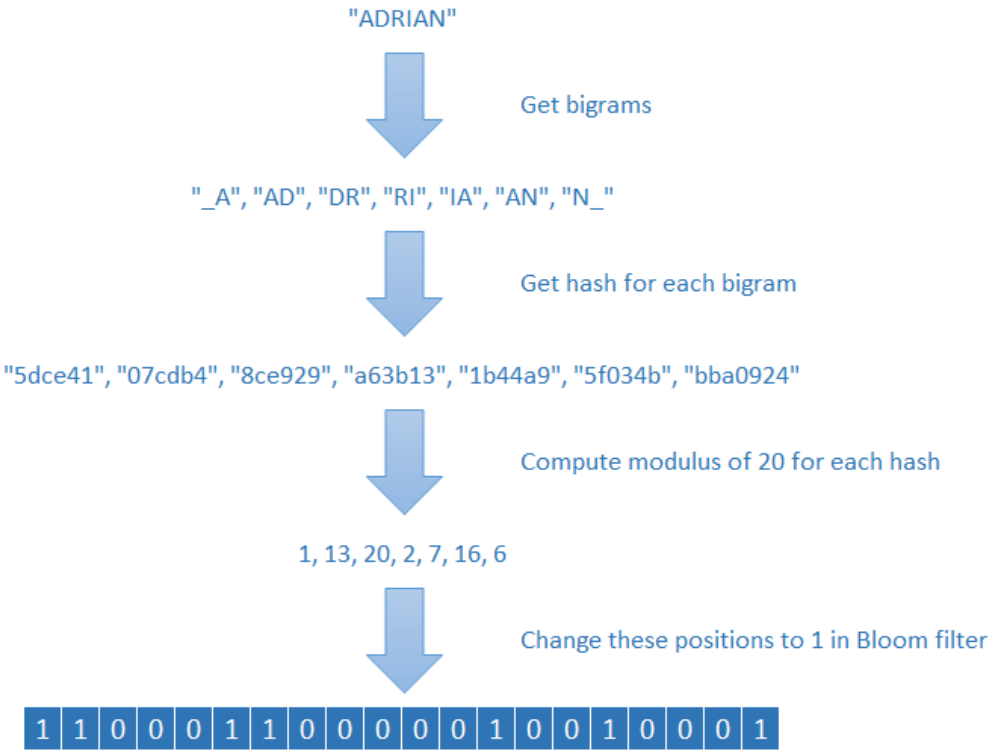


FIGURE 1.3: Creating a Bloom filter

that are different between values. Scores can range from 0 to the length of the Bloom filter. Lower values represent greater similarity with identical values having a score of 0.

Given two Bloom filters, A and B, similarities are calculated as follows:

$$Dice(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{1.3}$$

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1.4}$$

$$Hamming(A,B) = |A \oplus B| \tag{1.5}$$

These similarity comparisons should fit into existing linkage models in much the same way as standard string similarities. However, there is little mention in the literature of Bloom filters being used in the context of probabilistic record linkage, where a similarity value requires conversion to a field weight. The RBF and CLK composite Bloom filters both use a mechanism for relative field weighting based on hash counts and a simple threshold cut-off value on comparison to determine matches [86, 251].

While some privacy-preserving algorithms have been proposed [281], few have been evaluated regarding their privacy, efficiency and accuracy [238]. The ability to provide similarity comparisons on the data is highly desirable for accurate linkage, and evaluations of Bloom filters in

large-scale probabilistic record linkage have shown high linkage accuracy with relatively good efficiency [238].

However, the similarity-preserving nature of Bloom filters presents some security considerations, as this property can be exploited to launch an attack against the encryption and potentially reveal personal identifiers. In recent years, several attacks have been published. The first, proposed by Kuzu [173], revealed personal identifiers by performing a frequency analysis of individual fields. A discussion on the scope and limitations of the attack by Schnell and Borgs [254] revealed that this type of attack relies on aligning frequency distributions of the entire Bloom filter with unencoded identifiers; reducing, or eliminating, frequencies with the use of salted encodings will render this attack inneffective. A second attack, devised by Niedermeyer et al. [206] and extended by Kroll and Steinmetzer [172], focuses on the frequency distributions of the bit patterns of Bloom filters, including CLKs. The attack was successful in decoding CLKs using the double-hashing scheme as proposed in the original publication [252]. However, replacing the double-hashing scheme with full random hashing prevents such attacks [254]. Several other hardening techniques have been proposed to make Bloom filters more resilient against bit-pattern based attacks [249, 255]. Reducing the frequency of bit patterns by salting with record-specific values has also been suggested [67].

An extension called Counting Bloom filters uses three or more bits for every single bit in a standard Bloom filter to store the number of elements inserted [26, 94]. The counting of element insertions reduces the probability of false positives and also allows for elements to be deleted. The use of Counting Bloom filters has been evaluated in some studies to improve privacy for multi-party linkage and to reduce the computation and communication costs involved [146, 283].

An ideal PPRL technique would provide no possible mechanism for any individual to learn information about the personal identifiers used in the linkage process. Recent research has been published with some improvements to standard Bloom filters for the purposes of additional security [206, 260]; however, the impact on linkage quality has yet to be examined.

1.5.4 Secure multi-party computation (SMC)

Separation of data for linkage and data for research has been an important principle adopted by linkage organisations to help protect the privacy of the individuals in the data [158]. Privacy-preserving techniques such as cryptographic hashing and Bloom filters have helped in this regard. However, these techniques typically rely on a trusted third party to link the data. Linking datasets between two or more parties, without the need for a trusted third party, presents additional challenges [28].

Methods for sharing computation between two or more parties have been developed to ensure privacy between each party. These secure multi-party computation (SMC) methods make use of encryptions schemes to ensure privacy, removing the need for a trusted third party. Some encryption schemes used include commutative [183], functional [112] and homomorphic [76] encryption.

The foundations for SMC were laid by Yao [300] who developed a method for secure computation for two parties. This was extended to a generalised multi-party approach by Goldreich [109]. However, it is only recently that practical SMC solutions have been devised, as most work has focused on achieving security with zero disclosure [82]. Schemes like SPDZ [76] and the advances of homomorphic encryption have made SMC a viable solution for some scenarios.

Some SMC techniques have been developed and evaluated for use in privacy-preserving record linkage [150, 175, 281, 297]. However, their use appears limited [281], in part due to the computational overhead in applying these on datasets, which becomes impractical for record linkage with large data.

1.6 Quality assurance

Quality metrics provide a measure of the ability of a linkage technique to classify record-pairs correctly into matches and non-matches [62, 97]. In real-world scenarios, high linkage quality is critical, as research outcomes and policy decisions are dependent upon them. However, high accuracy can be difficult to achieve due to recoding errors, missing values, and outdated information in the data records [240]. The pre-processing phase of the data linkage process typically involves extensive data cleaning and standardisation to address these issues. This includes reformatting, correcting and removing data from fields based on their values [237]. The cleaning step has been shown to improve the quality of linkage [117, 126, 140]. However, it can take as much as seventy five percent of the effort of the data linkage process [104], and overdoing the data cleaning can result in decreased linkage quality [237].

The clerical review phase of the data linkage process is another time-consuming task that for some linkages can comprise months of work for dozens of clerks [126]. Clerical review involves manual examination of potential matches (record pairs) and a decision that determines whether these potential matches should be reclassified as matches or non-matches [140]. As the number of potential matches can be overwhelmingly large for human assessment, sampling methods for probabilistic linkages have been developed to reduce the number of manual checks required [43, 104, 118]. Quality metrics are determined on samples at different thresholds, and the observations on these samples are then used to determine cut-off values for the whole linkage.

The use of PPRL methods, where only encrypted identifiers are used for linkage, preclude the use of manual inspection, so must rely on the use of computerised methods to determine cut-off values. The only reference to threshold estimation in the literature is by Jaro [140], where he describes the estimation of a lower and upper threshold value, based on desired probabilities for a mismatch. Weights between these thresholds are undecided and are flagged for clerical review. As such, this method is not directly transferrable to PPRL.

A major challenge in the adoption of privacy-preserving methods is to achieve and maintain high accuracy of results. Little research has been undertaken on quality assessment and improvement of PPRL results. Most studies have focussed on improving the quality of the linkage through various automated processes such as indexing (blocking) methods [16, 27, 60, 282]. Some research has explored the use of graph theory for identifying errors that do not rely on identifiable data itself [124, 238]. The ability to correctly estimate linkage parameters is of paramount importance for these PPRL techniques to be practical [238].

1.7 Scaling for demand

As the demand for data linkage increases, the main challenge will be to ensure systems are scalable. An important focus for data linkage capability during the next decade is to develop the infrastructure capacity for the integration of cross-jurisdictional data across agencies, as the evaluation of many health issues involves the sophisticated analysis of data from many government sectors [224]. The increased demand for linkages to be conducted at a national or cross-jurisdictional level further impacts the size of datasets being compared. However, as the size of datasets increases linearly, the record linkage comparison space increases exponentially.

Record linkage is computationally expensive, with a potential comparison space equivalent to the Cartesian product of the record sets being linked, making linkage of large datasets (in the tens of millions of records or more) a considerable challenge. Optimising systems, removing manual operations and increasing the level of autonomy for such processes is essential.

Many industries have moved towards cloud computing as a solution for high computational workloads, data storage and analytics [243]. There are a number of business benefits to cloud computing which include usage-based costing, minimal upfront infrastructure investment, superior collaboration (both internally and externally), better management of data and increased business agility [70, 277]. Despite these advantages, the uptake of cloud infrastructure by the record linkage community has been slow. One reason for this is that the storage of identifiable information on cloud infrastructure is seen or assessed as high risk by data custodians. Even though security in cloud computing systems has been shown to be more robust than some in-house systems [119], the media reporting of data breaches has created an impression of insecurity and vulnerability [142, 161, 264]. This impression, coupled with a culture of risk aversion by data custodians [224], has left some record linkage units with expensive, dedicated equipment and computing resources that require managing, maintaining and upgrading or replacing regularly.

The rapid uptake of cloud computing by government has enabled some linkage units to modernise their infrastructure with a move to the cloud. The Victorian government's Department of Health and Human Services (DHHS) rebuilt much of its IT infrastructure in Microsoft Azure in 2017 [218], and is currently rebuilding its entire linkage capability to better suit the cloud

services that are available. This is an all-in approach by the department for its IT infrastructure, where other state governments are opting for digital strategies that have a preference for cloud-based solutions for all new projects [202, 78, 229].

To leverage the advantages of cloud computing, the record linkage community needs to explore operational cloud computing models for record linkage that address the specific concerns of all stakeholders. In addition, linkage infrastructure requires the development and implementation of robust security and information governance frameworks as part of a holistic cloud solution.

1.7.1 'Big data' in research

Unabating growth in the creation of data, coupled with advances in information technology and Internet connectivity, provides tremendous potential for data-driven breakthroughs in the understanding, treatment, and prevention of disease. These health research innovations are being complemented by data from non-traditional sources (i.e., from sources other than administrative health and survey records). Opportunities include the use of mobile phone records [90] and Google search histories [1, 106] for disease surveillance, patient data collected from wearable devices [216, 230], and manual journaling through mobile phone applications [165, 176]. Data from the private health sector and government administrative datasets that lie outside the health sector [266] are also of interest, as is spatial information that has direct application for understanding exposures and inequalities [288].

Genetic information unavailable a generation ago is already used in clinical decision making [5], and its importance is only likely to increase. The key to unlocking these data is in relating details at an individual patient level to provide an understanding of risk factors and appropriate interventions [162].

The Public Health Research Network (PHRN) in Australia and the Farr Institute in the UK are two examples of collaborations that have begun to invest in large-scale data linkage infrastructure to achieve national linkage objectives [95, 221]. The establishment of research centres specialising in the analysis of "big data" (e.g. the Centre for Big Data in Health at the University of New South Wales [201] also recognise the issue of increasing data size and complexity.

Through record linkage, it has been possible to construct and analyse population-wide datasets comprising "linked" administrative records pertaining to each individual. Health-based record linkage frameworks have been established, which routinely integrate data from hospital admissions, emergency departments, primary care facilities, birth, death, and disease registries, creating a rich analytic resource to support evidence-based decision making [49, 129, 185].

1.7.2 Commercial cloud

Businesses worldwide are rapidly adopting advances and innovation in scalable/elastic distributed systems provided by cloud computing. The Australian Government is also actively

promoting cloud computing for government, non-profits and research groups, requiring government agencies to consider cloud services for new ICT procurements [8].

The Australian Signals Directorate (ASD) established a Certified Cloud Services List (CCSL) in 2014 under a cloud services certification program (CSCP) and certified a number of cloud providers for data classified up to protected level. This included two of the largest commercial cloud providers in Amazon Web Services (AWS) and Microsoft Azure, conditional on agencies configuring the environment in line with the guidance in the ACSC Certification Report and Consumer Guide. The CSCP has since ceased operation, releasing guidance for cloud security in conjunction with industry [7], as well as separate requirements for providers that handle government data [11]. There is some debate as to what classification should be applied to raw personal identifiers, or privacy-preserved versions of personal identifiers. It is clear that governments and industry are working together to provide guidance for establishing scalable environments for secure workloads.

The last decade has seen the rise of cloud computing as a consequence of increased Internet bandwidth, an explosive growth of data and the convergence of two major trends in information technology: IT efficiency and business agility [188, 285]. The emergence and uptake of an Infrastructure as a Service (IaaS) service model (as defined by the US National Institute of Standards and Technology) by government agencies provide an immediate opportunity for data linkage units to provision the processing, storage and other computing resources as needed [195]. However, to fully utilise the capability of these cloud environments, development of native cloud solutions for data linkage is required, with linkage systems harnessing the rapid elasticity of cloud services for on-demand resource usage.

1.7.3 Distributed linkage algorithms

The record pair comparison and classification tasks are the most compute-intensive tasks in the data linkage process, though they are heavily affected by the indexing method used [217]. The single process limitation of most linkage applications makes it difficult to cater for increasingly large datasets, regardless of indexing. Increasing memory and CPU resources for these single-process applications provides some ability to increase capacity, but this may not be sustainable in the longer term.

Research on algorithms that address the computational burden of the comparison and classification tasks has been undertaken. Most work in this area has been on distributed and parallel algorithms for record linkage that are specific to the MapReduce paradigm [58, 92]. MapReduce was seen as a natural fit for parallel entity resolution, the map phase used for blocking and the reduce phase used for matching [168]. One of the issues with this approach is the generation of different block sizes causing an uneven distribution of workload across processing nodes [298]. Some nodes can be overloaded with processing while others sit relatively idle. This data skew problem is the focus of much of the work in this area, using different blocking methods, data partitioning and load balancing [58]. One method uses multiple MapReduce jobs; the first job identifies record pairs that should be compared, and the second job distributes these pairs

for comparison [151]. Another method that employs two MapReduce jobs uses the first job to analyse and estimate the processing cost of each block, while the second distributes these blocks evenly using this cost estimate [194].

Few sources detail the comparison and classification tasks themselves, and apart from a focus on load balancing algorithms to address data skew, most focus on different blocking methods to improve indexing efficiency. Blocking methods include standard blocking [68, 102, 167], density-based blocking [80], sorted neighbourhood [168] and LSH (locality sensitive hashing) [148, 147]. These blocking methods have had varying success in optimising workload distribution. However, little effort has been devoted to the accuracy of these methods within the MapReduce paradigm.

More recently, Apache Spark has gained traction over MapReduce as the processing engine of choice for big data analysis [262]. As a result, attention for distributing processing for entity resolution has slowly shifted to Spark. Pita et al. [220] have demonstrated good performance and linkage accuracy using a Spark-based workflow for probabilistic linkage. In this case, Spark was chosen for its in-memory processing, ease of programming and the new resilient distributed dataset (RDD) model. Like MapReduce, Spark continues to be used to address the issues with data skew on larger datasets. Full entity resolution solutions using Spark are being developed, with different indexing techniques used to address workload distribution [58]. The SparkER tool [100] uses LSH, meta-blocking and a block purging process to remove high-frequency blocking keys. Mestre et al. [196] present a sorted neighbourhood implementation with an adaptive window size, utilising three Spark transformation steps to minimise data skew during data distribution.

There have been some efforts to address linkage of larger datasets through parallel processing techniques outside of the Hadoop ecosystem. One example uses the specialised processing power of GPUs to parallelise record matching [260]. This modified version of PPJoin, called P4Join, claims an execution time improvement of up to twenty times. However, despite its potential for significant improvements on runtime performance, there has not been any further work published on P4Join using larger datasets or on clusters of GPU nodes. More recently, Boratto et al. evaluated a hybrid algorithm using both GPUs and CPUs with much larger datasets [31]. Though restricted to single (highly specified) machines, these evaluations show promise provided the approach can be applied within a compute cluster. No further results have been published.

1.8 Current PPRL systems/solutions

While privacy-preserving record linkage techniques are still relatively new, there are some software systems that claim to provide a range of privacy-preserving capability. Implementations may vary, although the majority of these systems appear to adopt the Bloom filter approach. A summary of current PPRL systems is provided.

1.8.1 Grhanite

Grhanite is software developed by the Health and Biomedical Informatics Centre at the University of Melbourne [44]. The software's primary application is to extract data from GP practices to store in a central repository maintained by the Health and Biomedical Informatics Centre. Individuals are identified between GP practices through the use of a hash-based PPRL protocol.

The protocol used by Grhanite involves concatenating elements from different identifiers together before they are hashed; however, the exact elements used are not known. To tolerate error and difference in identifiers, some level of pre-processing occurs, which includes the use of techniques such as phonetic encoding and nickname lookups. Grhanite states that it uses 'a variety of proprietary techniques to improve the sensitivity and specificity of record linkage far beyond that traditionally found in hashed deterministic linkage'. However, these techniques have not been independently evaluated.

The linkage algorithm is a fixed feature of the Grhanite system; it cannot take advantage of additional identifiers should they be available in certain datasets. It cannot be modified to take into account the nature of the data, for instance, if a dataset is particularly dirty, or missing certain identifiers, or has large numbers of missing values. As such, it is likely the linkage algorithm will perform poorly in these scenarios. The lack of string similarity measures, and the fact that field weights cannot be adjusted also suggest that linkage quality may be lower.

The protocol used by Grhanite is proprietary, and although high-level information is available in peer-reviewed publications, the details of the protocol are not available. No independent, published evaluations of the protocol can be found in the literature.

1.8.2 LinXmart

LinXmart is software developed by the Centre for Data Linkage at Curtin University [39] and has been purpose built for operational data linkage. The software is an 'end to end' system for record linkage and linkage key management. It performs PPRL as well as unencrypted linkage. The privacy-preserving linkage component utilises field-level and composite Bloom filters (the fields chosen are configurable). LinXmart also provides an application to transform raw data into a privacy-preserved state, which is typically run by the data custodians.

The default LinXmart Bloom implementation involves separately encoding each field into Bloom filters. Field-based weights are then computed and used for standard probabilistic linkages. It is also possible to combine groups of fields (or indeed the entire row) into composite Bloom filters that are then matched using deterministic passes. However, the separate field-based approach means that the protocol can operate on data containing large amounts of missing values.

The field-based approach implemented in LinXmart has been evaluated on large real-world Australian datasets. The first evaluation included a deduplication linkage of approximately 20 million records from the Admitted Patient Data Collection [238]. The evaluation found

no significant difference in linkage quality between a probabilistic record linkage carried out using personal identifiers, and a privacy-preserving record linkage approach using the same methodology.

1.8.3 SOEMPI

The Secure Open Enterprise Master Patient Index (SOEMPI) is a Java framework for PPRL developed and maintained by the Health Information Privacy Laboratory at Vanderbilt University [270]. It utilises a Bloom filter approach to PPRL, where all fields are combined into a single record-level Bloom filter. The software is open-source and available online [276].

In the record-level Bloom filter approach used by SOEMPI, each field is converted into a Bloom filter; these Bloom filters are then concatenated together to form a single Bloom filter for an individual [86]. The record-level Bloom filter approach allows individual fields to receive more or less of the 'space' in the Bloom filter, allowing individual fields to be weighted as more or less important.

The use of a record level Bloom filter approach means that identifiers that don't exist in all datasets being linked, could not be used in linkage. However the users do have the flexibility to utilise any identifiers that are available on all datasets being linked (unlike software which uses a pre-determined set of fields). Missing data in record level Bloom filters can result in lower linkage quality as compared to traditional unencrypted linkage or field level Bloom filters.

Data custodians must install the entire SOEMPI module in order to encrypt the data before sending it. This is a non-trivial installation, as it either requires using a downloaded virtual machine or installation of Java and a database package along with the actual application. Once installed, custodians must add all necessary parameters for encrypting the Bloom filters themselves (in practice these would be supplied by the linkage unit).

Encrypted data can be sent through the SOEMPI module to the linkage unit. In this software, the match configuration is determined and input by the data custodians rather than the linkage unit. Again, in practice, this information would need to be supplied to the custodian by the linkage unit.

Walkthrough tutorials for encrypting, linking and sending data are available online. The software does not appear to be currently maintained, with the last release occurring in 2014.

1.8.4 LSHDB

The LSHDB is an open-source software package developed by researchers at the Hellenic Open University, Greece for conducting record linkage, including privacy-preserving record linkage [147]. It utilises the locality sensitive hashing (LSH) technique for record comparison.

To use LSHDB for PPRL, data are first encoded utilising a Bloom filter approach; either field or record level Bloom filters. LSHDB has the advantages and disadvantages of whichever of these methods is chosen.

Comparison of records takes place using the locality sensitive hashing (LSH) algorithm. This algorithm essentially clusters the encoded records into groups of records thought to belong to the same person. Currently there is limited empirical information on how effective the LSH algorithm is in ensuring high linkage quality.

The LSH algorithm is designed to reduce the number of record-pair comparisons which need to be performed. As such, it is likely to be relatively efficient; however, few empirical evaluations of this method have been undertaken.

Data must first be encrypted prior by custodians prior to sending it to the linkage unit. However, the LSHDB contains no functionality for encoding of records into Bloom filters, either as a separate program provided to data custodians or within the main program itself. As such, a separate piece of software would need to be developed to carry out this encryption.

1.8.5 MERLIN

MERLIN is a multi-party privacy-preserving record linkage demonstration system, developed by the Australian National University in Canberra [232]. It currently exists as an online demonstration system only.

MERLIN implements a variety of methods for 'multi-party' privacy-preserving record linkage, whereby any number of data custodians can link together data without the need of an independent third party such as a linkage unit.

MERLIN uses a record-level Bloom filter approach to encode each record, similar to the approach used by SOEMPI. This method, therefore, has the same strengths and weaknesses as SOEMPI. Overall linkage quality achievable using this method is likely to be reasonably high, although not as high as is likely to be achieved with field-level Bloom filters. Missing values will increase the number of missed matches, so the quality and completeness of the data has a significant impact on the accuracy of the linkage.

The MERLIN system does not make use of an independent third party. As such, the entire linkage process would be carried out by data custodians, who are responsible for running the system and ensuring high linkage quality.

1.8.6 LinkIT

LinkIT (Linkage and Integration via data Transformations) is an open-source toolkit for privacy-preserving record linkage that is designed to work with FRIL (Fine-grained Record Integration and Linkage) to present a complete record linkage solution [30]. Last released in 2011, the LinkIT toolkit is not currently available online.

The LinkIT module is run on source datasets by the data custodians. It uses a novel embedding scheme, based on frequent variable-length n-grams, to securely encode specific fields into a privacy-preserved state. A third party utilises FRIL to perform probabilistic matching on the encoded data.

1.8.7 PPRL (R Package)

An R package published to the Comprehensive R Archive Network (CRAN) as 'PPRL', provides a set of functions for privacy-preserving fields and matching them using either a deterministic or probabilistic approach [258]. Privacy-preserving functions include encrypted SLK-581, several variations of single and composite Bloom filters, and other encryption methods.

Use of the package requires some knowledge of R, so is somewhat limited to users who are familiar with the R language and environment. While popular with statisticians, the R language includes some inherent limitations around memory management and multi-threading capabilities. Datasets must be loaded fully into memory before they can be analysed and only a single thread of computation can be run at a time. There are also some limitations to the linkage functions that might prevent optimal results. For example, the probabilistic linkage function does not allow field weights to be specified, relying on an internal estimation of weights prior to matching. However, for users who are already using R for the analysis of data, this package provides some extremely useful functionality for matching records prior to analysis.

1.9 Published manuscript(s)

1.9.1 Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage

Brown AP, Ferrante, AM, Randall, SM, Boyd, JH, Semmens, JB (2017). *Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage.* Frontiers in Public Health, 5 (March), 1–6.

frontiers in Public Health

PERSPECTIVE
published: 02 March 2017
doi: 10.3389/fpubh.2017.00034

Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage

Adrian P. Brown, Anna M. Ferrante, Sean M. Randall, James H. Boyd and James B. Semmens*

Centre for Population Health Research, Curtin University, Bentley, WA, Australia

In an era where the volume of structured and unstructured digital data has exploded, there has been an enormous growth in the creation of data about individuals that can be used for understanding and treating disease. Joining these records together at an individual level provides a complete picture of a patient's interaction with health services and allows better assessment of patient outcomes and effectiveness of treatment and services. Record linkage techniques provide an efficient and cost-effective method to bring individual records together as patient profiles. These linkage procedures bring their own challenges, especially relating to the protection of privacy. The development and implementation of record linkage systems that do not require the release of personal information can reduce the risks associated with record linkage and overcome legal barriers to data sharing. Current conceptual and experimental privacy-preserving record linkage (PPRL) models show promise in addressing data integration challenges. Enhancing and operationalizing PPRL protocols can help address the dilemma faced by some custodians between using data to improve quality of life and dealing with the ethical, legal, and administrative issues associated with protecting an individual's privacy. These methods can reduce the risk to privacy, as they do not require personally identifying information to be shared. PPRL methods can improve the delivery of record linkage services to the health and broader research community.

Keywords: record linkage, data integration, privacy, encryption, data quality, linkage quality

OPEN ACCESS

Edited by:
Paul Michael Kelly, ACT Health, Australia

Reviewed by:
Arnold Bosman, Transmissible, Netherlands
Ronan Foley, Maynooth University, Ireland

****Correspondence:***
Anna M. Ferrante
a.ferrante@curtin.edu.au

Specialty section:
This article was submitted to Public Health Policy, a section of the journal Frontiers in Public Health

Received: *30 November 2016*
Accepted: *15 February 2017*
Published: *02 March 2017*

Citation:
Brown AP, Ferrante AM, Randall SM, Boyd JH and Semmens JB (2017) Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage. Front. Public Health 5:34. doi: 10.3389/fpubh.2017.00034

INTRODUCTION

Unabating growth in the creation of data, coupled with advances in information technology and Internet connectivity, provides tremendous potential for data-driven breakthroughs in the understanding, treatment, and prevention of disease. These health research innovations are being complemented by data from non-traditional sources (i.e., from sources other than administrative health and survey records). Opportunities include the use of mobile phone records (1) and Google search histories (2) for disease surveillance, patient collected data from wearable devices (3), and manual journaling through mobile phone applications (4). Data from the private health sector and government administrative datasets that lie outside the health sector (5) are also of interest, as is spatial information that has direct application for understanding exposures and inequalities (6).

Genetic information unavailable a generation ago is already used in clinical decision making (7), and its importance is only likely to increase. The key to unlocking these data is in relating details at an individual patient level to provide an understanding of risk factors and appropriate interventions (8).

A key methodology that has supported health research is record linkage, a process of accurately bringing together records from multiple datasets that belong to the same person. Through record linkage, it has been possible to construct and analyze population-wide datasets comprising "linked" administrative records pertaining to each individual. Health-based record linkage frameworks have been established, which routinely integrate data from hospital admissions, emergency departments, primary care facilities, birth, death, and disease registries (1, 2), creating a rich analytic resource to support evidence-based decision making (9–11).

Present models of record linkage use trusted third parties (TTPs) or data linkage units (DLUs) to accurately match records using personal identifiers (12). Incorporating information from new and diverse data sources into these linkage frameworks are likely to have significant benefits to research; however, the operational and administrative overheads are substantial. Technical issues (i.e., scalability, efficiency) and effects on linkage quality (accuracy) will also be impacted and need to be assessed.

Sharing of public and private datasets also presents privacy and confidentiality challenges. Protecting the privacy of individuals is paramount in the record linkage process and essential to maintain community support and trust. There are serious ethical implications in combining information on individuals (generally without direct consent) from government and other sources; essentially a form of surveillance of an entire population. For some privacy advocates, this is a bridge too far, conjuring up images of an Orwellian dystopia or the excesses of totalitarian regimes (13, 14). Health researchers argue that privacy risks can be minimized and that the public benefit of utilizing these rich datasets outweighs the risk to privacy; that is, there is an ethical imperative to conduct record linkage for research (15). The public's view on this issue is not always clear; numerous surveys have been conducted in Australia, which sometimes return contradictory results regarding Australian views on the use of personal health information [see Ref. (16). for a review]. Similar contradictions have been observed in results from Canadian surveys (14).

While a number of existing processes and techniques are used to maintain patient privacy during record linkage (17), the development of new and improved linkage methods may provide an opportunity for alternative approaches that further reduce privacy risks without compromising on linkage quality.

This article discusses the emergence and potential benefit of record linkage techniques that limit the release of personal identifiers for linkage. These methods, collectively referred to as privacy-preserving record linkage (PPRL), operate in such a way that they do *not* require the release of personally identifying information by data custodians. PPRL methods work on information that has been permanently encoded, encrypted, or transformed before releasing the data for linkage. Through PPRL methods, the benefits of linkage can be realized without the risks associated with disclosure of personal information.

EXISTING RECORD LINKAGE FRAMEWORKS

There is a long history in Australia of record linkage supporting both jurisdictional level and national research and health decision making (10, 12, 18). Record linkage capabilities in all jurisdictions (19–21) have recently been strengthened, and in many cases expanded, through strategic national investment: through the National Collaborative Research Infrastructure Strategy in Australia; the Canadian Institutes of Health Research in Canada; and through the Farr Institute initiative in the United Kingdom (22).

The record linkage framework adopted by most of these jurisdictions is a TTP model, whereby dedicated linkage units undertake record linkage to service and support research. Administrative data collections (such as hospital discharges, emergency presentations, mortality, and cancer registers) have typically formed the backbone of enduring record linkage systems (18, 23). Such collections are highly confidential, containing sensitive personal information that is protected by law.

RECORD LINKAGE AND PRIVACY

Linkage of person-level records through the use of personally identifying information, and generally without consent, has significant ethical and legal implications that have been at the forefront of issues confronted and addressed by DLUs (12, 24).

The extent to which data can be used in record linkage depends on the applicable legislation in each jurisdiction. Some administrative collections are bound by specific laws which either prohibit or severely curtail the release of personal information from these systems.[1] It has been claimed that more than 500 secrecy and privacy provisions exist in Australian Commonwealth laws, imposing considerable limits on the availability and use of identifiable data (25). At Commonwealth level, privacy laws permit some level of disclosure of personal information by authorities for human research (*Commonwealth Privacy Act 1988 s 95*). The release of personal data for linkage can be authorized if public benefit outweighs the privacy of individuals (26).

Working within these legal frameworks, data custodians, DLUs, and the research community in Australia have developed secure data access and usage models that provide important safeguards to privacy. DLUs have also implemented best practice data governance policies and practices to minimize further the privacy risks posed by their operations (12, 18, 19, 27–29).

This includes utilizing the "separation principle" (30), a simple method for restricting the type of data received by each organization in the linkage process. Under this principle, the DLU receives only the personally identifying information required for linkage, but not the content data. The researcher, on the other

[1] In Western Australia, for example, both the WA Children's Court Act 1988 and the Young Offenders Act 1994 curtail the release of information for research in relation to juvenile offenders. In South Australia, state-based regulations restrict the release of information from the SA Perinatal Statistics Collection (SA Health Care Variation Regulations 2010, Reg 4). Similar legal barriers exist in other jurisdictions, both locally and internationally.

hand, receives only the content but not personal identifying information. Only the data custodian has access to both personal identifying information and clinical content data.

The use of the separation principle greatly enhances privacy. However, in many instances, the risk to privacy can be still large. For instance, knowledge that a particular individual has a record within a data collection is itself revealing, especially for specific data collections such as mental health inpatient datasets or cancer registries. This information will be still provided to the linkage unit under the separation principle.

The release of personally identifying information always carries some additional risk, as more individuals have access to this information. While rare, attempting to determine whether a person of interest is contained within a dataset does occur; for instance, US intelligence agents have used their surveillance capabilities to spy on romantic interests (31), as have Australian telecommunications workers (32).

Some custodians remain averse to the release of personal information for reasons that extend beyond privacy risks, such as discrimination, reputational damage and/or embarrassment, criminal misuse of the data, and commercial harm (25).

Legislative barriers and risk aversion by data custodians are currently being challenged by open data policies and a growing need by and for government to work with private industry to more effectively service community needs. A recent Productivity Commission Inquiry into the benefits and costs of increasing the availability and use of public and private sector data recognizes the barriers and risks associated with working with named data (25). The Inquiry outlines a framework for data sharing underpinned by legislative change, governance structures (to remove blocks and increase data access), and the development of "systems and processes […] to identify, assess, manage and mitigate risks related not just to data release and sharing, but also data collection and storage" [(25), p.9].

The issues being encountered in Australia are shared internationally. DLUs in the United States, Canada, and Europe face similar legal and risk-related hurdles (e.g., the United States: *Health Insurance Portability and Accountability Act 1996*, Canada: *Personal Health Information Protection Act 2004*, and Europe: *Data Protection Directive 95/46/EC*). German laws in relation to the disclosure of personal information are particularly restrictive (*Bundesdatenschutzgesetz—Federal Data Protection Act of Germany*) and, in some cases, only a single data item can be used for anonymous linkage (33).

PRIVACY-PRESERVING SOLUTIONS

Privacy-preserving record linkage protocols utilize algorithms and techniques to conduct linkage on encrypted or masked information; these methods do not require data custodians to release personal identifiers to third parties. This reduces the risks associated with the release of personal data. Three important attributes characterize all PPRL protocols: accuracy, efficiency, and privacy.

Different classes of privacy-preserving linkage methods provide differing levels of privacy protection. These range from techniques such as the statistical linkage key that simply amalgamates parts of a person's identifiers into a single variable (34) to methods that encrypt or encode the data so that those with access cannot learn any information directly from the encrypted values. The exact level of privacy required will always depend on context, but all things being equal, a protocol with higher privacy is preferred.

An important difference in PPRL protocols is the method of matching which impacts on linkage quality (accuracy). Protocols may perform matching on a particular set of identifiers, using either exact or similarity comparisons. Similarity matching enables records with slight differences to come together, which is vital for obtaining high-quality linkage results (accuracy). For this reason, PPRL protocols that utilize approximate matching are favored.

Efficiency can be often a concern for record linkage and will continue to present challenges to DLUs as the volume of data continues to grow. Although there are no established performance standards, record linkage is computationally slow, and for any PPRL protocol to be practical, it must complete within a reasonable time frame.

The extent to which these protocols are used in practice varies. To date, most PPRL implementations use exact matching on particular attributes of a dataset (35), which are typically irreversibly encoded to ensure privacy (36). Though efficient, these methods have reduced linkage quality and, therefore, are operationally unsuitable in DLUs.

Of all PPRL methods, the Bloom filter method appears to be the most promising for operational use (37). An advantage of the Bloom method over other PPRL methods is that it utilizes approximate matching while providing similar or superior privacy protection. The method has been evaluated on large-scale, real world health datasets, with results returning equal linkage quality and similar efficiency to traditional linkage methods (which use personal identifiers in the matching process) (38). No record linkage method, privacy preserving or not, achieves perfect accuracy—to be able to achieve equal accuracy to the standard non-privacy-preserving approach is a considerable accomplishment. The security of the protocol has been rigorously investigated (39–41). Cryptographic attacks on the algorithm found ways to reveal some identifiers (40). However, modifications to the protocol have rendered these attacks fruitless (42); there are currently no known security vulnerabilities with the protocol.

The introduction of the Bloom filter method brings new challenges (17). As well as operational requirements around designing optimal linkage strategies, new ways of validating record linkage results need to be developed. In traditional record linkage, linkage results are validated through clerical inspection (or "manual review") of personal identifiers; however, in a privacy-preserved context where all data are encoded, there is no way to manually review the data or correct possible data or linkage errors. New methods for validating linkage results under privacy-preserved linkage model are emerging, however (43).

PPRL: AN EXAMPLE

Consider the (hypothetical) scenario: to attempt to reduce the rate of youth suicide, the government of the day has invested in a comprehensive mental health care package for those who have

attempted suicide. The government wishes to see whether their program has worked in reducing the rate of suicide and attempted suicide.

To answer this question, two datasets will be required: a hospital admissions dataset and a mortality register. From the hospital admissions dataset, records will be required to be sent to the linkage unit for all those persons who have attempted suicide before and after the start of the health intervention; all records from the mortality register will be required by the linkage unit. The linkage unit will receive only the personal identifying information required for linkage (i.e., name, date of birth, gender, address). The linkage unit identifies which records from the supplied hospital dataset have associated mortality records. The linkage unit passes this information back to the data custodians, who then provide the content data (i.e., not personally identifying information) to the researcher for the hospital records, and any linked mortality records, along with a key that identifies which records belong to which individual. The researcher can then use this information to determine whether the intervention reduced suicide and attempted suicide rates.

The privacy risk in the aforementioned scenario is the delivery to the linkage unit of personal identifying information from hospital records of those who have attempted suicide. This extremely sensitive information has been made available to a third party. The use of privacy-preserving linkage methods would remove this risk; instead, the linkage unit would receive encrypted personal identifiers; they would have no means of identifying any of these individuals, but would still have the ability to determine which records belong to the same individual between datasets.

GROWING INTERNATIONAL INTEREST IN PPRL

With a growing demand for linked data from government and the university sector, interest in PPRL, particularly the Bloom filter method, is flourishing. Interest stems from two principal sources: at a technical level, by computer scientists and cryptographers with interests in information and data security, and at an operational level, by groups with interest in and responsibility for delivering record linkage services.

Several groups are actively developing and refining PPRL methods at the scientific level including the German Record Linkage Center (University of Duisburg-Essen) (44, 45), the Research School of Computer Science (Australian National University) (46–48), and the Health Information Privacy Laboratory (Vanderbilt University) (39, 49). Researchers from these groups and others recently participated in a 2016 Data Linkage and Anonymisation programme at the Isaac Newton Institute for Mathematical Sciences (Cambridge University, supported by EPSRC grant no EP/K032208/1)[2]; this 6-month international programme included seminars and workshops on linkage and privacy protection to share and advance knowledge in the mathematical sciences and related disciplines. A key goal of the forum was to "enhance opportunities for the analysis of data, especially obtained through linkage, whilst protecting privacy and taking account of related practical constraints."

At an operational level, PPRL featured prominently in the 2016 International Population Data Linkage Network Conference (Swansea University), with several presentations on the topic including a keynote session that described a collaboration between international research institutions in Canada, Australia, and Wales (44, 46, 50–53).

OPPORTUNITY AND CHANGE MANAGEMENT

In addition to reducing the privacy risks associated with record linkage, the advent of PPRL protocols potentially heralds a new era of population-focused research using linked data, bridging gaps, and opening up opportunities for new and different forms of linkage-based research. PPRL methods may provide an avenue to access previously "hard to get" datasets (i.e., those with significant legal or regulatory constraints). PPRL methods may also provide a mechanism for accessing and integrating data from new and emerging sources. As well as data from new technologies (e.g., wearable devices, smartphone apps), these new sources may include the private health sector that has, to date, had limited exposure to, and engagement with, data linkage frameworks (54, 55).

New methods may require new or adjusted models of operation. Some custodians have expressed a desire to have flexibility in record linkage models to accommodate the features of different data collections (50). However, different or altered data linkage operating models can have significant implications for end-user timeframes, operational efficiency, and linkage quality (50), and these need to be carefully managed and monitored. It is important that the strengths and limitations of the PPRL methods are understood. This will require conversations with stakeholders (i.e., data custodians, linkage units, researchers, and the community) around the risk–benefit of these new models and the expected realization of public benefit.

CONCLUSION

The implementation of PPRL methods that do not require the release of personal information but protect privacy through other mechanisms (e.g., encryption methods) represents a breakthrough in record linkage, substantially reducing privacy risks without negatively impacting on linkage quality. By utilizing methods that do not require the release of personally identifying information, concerns regarding personal surveillance and government overreach can be allayed. Supplementing traditional linkage methods with PPRL methods will increase the number and type of datasets that can be included in record linkage studies.

The advent of PPRL methods to protect patient privacy expands the toolkit of techniques that are available to DLUs. Used in conjunction with traditional linkage methods, PPRL widens the net of record linkage without compromising privacy or linkage quality. These methods will hopefully allow more diverse, patient-centered data sources to be utilized for health research,

[2]https://www.newton.ac.uk/event/dla.

bringing enormous opportunities to increase our understanding of disease and to tailor interventions and treatment to each individual.

AUTHOR CONTRIBUTIONS

AB and AF accept immediate responsibility for the manuscript. AF, AB, SR, JB, and JS each contributed to the conception and design of the paper. AF and AB drafted the first version of the article, with SR, JB, and JS providing important additional input and intellectual content. All authors were involved in revising the manuscript and approving its final form.

FUNDING

This work was discussed at the Isaac Newton Institute for Mathematical Sciences, Cambridge, supported by EPSRC grant no EP/K032208/1.

REFERENCES

1. Ebola and big data – call for help. *The Economist*. London: The Economist Group (2014).
2. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature* (2009) 457(7232):1012–4. doi:10.1038/nature07634
3. Pantelopoulos A, Bourbakis NG. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Trans Syst Man Cybern C Appl Rev* (2010) 40(1):1–12. doi:10.1109/TSMCC.2009.2032660
4. Klasnja P, Pratt W. Healthcare in the pocket: mapping the space of mobile-phone health interventions. *J Biomed Inform* (2012) 45(1):184–98. doi:10.1016/j.jbi.2011.08.017
5. Stanley PF. *Developmental Pathways in WA Children Project*. Perth, WA: Chief Investigator and Director, Telethon Institute for Child Health Research (2006–2007).
6. Waller LA, Gotway CA. *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: John Wiley and Sons (2004).
7. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. *Nature* (2015) 526(7573):336–42. doi:10.1038/nature15816
8. Khoury MJ, Iademarco MF, Riley WT. Precision public health for the era of precision medicine. *Am J Prev Med* (2015) 50(3):398–401. doi:10.1016/j.amepre.2015.08.031
9. Brook EL, Rosman DL, Holman CDAJ. Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System. *Aust N Z J Public Health* (2008) 32(1):19–23. doi:10.1111/j.1753-6405.2008.00160.x
10. Holman CDAJ, Bass AJ, Rosman DL, Smith MB, Semmens JB, Glasson EJ, et al. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Aust Health Rev* (2008) 32(4):766–77. doi:10.1071/AH080766
11. Lyons RA, Ford DV, Moore L, Rodgers SE. Use of data linkage to measure the population health effect of non-health-care interventions. *Lancet* (2014) 383(9927):1517–9. doi:10.1016/S0140-6736(13)61750-X
12. Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC Health Serv Res* (2012) 12(1):480. doi:10.1186/1472-6963-12-480
13. Green E, Ritchie F, Mytton J, Webber DJ, Deave T, Montgomery A, et al. *Enabling Data Linkage to Maximise the Value of Public Health Research Data*. London: Wellcome Trust (2015).
14. Upshur RE, Morin B, Goel V. The privacy paradox: laying Orwell's ghost to rest. *Can Med Assoc J* (2001) 165(3):307–9.
15. Hetzel D. Data linkage research – can we reap benefits for society without compromising public confidence? *Aust Health Consum* (2005) 2:27–8.
16. Holman CDAJ. *Anonymity and Research: Health Data and Biospecimen Law in Australia*. Perth: Uniprint, UWA (2012).
17. Boyd JH, Randall SM, Ferrante AM. *Application of Privacy-Preserving Techniques in Operational Record Linkage Centres. Medical Data Privacy Handbook*. Berlin: Springer (2015). p. 267–87.
18. Lawrence G, Dinh I, Taylor L. The centre for health record linkage: a new resource for health services research and evaluation. *Health Inf Manag* (2008) 37(2):60–2. doi:10.1177/183335830803700208
19. Ford DV, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, et al. The SAIL databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* (2009) 9(1):157. doi:10.1186/1472-6963-9-157
20. Martens PJ. Using the repository housed at the Manitoba centre for health policy: learning from the past, planning for the future. In: *Conference Proceedings of the Statistics Canada Conference: Longitudinal Social and Health Surveys in an International Perspective*. Montreal, Quebec (2006).
21. Gill LE. *OX-LINK: The Oxford Medical Record Linkage System. Record Linkage Techniques*. Oxford: University of Oxford (1997). 19 p.
22. Farr Institute. (2016). Available from: http://www.farrinstitute.org/
23. Rosman D, Garfield C, Fuller S, Stoney A, Owen T, Gawthorne G. Measuring data and link quality in a dynamic multi-set linkage system. In: *Symposium on Health Data Linkage*. Sydney (2002). Available from: http://www.phidu.torrens.edu.au/pdf/1999-2004/symposium-proceedings-2003/rosman_a.pdf
24. Hobbs M, McCall M. Health statistics and record linkage in Australia. *J Chronic Dis* (1970) 23(5):375–81. doi:10.1016/0021-9681(70)90020-2
25. Productivity Commission. *Data Availability and Use, Draft Report*. Canberra: Australian Government (2016).
26. Allen J, Holman CDJ, Meslin E, Stanley F. Privacy protectionism and health information: is there any redress for harms to health? *J Law Med* (2013) 21(2):473–85.
27. Harris J. Next generation linkage management system. In: *Sixth Australiasian Workshop on Health Informations and Knowledge Management*. Adelaide: Australian Computer Society (2013).
28. Trutwein B, Holman D, Rosman D. Health data linkage conserves privacy in a research-rich environment. *Ann Epidemiol* (2006) 16(4):279–80. doi:10.1016/j.annepidem.2005.05.003
29. Roos LL, Brownell M, Lix L, Roos NP, Walld R, MacWilliam L. From health research to social research: privacy, methods, approaches. *Soc Sci Med* (2008) 66(1):117–29. doi:10.1016/j.socscimed.2007.08.017
30. Kelman CW, Bass AJ, Holman CDJ. Research use of linked health data – a best practice protocol. *Aust N Z J Public Health* (2002) 26(3):251–5. doi:10.1111/j.1467-842X.2002.tb00682.x
31. NSA officers spy on love interests. *Wall Street J* (2013). Available from: http://blogs.wsj.com/washwire/2013/08/23/nsa-officers-sometimes-spy-on-love-interests/
32. Vodafone sacks staff over alleged security breach. *IT News* (2011). Available from: http://www.itnews.com.au/News/244672,vodafone-sacks-staff-over-alleged-security-breach.aspx
33. Bundestag D. Gesetz ueber Krebsregister (Krebsregistergesetz KRG). *Bundesgesetzblatt* (1994) 79:994.
34. Karmel R. *Data Linkage Protocols Using a Statistical Linkage Key*. Canberra: Australian Institute of Health and Welfare (2005).
35. Karmel R, Anderson P, Gibson D, Peut A, Duckett S, Wells Y. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. *BMC Health Serv Res* (2010) 10(41):1–13.
36. Quantin C, Bouzelat H, Allaert F, Benhamiche A-M, Faivre J, Dusserre L. How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure. *Int J Med Inform* (1998) 49(1):117–22. doi:10.1016/S1386-5056(98)00019-7
37. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. *BMC Med Inform Decis Mak* (2009) 9:41. doi:10.1186/1472-6947-9-41
38. Randall SM, Ferrante AM, Boyd JH, Semmens JB. Privacy-preserving record linkage on large real world datasets. *J Biomed Inform* (2014) 50:205–12. doi:10.1016/j.jbi.2013.12.003
39. A constraint satisfaction cryptanalysis of Bloom filters in private record linkage. In: Kuzu M, Kantarcioglu M, Durham E, Malin B, editors. *Privacy Enhancing Technologies*. Springer (2011).

40. Niedermeyer F, Steinmetzer S, Kroll M, Schnell R. Cryptanalysis of basic Bloom filters used for privacy preserving record linkage. *J Pri Confidentiality* (2014) 6(2):59–79.
41. Kroll M, Steinmetzer S. Automated cryptanalysis of bloom filter encryptions of health records. (2014). arXiv preprint arXiv:14106739.
42. Schnell R, Bachteler T, Reiher J. *A Novel Error-Tolerant Anonymous Linking Code. Working Paper Series No. WP-GRLC-2011-02*. Nürnberg, Germany: German Record Linkage Center (2011).
43. Randall SM, Boyd JH, Ferrante AM, Bauer JK, Semmens JB. Use of graph theory measures to identify errors in record linkage. *Comput Methods Programs Biomed* (2014) 115(2):55–63. doi:10.1016/j.cmpb.2014.03.008
44. Schnell R, Borgs C. Secure privacy preserving record linkage of large databases by modified Bloom filter encodings. *2016 International Population Data Linkage Conference*. Swansea, Wales: Swansea University (2016).
45. Schnell R, Borgs C. Randomized response and balanced Bloom filters for privacy preserving record linkage. In: *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*. IEEE (2016). p. 218–24. doi:10.1109/ICDMW.2016.0038
46. Christen P. Advanced computational and privacy methods for data linkage. *2016 International Population Data Linkage Conference*. Swansea, Wales: Swansea University (2016).
47. Vatsalan D, Christen P, O'Keefe C, Verykios V. An evaluation framework for privacy-preserving record linkage. *J Pri Confidentiality* (2014) 6(1):35–75.
48. Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. *Inf Syst* (2013) 38(6):946–69. doi:10.1016/j.is.2012.11.005
49. Durham EA, Kantarcioglu M, Xue Y, Toth C, Kuzu M, Malin B. Composite bloom filters for secure record linkage. *IEEE Trans Knowl Data Eng* (2014) 26:2956–68. doi:10.1109/TKDE.2013.91
50. Irvine K, Hollis S. Multiple operating models for data linkage: a privacy positive. *2016 International Population Data Linkage Conference*. Swansea, Wales: Swansea University (2016).
51. Pow C, Iron K, Boyd J, Brown A, Thompson S, Chong N, et al. Privacy-preserving record linkage: an international collaboration between Canada, Australia and Wales. *2016 International Population Data Linkage Conference*. Swansea, Wales: Swansea University (2016).
52. Adrian Brown CB, Randall S, Schnell R. High quality linkage using multibit trees for privacy-preserving blocking. *2016 International Population Data Linkage Conference*. Swansea, Wales: Swansea University (2016).
53. Boyd J, Ferrante A, Brown A, Randall S, Semmens J. Implementing privacy-preserving record linkage: welcome to the real world. *2016 International Population Data Linkage Conference*. Swansea, Wales: Swansea University (2016).
54. Holman D, Bass A, Rouse I, Hobbs M. Population-based linkage of health records in Western Australia: development of a health services research linked database. *Aust N Z J Public Health* (1999) 23(5):453–59. doi:10.1111/j.1467-842X.1999.tb01297.x
55. Magnusson RS. Data linkage, health research and privacy: regulating data flows in Australia's health information system. *Syd Law Rev* (2002) 24(1):5–55.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

1.9.2 Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581?

Randall SM, Ferrante AM, Boyd JH, **Brown AP**, Semmens JB (2016). *Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581?* Health Information Management Journal.

Article

HIMJ

Health Information Management Journal
2016, Vol. 45(2) 71–79

DOI: 10.1177/1833358316647587
himj.sagepub.com
SAGE

Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581?

Sean M Randall, *BSc*,
Anna M Ferrante, *PhD*,
James H Boyd, *BSc*,
Adrian P Brown, *BSc*,
James B Semmens, *PhD*

Abstract

Background: The statistical linkage key (SLK-581) is a common tool for record linkage in Australia, due to its ability to provide some privacy protection. However, newer privacy-preserving approaches may provide greater privacy protection, while allowing high-quality linkage. **Objective:** To evaluate the standard SLK-581, encrypted SLK-581 and a newer privacy-preserving approach using Bloom filters, in terms of both privacy and linkage quality. **Method:** Linkage quality was compared by conducting linkages on Australian health datasets using these three techniques and examining results. Privacy was compared qualitatively in relation to a series of scenarios where privacy breaches may occur. **Results:** The Bloom filter technique offered greater privacy protection and linkage quality compared to the SLK-based method commonly used in Australia. **Conclusion:** The adoption of new privacy-preserving methods would allow both greater confidence in research results, while significantly improving privacy protection.

Keywords (MeSH)

medical record linkage; privacy; algorithms; evaluation studies; data linkage

Introduction

The last decade has seen a significant increase in population-based research that uses large linked datasets to monitor diseases and assess the effects of treatment. For most of these purposes, complete and accurate data are important; missing or wrongly linked data can bias results and lead to wrong conclusions.

Despite the significant research benefits, the creation of linked datasets carries some risk to individual privacy, as best practice protocols typically require the release of personally identifying information to third-party linkage units (Boyd et al., 2012; Kelman et al., 2002). There is no 'one-size-fits-all' approach to undertaking data linkage for research purposes, and alternative methods have emerged which attempt to reduce privacy risks. The challenge in adopting these methods in operational environments is in achieving high levels of privacy protection without negatively impacting on linkage quality.

In Australia, statistical linkage key (SLK) protocols have proved to be a popular method of providing privacy protection in linkage operations. An SLK is a derived variable generated from components of an individual's personal demographic data (Community Services Ministers Advisory Council, 2004). One of the most well-used SLKs is the SLK-581 – a key originally developed by the Australian Institute of Health and Welfare (AIHW) and used to enable linkage of records from the Home and Community Care (HACC) data collection (Ryan et al., 1999). The SLK-581 was developed to enable linkage of a person's records within and between datasets without explicitly identifying individuals. A number of variations were assessed during the development of the key; however, the SLK-581 proved to be the most effective for the linkage of HACC data, given the particular properties of that data set (Ryan et al., 1999).

The SLK-581 consists of the second and third letters of the individual's first name, the second, third and fifth letters

Centre for Population Health Research, Curtin University, Australia

Accepted for publication April 12, 2016.

Corresponding author:
Sean M Randall, Centre for Population Health Research, Faculty of Health Sciences, Curtin University, Bentley, 6012 Western Australia, Australia.
Email: sean.randall@curtin.edu.au

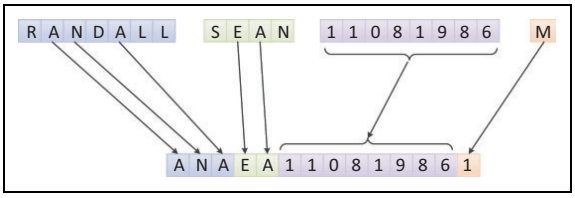


Figure 1. An example of an SLK formed from personal identifiers. SLK: statistical linkage key.

of the surname, the individual's full date of birth and their sex (Ryan et al., 1999). These are amalgamated to form a single field (an example is shown in Figure 1).

Since its introduction, the SLK-581 has been used both to link between the HACC datasets and to link HACC data to other datasets for which personally identifying information is typically available (i.e. hospital admissions datasets; Karmel and Rosman, 2008). The SLK-581 has been implemented across a range of additional datasets, including national datasets on alcohol and drug treatment services (AIHW, 2009), disability services (AIHW, 2013), early childhood education (Australian Bureau of Statistics, 2013) and homelessness (AIHW, 2014).

While the SLK has been used for linkage in numerous research projects, the ability of the key to achieve high linkage quality has been an issue of continuing concern. The limited amount of information contained in the SLK-581 restricts its ability to find all possible matches in a data set, often resulting in a high rate of 'missed links' (i.e. poor sensitivity or recall; Christen, 2012). Early investigations confirmed that achieving high recall was problematic, with multiple SLKs for a single individual occurring in 2–6% of the study population, depending on the length of the study period (Bass and Garfield, 2002). Linkage to death data also showed a lower recall (88.4%) when compared with that achieved through linkage with full named information, with poorer recall affecting research results (Karmel, 2005). The SLK-581 has also been shown to produce substantial rates of missed links that increase over time, resulting in underestimates of hospitalisation rates which vary by health condition (Taylor et al., 2014). Matching rules involving SLKs typically only identify matches where a large proportion of attributes are identical; as a result, low rates of false positive links have been observed (Ryan et al., 1999). In practice, intricate linkage techniques (Karmel et al., 2010), along with additional identifying information that does not form part of the SLK-581, such as postcode, language spoken at home and country of birth (Karmel, 2005), have been used to improve the linkage quality above levels achievable with the standard SLK-581.

While a number of studies have investigated the linkage quality of the SLK-581, much less has been written about the privacy risks associated with use of the key. Although the SLK-581 was developed to reduce the risk of identification of individuals, it consists entirely of unencrypted identifying information specifically selected to be unique for each individual. As such, the risks associated with reidentification remain. The limited privacy safeguards provided by the SLK-581 were noted during its development: '... it is a very common misconception that an SLK by itself does not allow an individual to be identified ... (the SLK) technically could be re-constructed to allow an individual to be re-identified with some degree of accuracy ...' (Community Services Ministers Advisory Council, 2004: 12). Although the SLK-581 can be encrypted (hashed) to improve privacy protection, concerns over the resulting linkage quality have meant that this method was not recommended for use by its original developers (Community Services Ministers Advisory Council, 2004). At the time of its development, the SLK-581 offered a reasonable option regarding safeguarding privacy, while still allowing linkage (of HACC data) to occur.

In recent years, there has been an increased focus on the protection of individual privacy, resulting in the development of new privacy-preserving record linkage methodologies. These methods, which enable record linkage using encrypted or encoded data, have become a popular area of research. One of the most promising privacy-preserving techniques to emerge is the use of Bloom filters (Schnell et al., 2009). This method has the advantage that identifiers are encrypted, yet allow a range of standard probabilistic record linkage techniques to be applied to the encrypted data. For instance, comparison of similarity between (encrypted) names can be undertaken. A recent evaluation of Bloom filter methods on Australian datasets has shown the linkage quality to be equal to that achieved through probabilistic record linkage on full personal identifiers (Randall et al., 2014).

Bloom filter method

The Bloom filter method works by encoding each individual data field into a structure called a Bloom filter (essentially a list of ones and zeros), which are then compared. The encoding process uses a series of hash functions to map elements of the data field to positions within the Bloom filter. A hash function is an algorithm that produces a fixed length output with several important properties. Firstly, given the same input, it will always produce the same output (i.e. the same bigram will always produce the same hash value). The hash functions are also one way, meaning it is not possible to determine the original bigram from the given hash value.

An initialised Bloom filter has a set length (number of positions) with each position set to 0. The data field to be placed into the Bloom filter is split into bigrams – overlapping sets of letters (e.g. the name GLEN could become 'GL' 'LE' 'EN'). Each of these sets of letters is processed by the hash functions, using a secret key. The modulus of these hash functions is then computed on the length of the Bloom filter. This results in each bigram having a number that corresponds to a position in the Bloom filter. These positions are then set to 1 (with all other positions originally being set to 0; see Figure 2). When all required bigrams are added in this way, the Bloom filter is completed and ready for comparison. Each bigram can be hashed multiple times, resulting in multiple positions in the Bloom filter being set to 1 for each bigram. This can be useful to

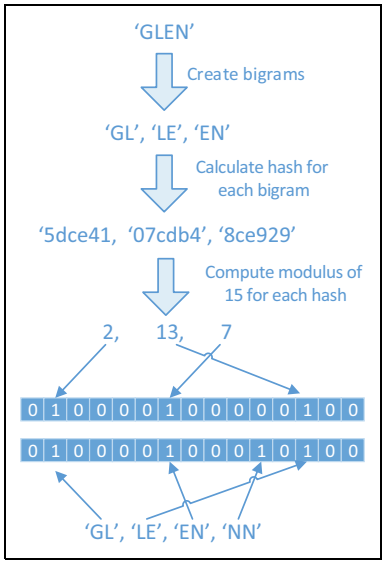


Figure 2. An example of the creation of a Bloom filter of length 15 for the first name Glen and its comparison to another Bloom filter for the name Glenn.

reduce the effects of false positives (which occur when two bigrams map to the same position in the Bloom filter).

Pairs of Bloom filters are compared using the Dice coefficient. The number of positions that are set to '1' in both Bloom filters is totalled, multiplied by 2 and then divided by the total number of positions set to 1 across the two Bloom filters. The Dice coefficient results in a score between 0 and 1, where a higher score reflects greater similarity. In Figure 2, a score of 0.857 (6/7) is returned for the comparison of the Bloom filters for 'Glen' and 'Glenn'.

Record linkage using Bloom filters involves the same techniques used in traditional probabilistic record linkage. Weights are given to each field based on the likelihood of these agreeing, scores are summed across field comparisons, and where this sum is greater than a set threshold, a record-pair comparison is designated a match.

For Bloom filters to be used in practice, it is necessary for records to be encrypted into Bloom filters prior to being released for data linkage (i.e. by data custodians). However, unlike the creation of the SLK-581 which can be undertaken relatively easily, the encryption of data into Bloom filters is more complex; in practice, computer programmes are required to be run by data custodians, potentially making the process more involved. The data custodians involved in the project would agree on the secret key used to hash the data, which would not be shared with the linkage unit. The custodians must also agree on a number of parameters (including the length of the Bloom filter, the number of hashes, the particular hash function and whether to use padding), which all custodians must use.

Further detailed technical information on how to implement Bloom filters can be found in the original paper by Schnell et al. (2009).

Privacy risks in record linkage

While the aim of many privacy-preserving methods is to reduce the privacy risks associated with record linkage, the precise nature of these risks is not always conveyed. Algorithms are often presented without reference to privacy requirements or to any privacy protection standards that may be in operation. This makes it difficult to measure and compare the privacy protections offered by alternative linkage protocols. In the case of the SLK-581, for example, no statement of privacy standards and/or legal obligations as per Australian law has ever been laid down.

The *Privacy Act* (1998) and, more recently, the *Privacy Amendment (Enhancing Privacy Protection) Act* (2012) are Australian Commonwealth laws that articulate a number of privacy principles (APPs) regarding the collection, use, storage and disclosure of personal information. APP6, on data disclosure, is most relevant to data linkage, as current practices involve the release of identifiable information to a trusted third party (a data linkage unit). Most Australian states have similar laws with roughly equivalent principles that apply at a local level (Lovett et al., 2008).

What constitutes personal information is of primary importance in discussions about privacy. According to the Commonwealth Acts, personal information is defined as information about an identified individual or about an individual who is 'reasonably identifiable'. If personal information can be transformed in such a way that it is no longer reasonably identifiable, then it may be possible for organisations to release information to third parties to enable linkage to occur. The critical test in these situations is to determine whether the transformed information is reasonably identifiable. Different interpretations of this term exist (O'Keefe and Connolly, 2011), and the cost, difficulty and practicality of identification often factor into the decision-making process (Australian Government, 2012).

With this concept of personal information as reasonably identifiable, we present three scenarios which we use to both understand and assess the privacy risks associated with the process of record linkage. These scenarios have been previously used in evaluating reidentification risk (El Emam et al., 2009). In describing these scenarios, we assume that record linkage is conducted by a third party using the separation principle (Kelman et al., 2002), whereby clinical or content data are not made available to the linkage unit. We note that while the separation principle can remove some privacy risk, the determination of an individual's existence within certain datasets (for instance a mental health data set) can still be considered highly sensitive.

Scenario 1: Risk of accidental recognition

Those with access to the data set may, in the course of their normal duty, unwittingly stumble upon a record of an individual whom they know. Such accidental disclosure is most likely to occur where data processing is heavily manual, and where manual review of individual records forms a key part of business processes (Lawrence et al., 2008; Rosman et al., 2002).

Scenario 2: Risk of determining whether an individual is contained within a database

Those with access to the data set may attempt to determine whether someone they know is contained within the data set. Such acts though rare have been known to occur; staff in the US intelligence services have used their surveillance capabilities to spy on their romantic interests (The Wall Street Journal, 2013), while Australian Vodafone staff reportedly examined call logs of spouses (IT News, 2011). The extent to which these actions occur in a data linkage context is difficult to gauge, given that such breaches are usually discovered through detailed audits or self-reporting.

Scenario 3: Risk of reidentifying individuals contained in the data set

In this scenario, those with access to the data attempt to determine the identity of individuals contained within the data set using both the available records in the data set and any available public information, such as that which could be found on the Internet or in public documents. The Information Commissioner's Office in the United Kingdom suggests designing safeguards based on this motivated intruder scenario (Graham, 2012).

Objective

In this article, we compare the privacy protection and linkage quality offered by three linkage methods – privacy-preserving record linkage using Bloom filters, the standard SLK-581 and the encrypted SLK-581. To assess privacy risks, we evaluate the ability of each method to reduce or mitigate privacy risks, in relation to the three privacy scenarios presented above. To assess linkage quality, we undertake a quantitative study of the linkage accuracy of both methods when applied to a range of real-world Australian health datasets.

Method

Assessment of privacy risk

To measure privacy, we undertake a qualitative approach rather than a quantitative approach. Although quantitative metrics for measuring privacy have been proposed in the literature, these do not appear appropriate for our work. For instance, measures of entropy can be used to estimate privacy, with higher entropy indicative of higher privacy. However, this metric is more suitable for measuring privacy in techniques that seek to add noise to the original data (Bertino et al., 2008); entropy cannot distinguish between plainly readable private data and garbled data with the same predictability.

A second method (Vatsalan et al., 2014) proposes to measure privacy by examining, for each encoded element of a record, the proportion of other records with the same value of this element – the larger the number, the greater the privacy. This metric again appears inappropriate – Bloom filtered data would receive the same score as plain readable data, despite the obvious privacy improvements.

Instead, to assess the linkage protocols against the three scenarios described above, we appraise each along two dimensions: the impact of reidentification (i.e. the risk of harm if an individual is identified) and the level of privacy risk remaining. While alternate quantitative measures of information disclosure exist in the literature (Vatsalan et al., 2014), these measures focus on the amount of information disclosure technically possible, without distinguishing the practicality or significance of any disclosure.

Numerous factors determine the overall impact of a privacy breach. More sensitive data pose a greater risk of harm (Office of the Australian Information Commissioner, 2014), with health data widely accepted as being among the most sensitive (Office of the Australian Information Commissioner, 2011). The amount of disclosed information (both in terms of the number of total individuals affected and in terms of the breadth of information disclosed on a particular individual) will also directly affect the overall impact of a breach. The motivation of any privacy breach (whether accidental or malicious) to some extent also determines the overall impact (Office of the Australian Information Commissioner, 2014).

The level of privacy risk remaining after a privacy-preserving technique has been used (i.e. residual privacy risk) is determined by the difficulty of reidentification, and the extent of reidentification possible. This can range from an individual being able to reidentify all data without any difficulty (such as with un-obfuscated data) to a technique that would make it impossible to reidentify at all.

We utilise decision matrices (Figure 3), measuring the level of residual privacy risk remaining, and the overall impact of any reidentification, to evaluate the privacy impact of each linkage method. These risks are measured with respect to the three scenarios outlined above.

Linkage quality evaluation

To evaluate linkage quality, deduplications of four datasets (described below) were conducted using both the SLK-581 method and the Bloom filter method. A standard linkage strategy was employed (also described below).

Datasets

Five large-scale Australian hospital datasets were used in this evaluation. These were hospital admissions records for New South Wales (NSW) from both public and private hospitals, emergency presentations for New South Wales and hospital admissions and emergency presentations for South Australia (SA). Only public hospital data were available from SA. Each data set contained all records for the years 2008–2010. The data (personal identifiers and not clinical information) were made available for data linkage as part of a Population Health Research Network Proof of Concept project (Mitchell et al., 2015). Ethical approvals for the Proof of Concept research project were obtained from health and human ethics committees in SA Health

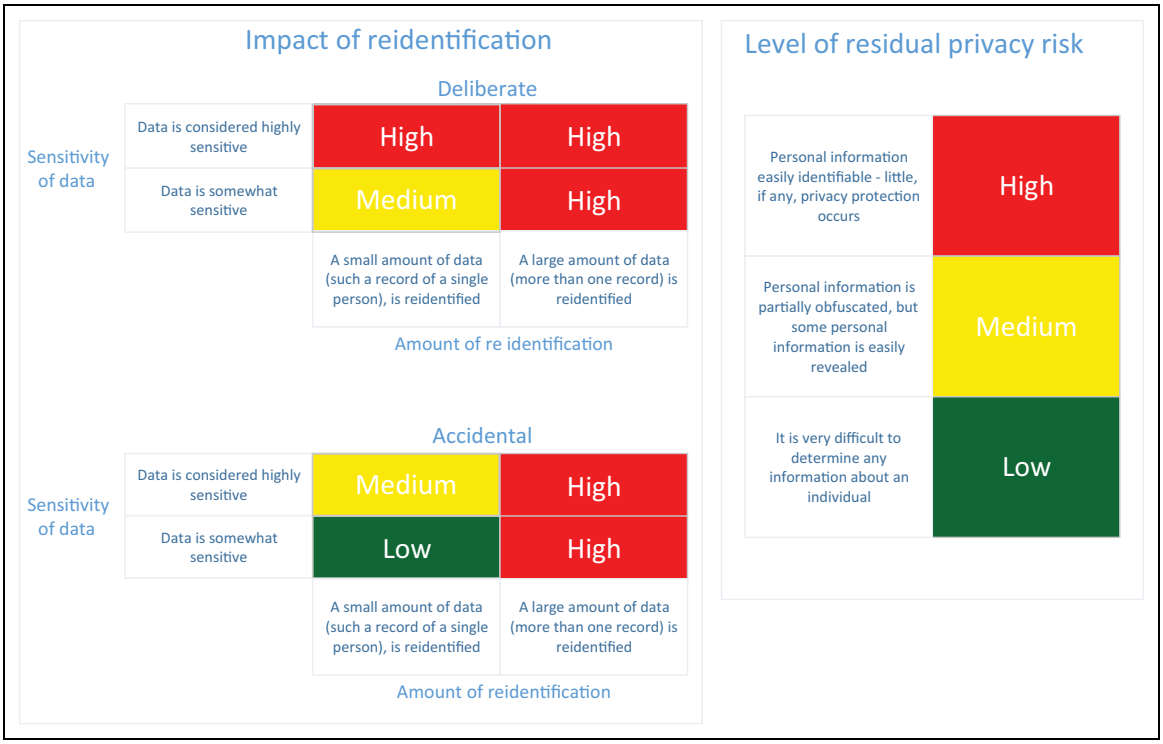


Figure 3. Risk and reidentification assessment framework.

and the Cancer Institute NSW. Additional approval to conduct the linkages and to undertake research into record linkage methods was obtained from Curtin University (WA). Formal Data Agreements between the relevant data custodians in each jurisdiction and Curtin University enabled the release of personal identifiers. The quality of these datasets differed, with NSW private hospital admissions records missing all name information and a quarter of address data.

Linkage strategy

The SLK-581 was computed for each record based on described methods. Records that had the same SLK-581 were marked as belonging to the same person.

The Bloom filter was computed based on methods previously reported (Randall et al., 2014). Individual fields were transformed into Bloom filters, weights were estimated using Jaro's method (Jaro, 1989) and an appropriate threshold was estimated. All available variables (name, address and date of birth) were used in the comparisons.

Measuring linkage quality

The results from the linkages were compared using three standard linkage metrics: precision, recall and F-measure (Christen and Goiser, 2007). Each of these metrics returns a score between 0 and 1, where a higher number indicates higher quality. Precision (otherwise known as positive predictive value) indicates the percentage of found record pairs that were correct (i.e. measuring the rate of false positives). Recall (otherwise known as Sensitivity) measures the percentage of correct record pairs that were identified (i.e. a measure of false negatives). The F-measure is the harmonic mean of these previous measures and gives a single score with which to compare results.

To estimate the values of these metrics, random samples of record pairs, which were identified by one linkage method but not by the other, were selected and manually reviewed. Decisions made by this manual review process constituted our 'gold standard'. No evaluation was carried out of the pairs deemed to be correct by both methods or incorrect by both methods. We assumed in our calculations that all pairs matched by both methods were correct and that all pairs deemed incorrect by both methods were thus so. It should be noted, therefore, that by adopting this approach, our results are only suitable for comparing methods, and they do not tell us anything about the overall quality of these methods for record linkage. In other words, our results are thus a measure of *relative*, and not *absolute*, linkage quality. The relative precision, recall and F-measure of each privacy-preserving linkage method were further subtracted to determine the percentage difference found between the two methods.

Results

Privacy

Results from the assessment of privacy protection are shown in Table 1. Due to the highly sensitive nature of

Table 1. Privacy assessment of SLK-581 and Bloom filter methods.

	Standard SLK-581	Encrypted SLK-581	Bloom filter method
Scenario 1			
Level of residual privacy risk	Medium	Low	Low
Impact of reidentification	Medium	Medium	Medium
Scenario 2			
Level of residual privacy risk	High	Medium	Low
Impact of reidentification	High	High	High
Scenario 3			
Level of residual privacy risk	Medium	Low	Low
Impact of reidentification	High	High	High

SLK: statistical linkage key.

health datasets, the impact of deliberate reidentification arising from scenarios 2 and 3 was deemed high, while the accidental reidentification of a small amount of information in scenario 1 was considered to be of medium-level impact.

Both the encrypted SLK-581 and the Bloom filter method provide reasonable protection against accidental recognition. As the standard SLK-581 contains unencrypted identifiable data, there is still a possibility of accidental recognition; however, the removal of specific characters from name information makes immediate identification less likely. For the Bloom filter method and the encrypted SLK-581, all data are completely garbled; as such, there is no possibility of identifying individuals this way.

The standard SLK provides no security against determining whether a particular individual is contained within a data set. An SLK or partial SLK can be created from a person's known identifiers and the data set searched to determine whether this SLK exists. An encrypted SLK requires one further step (for the created SLK to be encrypted) before the data set is then searched. The Bloom filter method provides a higher level of security against this privacy breach, as knowledge of a secret key is required to convert identifiers into the correct format for potential comparison.

Both the SLK-581 and the Bloom filter method provide some level of privacy against determining the identity of an individual from a single record. As the standard SLK reveals personal information such as date of birth, gender and components of the individual's name, it provides less privacy than the encrypted SLK and the Bloom filter method, which do not easily reveal any of this information. Several sophisticated frequency-based attacks against the Bloom filter protocol have been reported. Niedermeyer et al. (2014) presented an attack that utilises a weakness in the original hashing implementation to reveal personally identifying information – by adopting a more appropriate hashing method, this attack is nullified. Kuzu (Kuzu et al., 2011) presented a constraint satisfaction frequency attack which using the default parameters was able to reidentify 11% of first names. However, by modifying the parameters used in creating the Bloom filters, the encoding was made more resistant, with the computation time required in particular increasing exponentially. Assuming the parameters are set appropriately, given the difficulty of successfully carrying out such attacks, the risk of reidentification is classified as low (see Figure 3).

Linkage quality

The results found from the deduplications using SLK and Bloom filter methods are shown in Table 2. As both the standard SLK and the encrypted SLK return the same results, only one is shown here. The vast majority of identified record pairs were found using both linkage methods. Both methods identified record pairs as correct links that were not found by the other method, with the Bloom filter method identifying a larger number. Manual review of random samples of record pairs showed that the majority of record pairs only identified by the SLK were incorrect (between 60% and 100%, depending on the data set), while the majority of record pairs only identified by the Bloom filter method were correct (between 80% and 97.5%).

The NSW private hospital admissions data set was an outlier due to the absence of name information; the SLK linkage method was not appropriate for this data set. The Bloom filter method, which could utilise available address information, provided more realistic results.

Analysis of the reviewed record pairs provided reasons for the difference in linkage quality. The correct links missed by the SLK method typically involved women who had changed surname within the 3-year period, name misspellings and slight differences in date of birth. Incorrect links found by the SLK often involved completely different name information that had the specific letters of the SLK in common (an invented example; WARREN BEATIE and MARCO DEALIANO). The incorrect links found by the Bloom filter method often involved records of twins.

The Bloom filter method provided superior linkage quality for all datasets (Table 2). For datasets excluding the NSW private hospital data, it found an additional 1–2% correct links. To find these additional links, the Bloom filter method introduced a small number of errors, reducing precision by on average 0.1% (excluding NSW private hospital data). Again, the NSW private hospital data were a large outlier due to its lack of names.

While the difference in overall percentages between methods was small, this was due to the very large number of record pairs found by both the SLK and Bloom filter methods, which somewhat obscured the often large number of additional correct pairs found by the Bloom filter method (approximately 1.8 million additional record pairs for the NSW public hospital data set).

An additional linkage was performed using individual components of the SLK-581 to improve linkage quality (as described by Karmel et al. 2010). This linkage strategy is only possible using the unencrypted SLK method. However, utilising this linkage strategy resulted in linkage quality worse than that achieved by the SLK on its own.

Table 2. A comparison of linkage quality between the Bloom filter method and SLK-581.

	NSW hospital public	NSW hospital private	NSW emergency	SA hospital	SA emergency
Number of record pairs which were identified by:					
Both linkage methods (a)	142,320,572	17,422,896	10,158,434	33,798,287	2,359,573
SLK-581 method only (b)	53,558	52,016,277	11,794	1,282	2,968
Estimated % of these which were correct (c)	10	0	23	25	40
Bloom filter method only (d)	1,911,435	1,053	229,171	393,481	62,341
Estimated % of these which were correct (e)	97.5	81	89	89	90
Estimated total correct pairs[a] (f)	144,189,577	17,423,749	10,365,109	34,148,806	2,416,867
Bloom filter method as compared with SLK-581 method					
Estimated difference in Precision[b] (g)	0.0	+74.9	−0.2	−0.1	−0.2
Estimated difference in Recall[c] (h)	+1.3	0.0	+1.9	+1.0	+2.3
Estimated difference in F-Measure[d]	+0.6	+59.9	+0.9	+0.5	+1.1

SLK: statistical linkage key.

[a]The number of pairs found in both methods plus the estimated number of correct pairs found from either method. Calculated as $f = a + (b \times c) + (d \times e)$.

[b]Precision is first calculated for both SLK and Bloom filter method. $Prec_{SLK} = \frac{a + (b \times c)}{(a + b)}$, $Prec_{BF} = \frac{a + (d \times e)}{a + d}$. Result calculated as $(Prec_{BF} - Prec_{SLK}) \times 100$.

[c]Recall is first calculated for both SLK and Bloom filter method. $Recall_{SLK} = \frac{a + (b \times c)}{f}$, $Recall_{BF} = \frac{a + (d \times e)}{f}$. Result calculated as $(Recall_{BF} - Recall_{SLK}) \times 100$.

[d]Result calculated as $(FMeasure_{BF} - FMeasure_{SLK}) \times 100$. Standard formulae used for calculated F-measure from precision and recall. $FMeasure = \frac{2 \times Precision \times Recall}{Precision + Recall}$.

Discussion

As demand for record linkage services grows, it is imperative for linkage units to implement methods that protect privacy and safeguard security, while maximising the benefits that can be derived from administrative data.

Privacy protection

Assessment of privacy protection showed the Bloom method to be superior to the standard SLK-581 and similar to the encrypted SLK method. It was clear that the standard SLK-581 has constraints regarding privacy. Most notably, the SLK provides no mechanism to stop individuals from determining whether a specific individual is located within a data set, and individual data fields are easily identifiable from the provided data. Given this, and by Australian legislation, it is likely that these data could be classed as reasonably identifiable.

The Bloom filter method does not suffer from these deficiencies. While complex attacks on Bloom filters have shown that in certain circumstances some information (for instance, that a record has a common surname, such as 'SMITH') can be leaked (Kuzu et al., 2011), these attacks are both difficult to implement and provide limited information typically already available to an individual with access to an SLK (an SLK with MIH as the second, third and fifth letters of surname will be very likely to have the surname 'SMITH'). The difficulty of attacking these protocols means that the Bloom filter method, along with the encrypted SLK, would probably not be classified as reasonably identifiable by current Australian law.

Linkage quality

From the perspective of linkage quality, our results showed that the Bloom filter method provided superior quality, finding additional links missed by SLK matching. Manual assessment suggests that a large proportion of these missed links were caused by surname changes among females (presumably due to marriage). The datasets used for this study only spanned 3 years; when linking datasets over longer periods, it is likely that the SLK will miss an even greater proportion of links, due to the increased proportion of individuals changing their name. The SLK method was also more restrictive, requiring datasets to contain specific fields; where these were not recorded or available for linkage (such as in the NSW private hospital data set), this method is no longer practical. By utilising probabilistic linkage techniques rather than exact matching, the Bloom filter method can cope with poorer data quality and utilise additional data elements as available, allowing it to be useful in a much greater range of scenarios.

The SLK uses fewer data elements (no address information) than the Bloom filter approach demonstrated in this article, and this is undoubtedly part of the reason for the lower linkage quality. The SLK was purposely designed to use only a few data elements, as it cannot handle any difference in attributes; only records with identical matches on all selected attributes are accepted. For this reason, attributes such as address that can regularly change cannot be used within an SLK. The SLK's lack of flexibility with regard to available personal identifiers can be considered part of its weakness.

Although differences in data quality will vary between linkages and datasets, it is clear that any improvement will have practical significance. For record linkage units in Australia, a large amount of time and money is invested into achieving the maximum possible linkage quality; essentially any improvement in linkage quality is considered worthwhile. The use of extensive clerical review to improve quality is common, despite the large overheads of this approach. A quality improvement of the size found between SLKs and Bloom filters would require many thousands of man-hours of clerical review.

Our linkage quality results provided only relative and not absolute measures of linkage quality. Absolute measures of linkage quality require comparisons with an answer sheet; for instance, answers derived from full probabilistic linkage with clerical review. Several previous studies have provided absolute measures of linkage quality for both SLKs and Bloom filter–based record linkage; the SLK method, in particular, has received numerous evaluations due to its wide implementation. Privacy-preserving linkage using Bloom filters has been shown to achieve equal results to those found with linkage on un-encoded data (Randall et al., 2013), while poorer linkage quality has been found when SLKs have been tested against full probabilistic linkage (Bass and Garfield, 2002; Karmel, 2005; Taylor et al., 2014).

Despite deficiencies in linkage quality, the SLK method has several advantages. The Bloom filter method is more complex, both to implement and to understand, than the SLK method. In practice, tools need to be provided to data custodians to allow them to encrypt data in the required way. The SLK's simplicity has meant that it can be calculated directly at the point of collection, on paper if necessary. It has been noted that SLKs calculated at the point of collection typically have higher error rates, which can seriously affect the quality of any analysis performed (Community Services Ministers Advisory Council, 2004). However, for particularly sensitive datasets (such as the SAAP homelessness data set; Community Services Ministers Advisory Council, 2004), data custodians themselves may not want to hold further identifying information. The Bloom filter method is not equipped to be computed by hand, and in scenarios where collection on paper is required, the SLK method is advantageous. For all datasets in which personally identifying information is held by data custodians, or in which all recording is computerised, data released for linkage should be encrypted to decrease the likelihood of identification of individuals. Where ensuring high linkage quality is also important, the use of Bloom filters instead of SLKs should be recommended.

There are few current privacy-preserving alternatives to the SLK and Bloom filter methods. While numerous alternative methods have been presented in the literature (Vatsalan et al., 2013), these are in an earlier stage of development and require further testing on large real-world datasets to evaluate their potential before they can be considered practical alternatives. One alternative is Grhanite (Boyle and Rafael, 2011). Grhanite's privacy-preserving method is proprietary and is understood to be based on a similar principle to the encrypted SLK.

Conclusion

Administrative health data are highly sensitive, containing personal information about individuals that should not be disclosed. At the same time, it has been argued there are economic and moral imperatives to utilise this rich source of information through linked health research (Hetzel, 2006). The emergence of new methods that are both privacy-preserving and highly accurate can satisfy both of these demands. While there is no 'one-size-fits-all' method for privacy-preserving record linkage, there appear very few scenarios in which the regular SLK-581 method would be preferable to use over privacy-preserving linkage using Bloom filters. While the adoption of a new protocol would undoubtedly have short-term costs, it would ensure strong privacy protection and highly accurate research in the longer term.

Acknowledgements

This study would not have been possible without the collaboration, assistance and expertise provided by the NSW Ministry of Health and by the South Australian Department for Health and Ageing.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy's Population Health Research Network.

References

Australian Bureau of Statistics (2013) *National Early Childhood Education and Care Collection: Data Collection Guide. Cat. 4240.0.55.002*. Canberra:Australian Bureau of Statistics.

Australian Government (2012) *Privacy Amendment (Enhancing Privacy Protection) Bill 2012, Explanatory Memorandum*. Canberra: Australian Government.

Australian Institute of Health and Welfare (AIHW) (2009) *Enhancing the Alcohol and Other Drug Treatment Services National Minimum Data Set*. Canberra: AIHW.

Australian Institute of Health and Welfare (AIHW) (2013) *Disability Services Minimum Data Set: Data Guide*. Canberra: AIHW.

Australian Institute of Health and Welfare (AIHW) (2014) *Specialist Homelessness Services Collection Data Quality Statement 2013-14*. Available at: http://meteor.aihw.gov.au/content/index.phtml/itemId/593778 (accessed 26 April 2016).

Bass J and Garfield C (2002) Statistical linkage keys: how effective are they? *Symposium on Health Data Linkage, Sydney 2002*. Available at: http://www.publichealth.gov.au/symposium.html (accessed 26 April 2016).

Bertino E, Lin D and Jiang W (2008) A survey of quantification of privacy preserving data mining algorithms. *Privacy-preserving Data Mining* 34: 183–205.

Boyd JH, Ferrante AM, O'Keefe CM, et al. (2012) Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC Health Services Research* 12: 480.

Boyle DIR and Rafael N (2011) BioGrid Australia and GRHANITE: privacy-protecting subject matching. *Studies in Health Technology and Informatics* 168: 24–34.

Christen P (2012) *Data Matching*. New York: Springer.

Christen P and Goiser K (2007) Quality and complexity measures for data linkage and deduplication. In: Guillet F and Hamilton H (eds) *Quality Measures in Data Mining Studies in Computational Intelligence*. New York:Springer, pp. 127–151.

Community Services Ministers Advisory Council (2004) *Statistical Data Linkage in Community Services Data Collections*. Canberra: Community Services Ministers Advisory Council.

El Emam K, Dankar FK, Vaillancourt R, et al. (2009) Evaluating the risk of re-identification of patients from hospital prescription records. *The Canadian Journal of Hospital Pharmacy* 62: 307.

Graham C (2012) Anonymisation: managing data protection risk code of practice. Cheshire, UK: Information Commissioner's Office. Available at: https://ico.org.uk/media/1061/anonymisation-code.pdf (accessed 26 April 2016).

Hetzel D (2006) Data linkage research—can we reap benefits for society without compromising public confidence? *Australian Health Consumer* 2: 27–28.

IT News (2011) *Vodafone Sacks Staff Over Alleged Security Breach*. Available at: http://www.itnews.com.au/News/244672,vodafone-sacks-staff-over-alleged-security-breach.aspx (accessed 26 April 2016).

Jaro MA (1989) Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* 89: 414–420.

Karmel R (2005) *Transitions between Aged Care Services. Data Linkage Series no. 2*. Canberra: AIHW.

Karmel R, Anderson P, Gibson D, et al. (2010) Empirical aspects of record linkage across multiple datasets using statistical linkage keys: the experience of the PIAC cohort study. *BMC Health Services Research* 10: 41.

Karmel R and Rosman D (2008) Linkage of health and aged care service events: comparing linkage and event selection methods. *BMC Health Services Research* 8: 149.

Kelman C, Bass A and Holman D. (2002) Research use of linked health data: a best practice protocol. *Australian and New Zealand Journal of Public Health* 26: 5.

Kuzu M, Kantarcioglu M, Durham E, et al. (2011) A constraint satisfaction cryptanalysis of Bloom filters in private record linkage. Privacy Enhancing Technologies pp. 226–245. Berlin Heidelberg: Springer.

Lawrence G, Dinh I and Taylor L (2008) The centre for health record linkage: a new resource for health services research and evaluation. *Health Information Management Journal* 37: 60–62.

Lovett R, Fisher J, Al-Yaman F, et al. (2008) A review of Australian health privacy regulation regarding the use and disclosure of identified data to conduct data linkage. *Australian and New Zealand Journal of Public Health* 32: 282–285.

Mitchell RJ, Cameron CM, McClure RJ, et al. (2015) Data linkage capabilities in Australia: practical issues identified by a population health research network 'proof of concept project'. *Australian and New Zealand Journal of Public Health* 39: 319–325.

O'Keefe CM and Connolly C. (2011) Regulation and perception concerning the use of health data for research in Australia. *Electronic Journal of Health Informatics* 6: 16.

Office of the Australian Information Commissioner (2011) Privacy Fact Sheet 2: National Privacy Principles. Sydney. Available at: https://www.oaic.gov.au/privacy-law/privacy-archive/privacy-resources-archive/privacy-fact-sheet-2-national-privacy-principles (accessed 26 April 2016).

Office of the Australian Information Commissioner (2014) Data breach notification guide: a guide to handling personal information security breaches. Available at: http://www.oaic.gov.au/images/documents/privacy/privacy-resources/privacy-guides/data-breach-notification-guide-august-2014.pdf (accessed 26 April 2016).

Randall SM, Ferrante AM, Boyd JH, et al. (2013) Privacy-preserving record linkage on large real world datasets. *Journal of Biomedical Informatics* 50: 205–212.

Randall SM, Ferrante AM, Boyd JH, et al. (2014) Privacy-preserving record linkage on large real world datasets. *Journal of Biomedical Informatics* 50: 205–212.

Rosman D, Garfield C, Fuller S, et al. (2002) Measuring data and link quality in a dynamic multi-set linkage system. *Sydney (NSW): Symposium on Health Data Linkage*. Available at: http://www. publichealth. gov. au/symposiumpdf/rosman_a.pdf (accessed 26 April 2016).

Ryan T, Holmes B and Gibson D. (1999) *A National Minimum Data Set for Home and Community Care*. Canberra: Australian Institute of Health and Welfare.

Schnell R, Bachteler T and Reiher J (2009) Privacy-preserving record linkage using Bloom filters. *BMC Medical Informatics and Decision Making* 9(1): 41.

Taylor LK, Irvine K, Iannotti R, et al. (2014) Optimal strategy for linkage of datasets containing a statistical linkage key and datasets with full personal identifiers. *BMC Medical Informatics and Decision Making* 14: 85.

The Wall Street Journal (2013) *NSA Officers Spy on Love Interests*. Available at: http://blogs.wsj.com/washwire/2013/08/23/nsa-officers-sometimes-spy-on-love-interests/ (accessed 26 April 2016).

Vatsalan D, Christen P, O'Keefe CM, et al. (2014) An evaluation framework for privacy-preserving record linkage. *Journal of Privacy and Confidentiality* 6: 3.

Vatsalan D, Christen P and Verykios VS (2013) A taxonomy of privacy-preserving record linkage techniques. *Information Systems* 38: 946–969.

Chapter 2

Maximising linkage quality

Included Manuscript(s):

2. **Brown AP**, Randall SM, Ferrante AM, Semmens JB, Boyd JH (2017). *Estimating parameters for probabilistic linkage of privacy-preserved datasets* BMC Medical Research Methodology, 17(1), 95. https://doi.org/10.1186/s12874-017-0370-0

3. **Brown AP**, Randall, SM, Boyd, JH, Ferrante, AM (2019). *Evaluation of approximate comparison methods on Bloom filters for probabilistic linkage.* International Journal of Population Data Science, 4(1). https://doi.org/10.23889/ijpds.v4i1.1095

Conference proceeding(s):

8. Randall SM, Ferrante AM, Boyd JH, **Brown AP**, Semmens JB (2015). *Grouping methods for ongoing record linkage* (2015) Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference.

Letter(s):

10. Boyd JH, Ferrante AM, Irvine K, Smith M, Moore E, **Brown AP**, Randall SM (2016). *Understanding the Origins of record linkage error and how they affect research outcomes* (2016) Australia and New Zealand Journal of Public Health. https://doi.org/10.1111/1753-6405.12597

Aim 2 of this thesis is to *identify methods for maximising the quality of privacy-preserving record linkage that do not rely on manual clerical review*. This chapter looks at quality techniques that can be applied during different parts of the record linkage pipeline to reduce the need for post-linkage quality processes. As datasets become larger, the effectiveness of traditional manual quality processes following linkage is reduced as they are unable to scale. Techniques that can scale and can maximise quality early in the pipeline will have the biggest impact in supporting automation of massive data linkages on commercial cloud infrastructure.

2.1 The effect of linkage error on research outcomes

Data linkage units strive to maximise the quality of their linkages as the results are used directly in research and for policymaking. From cleaning of the data through to manual inspection of records at the completion of linkage, the minimisation of linkage error is critical. Linkage errors have been found to affect research results [72], although the direct effects of different types of errors on particular analysis methods are unclear.

The paper, *Understanding the origins of record linkage errors and how they affect research*, included as a supporting manuscript for this thesis, describes the impact of linkage quality on research outcomes and makes some recommendations for both researchers and linkage units. Researchers should take time to understand the data used in their studies as well as how the data was brought together through data linkage. Data linkage units should provide greater transparency and improve the reporting of linkage processes and results. Together, these will improve study design, help researchers understand the impact of analytical techniques and strengthen the interpretation of results. However, ensuring the quality and integrity of research based on linked data ultimately requires data linkage processes and methods that can achieve high-quality linkage.

2.2 Factors affecting linkage quality

All of the methods and techniques used in the data linkage process can influence the resulting quality of the linkage. Refining the linkage methods and techniques used throughout the data linkage process will all impact on the resulting linkage quality. This does not mean, for example, that more cleaning and standardisation of data during the very beginning of the linkage process will result in better quality. In fact, it has been shown that too much cleaning can reduce the overall quality of the linkage [237]. However, it is essential that the methods used are optimised to achieve the best possible linkage quality. While this may be relatively straightforward with traditional unencrypted linkage, the use of privacy-preserving techniques adds another set of challenges in achieving high quality. Most research on privacy-preserving techniques has focussed primarily on security and privacy. The resultant accuracy or 'quality' of these techniques has often been overlooked.

Of all PPRL methods, the Bloom filter method appears to be the most promising for operational use. As such, the research in this thesis focuses on the Bloom filter method as a basis for new work.

2.2.1 Data pre-processing

Datasets can vary significantly in quality. Good quality datasets may result from a total survey error (TSE) approach to minimising error during the original collection and processing of data [23]. However, it is unlikely that a TSE approach is used for all datasets, particularly those administrative datasets sourced from operational environments. Missing values, placeholders, default values, spelling mistakes and poor management all lead to a degradation in the quality of data. These poor quality characteristics create challenges for data linkers. Placeholders and default values can be found during data cleaning, with the use of lookup tables helping to eliminate the majority of these. For privacy-preserved data, these manual quality methods are not available, and every effort must be made to identify data quality issues before the data are encoded. Standard and reproducible pre-processing routines are required during the encoding process to ensure data cleanliness and consistency between datasets.

Missing values in some fields may be imputed. For example, a missing sex field could be imputed through the use of a lookup table based on the given name. For most fields, however, a missing value remains so and requires consideration during the matching process. For field-level Bloom filters, the missing value can still be represented as a missing value. Composite Bloom filters do not account for the number of identifiers for which valid information is present, and the calculation of similarities between composite Bloom filters will be particularly affected by asymmetrically missing identifiers.

As one of the very first steps in the data linkage process, efficient and effective data pre-processing is a significant factor affecting linkage quality. Poor or insufficient pre-processing techniques at this stage can exacerbate quality issues for future steps in the process. Moreover, for privacy-preserving record linkage, there is no opportunity to correct this at a later time.

2.2.2 Optimising matching strategies in linkage

The matching strategy employed in linkage is typically designed to take into account the quality and attributes of the data being linked. If the quality of the data are excellent, a simple deterministic strategy may be sufficient. This assumes that there are no missing values, spelling errors or typographical errors. However, perfectly formed data is rarely the case, so a more sophisticated approach is often required.

One alternative privacy-preserving approach to deterministic linkage is the use of composite Bloom filters. The collection of identifiers usually used for deterministic matching can be encoded into a single Bloom filter. Approximate comparisons between these composite Bloom filters can then be used with a suitable threshold value to determine matches and non-matches.

This approach provides better privacy and allows for small variations in field values. Composite Bloom filters may be useful for private indexing or in situations where a single linkage field is desirable, but handling missing values and identifiers that change over time (such as address) remain issues.

Many data linkage units prefer probabilistic record linkage due to its proven track record of producing high linkage quality from unencrypted identifiers [2, 42, 114]. Single field Bloom filters used within a probabilistic framework have shown to produce high linkage quality with reasonable efficiency [238]. However, one of the challenges in a practical probabilistic PPRL approach is how to accurately estimate parameter settings. Typical methods for estimating parameters rely on manually examining small samples of data. In the privacy-preserving case, the data are not available for examination, so alternate estimation methods are required.

The paper, *Estimating parameters for probabilistic linkage of privacy-preserved datasets*, included as part of this thesis, presents a method for estimating the probabilities and threshold values required when using privacy-preserved record linkage using Bloom filters in a probabilistic record linkage framework. This probability estimation method produced linkage results comparable to that of the calculated probabilities, even with datasets with as much as 20% introduced error. The threshold estimation technique produced values higher than the optimum for calculated probabilities, but values lower than the optimum for estimated probabilities. The linkage quality at the estimated threshold levels was still relatively high; however, additional simulation studies may help produce threshold values that are closer to the optimal value.

An extension to the Fellegi-Sunter model of record linkage allows for approximate matches between fields. An approximate match is typically assigned a 'similarity score'. These scores are then converted into partial weights; somewhere between the full disagreement weight and the full agreement weight, depending on the similarity. The use of partial agreement has been shown to improve the linkage quality greatly when compared to the use of exact comparisons [89, 222, 296].

The paper, *Evaluation of approximate comparison methods on Bloom filters for probabilistic linkage*, included as part of this thesis, evaluates the effectiveness of three approximate comparison methods for Bloom filters within the context of the Fellegi-Sunter model of recording linkage: Sørensen–Dice coefficient, Jaccard similarity and Hamming distance. The results of these evaluations showed the use of these approximate comparison methods over exact comparisons within a probabilistic framework provides a consistent improvement in recall while maintaining a high level of precision. However, there is still a trade-off between missed matches and incorrect matches, and care must be taken in selecting appropriate threshold values during linkage.

2.2.3 Leveraging good quality data

An essential aspect of the data linkage process is the grouping of matches and potential matches. This process determines how to combine record-pairs into groups representing individuals. A common approach is to use transitive closure to merge all record-pairs, with all connected records being assigned to the same individual. However, if some of these datasets have already been linked (as part of an enduring linkage, for example), it is likely that the merging of existing groups will be undesirable. An alternative to merge grouping has been suggested previously [159]. The best-link method groups records in the order in which they are matched, finding the best existing group in the repository to join using the highest pair weight for the record. This ensures existing groups are never brought together, relying on the quality of the groupings.

The paper, *Grouping methods for ongoing record linkage*, included as a supporting manuscript for this thesis, presents a new grouping strategy (referred to as weighted best-link) and evaluates this algorithm against the standard merge based method and an alternative best-link algorithm. Using real-world datasets, the evaluation showed that the best-link methods were superior to the merge method when the existing data repository had an error rate of 2.5% or less; beyond this error rate results were mixed. Adopting best-link grouping methods can help leverage the high quality of existing linked datasets to produce better linkage results with new datasets.

There are other opportunities to leverage the quality of data during the matching process. As previously stated, the matching of high-quality data tends to produce high-quality linkage. Also, best-link grouping methods produce superior quality when linking to high-quality datasets. Therefore, splitting datasets into good quality and poor quality data to process separately should produce better results (if good quality data is linked first).

Identifying what constitutes good quality data can be challenging. It is far easier to identify poor quality data and classify everything else as 'good'. Missing values, high-frequency values and placeholders are three data attributes that can be used to identify poor quality records. While this may be relatively easy for unencrypted data and for field-level Bloom filters, some encoding techniques may require these records to be identified at the source, during the encoding process.

Exclusion of missing values and high-frequency values from the good quality data ensures a reliable estimation of parameters for probabilistic linkages. While the poor quality data typically represents a small portion of the entire dataset, a modified matching strategy may be required to account for missing field data before the best-link grouping algorithm can be run.

2.3 Conclusion

Linkage accuracy is paramount in data linkage and is particularly important for operational use of PPRL as current post-linkage practices to improve accuracy can not be used. There are linkage quality techniques applicable to PPRL that will help maximise the accuracy of the

linkage, minimising the need for manual clerical review and providing a foundation for reliable operational use.

2.4 Published manuscript(s)

2.4.1 Estimating parameters for probabilistic linkage of privacy-preserved datasets

Brown AP, Randall SM, Ferrante AM, Semmens JB, Boyd JH (2017). *Estimating parameters for probabilistic linkage of privacy-preserved datasets* BMC Medical Research Methodology, 17(1), 95.

Brown et al. BMC Medical Research Methodology (2017) 17:95
DOI 10.1186/s12874-017-0370-0

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Estimating parameters for probabilistic linkage of privacy-preserved datasets

Adrian P. Brown*, Sean M. Randall, Anna M. Ferrante, James B. Semmens and James H. Boyd

Abstract

Background: Probabilistic record linkage is a process used to bring together person-based records from within the same dataset (de-duplication) or from disparate datasets using pairwise comparisons and matching probabilities. The linkage strategy and associated match probabilities are often estimated through investigations into data quality and manual inspection. However, as privacy-preserved datasets comprise encrypted data, such methods are not possible. In this paper, we present a method for estimating the probabilities and threshold values for probabilistic privacy-preserved record linkage using Bloom filters.

Methods: Our method was tested through a simulation study using synthetic data, followed by an application using real-world administrative data. Synthetic datasets were generated with error rates from zero to 20% error. Our method was used to estimate parameters (probabilities and thresholds) for de-duplication linkages. Linkage quality was determined by F-measure. Each dataset was privacy-preserved using separate Bloom filters for each field. Match probabilities were estimated using the expectation-maximisation (EM) algorithm on the privacy-preserved data. Threshold cut-off values were determined by an extension to the EM algorithm allowing linkage quality to be estimated for each possible threshold. De-duplication linkages of each privacy-preserved dataset were performed using both estimated and calculated probabilities. Linkage quality using the F-measure at the estimated threshold values was also compared to the highest F-measure. Three large administrative datasets were used to demonstrate the applicability of the probability and threshold estimation technique on real-world data.

Results: Linkage of the synthetic datasets using the estimated probabilities produced an F-measure that was comparable to the F-measure using calculated probabilities, even with up to 20% error. Linkage of the administrative datasets using estimated probabilities produced an F-measure that was higher than the F-measure using calculated probabilities. Further, the threshold estimation yielded results for F-measure that were only slightly below the highest possible for those probabilities.

Conclusions: The method appears highly accurate across a spectrum of datasets with varying degrees of error. As there are few alternatives for parameter estimation, the approach is a major step towards providing a complete operational approach for probabilistic linkage of privacy-preserved datasets.

Keywords: Record linkage, Probabilistic, Privacy, Data quality, Linkage quality

Background

Record linkage is a process that allows us to gather together person-based records that belong to the same individual. In situations where unique identifiers are not available, personally identifying information such as name, date of birth and address are used to link records from one or more data collections. As administrative collections typically capture information for large portions of the population, the linked data allows researchers to answer numerous health questions for the whole population at relatively low cost.

Privacy-preserving record linkage

Legal, administrative and technical issues can prevent the release of name-identified data for record linkage. New methods have emerged that do not require the

* Correspondence: adrian.brown@curtin.edu.au
Centre for Population Health Research, Curtin University, Kent Street, Bentley, Western Australia 6102, Australia

release of personally identifying information by data custodians; rather, data custodians use specific encoding processes to transform personally identifying information into a permanently non-identifiable state (an irreversible 'privacy-preserved' state). These methods are collectively referred to as privacy-preserving record linkage (PPRL). Under a trusted third party linkage model [1], this operation occurs *before* the release of any data to record linkage units. Thus, personally identifying information is not disclosed by the data custodian. These PPRL methods can be used within existing record linkage frameworks, and are subject to some of the same challenges [2].

One of the most promising PPRL techniques to emerge is a method which uses Bloom filters in record linkage [3]. A Bloom filter is a probabilistic data structure originally developed to check set membership that can also be used to approximate the similarity of two sets. The ability to provide similarity comparisons on two sets of data is highly desirable for accurate record linkage.

An evaluation of Bloom filters in large-scale probabilistic record linkage has shown high linkage quality (equal to that achieved with unencrypted linkage) with relatively good efficiency [4]. This evaluation utilised single field Bloom filters as opposed to record-level Bloom filters, where all identifiers are added into a single Bloom filter [5]. One of the outstanding challenges for a practical probabilistic PPRL approach is to accurately estimate parameter settings [4]. Typical methods to estimate parameters involve manually examining small samples of data. In the privacy-preserving case, this data is not available to examine so alternate parameter estimation methods are required.

Probabilistic record linkage

In probabilistic record linkage, individual records are compared on a pairwise basis. This process makes the number of possible comparisons extremely large for all but small data files. To reduce computation overhead, records are usually only compared if they have information in common i.e. they have the same value in a particular field or set of fields. Known as blocking, this method reduces the computational comparison space. Pairs of records in each block are compared and assessed through comparison of the values in each matching field (e.g. first name, surname, address, etc.). As shown in Fig. 1, each field comparison is assigned a field score, the value of which depends on whether the field value agrees or disagrees. These agreement and disagreement scores (weights) are computed separately for each field. All field scores are then summed to produce a final score. If this score is greater than a set threshold value, the record pair is designated a match. The set of fields used in the linkage are chosen based on characteristics such as completeness, consistency and discriminating power within each dataset. The discriminating power is a measure of entropy, indicating how useful an identifier might be in the record linkage process [6, 7].

In the Fellegi-Sunter model of record linkage [8], the agreement and disagreement scores used in field comparisons are based on the calculation of two specific probabilities, called the m-probability and u-probability [8]. The m-probability is the likelihood of two fields matching if the records belong to the same individual. The u-probability is the likelihood of two fields matching if the records do **not** belong to the same individual. These two probabilities are converted into agreement and disagreement weights for each field as follows:

$$Agreement\ Weight = \log\left(\frac{m}{u}\right),$$
$$Disagreement\ Weight = \log\left(\frac{1-m}{1-u}\right)$$

The Fellegi-Sunter model incorporates a simplifying assumption where the chances of agreement or disagreement for one field is independent of the chances of agreement or disagreement for another field [8]. This independence assumption allows us to calculate agreement and disagreement weights for each field separately. Extensions to the Fellegi-Sunter model have been developed for approximate comparisons, allowing the assignment of a partial weight for partial agreement that lies somewhere between agreement and disagreement [9]. While there are many types of approximate comparisons for various types of data, most deal with the distance between two strings [10–12]. To fit these approximate comparisons into a probabilistic model, the distance is converted into a partial weight [13].

Missing values can be problematic in probabilistic record linkage. Comparisons are typically treated in one of three ways: a missing value is assigned the disagreement weight, a zero weight, or a separate weight accounted for explicitly. The last option extends the independence assumption to include probabilities for missing values, altering the calculations for weights. Other approaches involve removing the field from matching or even removing the entire record [10, 14].

Parameter estimation

Several methods have been developed to estimate m- and u-probabilities [15, 16]; in practice, most methods

Threshold	11					
Record 1	Robert	Smith	176B Hill View Tce	Bentley	27/03/1979	
Record 2	Bob	Smith	40 Dunedin St	Mount Hawthorn	27/03/1979	
Agreement Weight	5	10	16	3	12	
Disagreement Weight	-2	-4	-3	-3	-4	
						Total
	-2	10	-3	-3	12	14

Fig. 1 Record comparison example

are based on investigations around data quality and prior knowledge, such as the iterative refinement procedure [17].

Automated methods for deriving m-probabilities, such as through EM (expectation-maximisation) estimation have been devised [16, 18, 19]. The EM algorithm has the potential to provide accurate estimates for m-probabilities, in some cases outperforming the probabilities obtained via the iterative refinement procedure [13]. Other estimation methods do exist, such as an algebraic solution by Fellegi and Sunter [8] and the IMSL routine ZXSSQ (an implementation of the Levenberg-Marquardt algorithm) [20]; however, these are more sensitive to initial parameters and require adjustment functions to keep estimates within bounds [21]. An extensive analysis of parameter estimation techniques for the Fellegi-Sunter model of linkage has been detailed by Herzog et al. [15].

Determination of the appropriate threshold setting above which to accept record-pairs as valid matches typically occur through manual inspection of record-pairs within a range of weight scores [22]. The use of PPRL methods within a probabilistic linkage framework, where only encrypted identifiers are used for linkage, preclude the use of any manual, clerical review and so must rely on the use of alternative, computerised methods to determine the best cut-off values. This ability to correctly estimate parameters is of paramount importance if PPRL techniques are to be practical [4].

In this paper, we present a method for accurately estimating probabilities and an optimal threshold cut-off value that can be applied when using Bloom filters within the Fellegi-Sunter model for record linkage. The work builds on a previous privacy-preserving study, which utilised a probabilistic record linkage framework [4]. In this paper, we evaluate our parameter estimation method in two ways: firstly, in a simulation study using synthetic datasets with varying degrees of error; and secondly, on three large-scale administrative datasets, comparing the resultant linkage quality against the quality achieved using calculated m- and u-probabilities.

Methods

Simulation study using synthetic datasets

A series of synthetic datasets were created for our simulation study. Firstly a single 'master' dataset was created, containing 1 million records, with multiple records belonging to the same individual. This dataset did not contain any missing values, or errors typical of what would be seen in administrative data. Then, a series of new datasets were created by first taking the error-free master dataset, and removing or degrading the quality of particular fields.

The synthetic data was generated using an amended version of the FEBRL data generator [23]. The distribution of duplicate records (how many records pertain to each individual) was based on the distribution found in the Western Australian hospital morbidity data collection. The values found in the master dataset were based on frequency distributions found in the Western Australian population. Each record in the dataset contained first name, middle name, surname, sex, date of birth, address, suburb, and postcode information. Address information was randomly selected from the National Address File, a public dataset containing all valid Western Australian addresses.[1]

Additional 'corrupted' datasets were created by modifying the master dataset with a set level of error. In the 1% error file, 1% of field values to be used for linkage were randomly selected to have their values set to missing; a further 1% were randomly selected to have their values corrupted, through the use of typographical errors, misspellings, truncation and replacement of values. In this way, each record could potentially have multiple fields set to missing or corrupted. The same procedure was used to generate a 5% error file, 10% error file and 20% error file. A privacy-preserved version of each dataset was created, using single field Bloom filters.

Testing using administrative datasets

Three datasets comprising real administrative data (hospital admissions records from New South Wales (NSW), Western Australia (WA) and South Australia (SA)) were used to demonstrate the applicability of

the method to real-world data. These datasets have previously been de-duplicated to a very high standard using full identifiers. The results of those de-duplication linkages are used in this study and act as our 'truth set'. The information in this 'truth set' was not used during the linkage process or the estimation of parameters, but was used only as a standard by which to evaluate our results. This data was made available as part of the Population Health Research Network Proof of Concept 1 project [24].

Privacy-preserved versions of each administrative dataset were created, using single field Bloom filters, in the same way as the synthetic datasets. Due to the size of these administrative datasets, five samples (a random 10%) of each privacy-preserved dataset were created; probabilities are estimated for each sample. A de-duplication linkage was performed on each sample and also against the full dataset. The resulting quality was calculated using the 'truth set'.

Application of Bloom filters

The privacy-preserved versions of the synthetic and administrative datasets were created using Bloom filters. Bloom filters were constructed in line with previous work [3]. An empty (or missing) field in the original datasets was left as empty in the privacy-preserved versions.

Matching strategies used for the datasets were based on the strategies used in a published evaluation of linkage software [25]. Two blocking strategies were used; last name Soundex with first name initial, and date of birth with sex. The matching identifiers included Bloom filters for names, address and suburb, using the Sørensen-Dice coefficient comparison for similarity [3]. Sørensen-Dice coefficient values are converted to partial agreement values using a piecewise linear curve, created using Winkler's [13] method. All other fields, including blocking variables, which are created at the same time as the Bloom filters, used exact matches on cryptographically hashed values. Missing value comparisons were assigned a zero weight.

Measuring linkage quality

In line with earlier work [3, 26], we used precision, recall and F-measure as our linkage quality metrics. Precision (also known as positive predictive value) measures the proportion of true positive pairs (correct matches) found from all *classified* matches. Recall (also known as sensitivity) measures the proportion of true positive pairs found from all *true* matches. Both precision and recall return a score between 0 and 1, with higher scores indicating less false positives and false negatives (missed matches) respectively. The F-measure is the harmonic mean between precision and recall, providing a single figure with which we can compare results. Typically, a middle-ground is sought between precision and recall, as there is a trade-off between these values. As the probabilistic linkage threshold is increased, the number of false positives decreases (and so precision increases); however, the number of correct matches missed will also increase, leading to a decrease in recall.

The calculations for these metrics are provided below.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Estimating m and u probabilities

The EM algorithm has been used to calculate the m-probabilities ($\boldsymbol{m}$), u-probabilities ($\boldsymbol{u}$) and the proportion ($\boldsymbol{p}$) of record pairs that match in probabilistic linkage [21]. It is an iterative algorithm that uses the output values of one iteration as the input to the next. We added two additional variables to the EM algorithm as described by Jaro [21], the *missing m-probability* and *missing u-probability* values (denoted by $\boldsymbol{m_m}$ and $\boldsymbol{u_m}$ respectively), to more accurately estimate a single threshold cut-off value (discussed later).

Jaro [21] suggests the algorithm is not particularly sensitive to the starting values ($\boldsymbol{m}$, $\boldsymbol{u}$, $\boldsymbol{m_m}$, $\boldsymbol{u_m}$,$\boldsymbol{p}$). However, the starting values for $\boldsymbol{m}$ should be higher than those for $\boldsymbol{u}$. We thus set an initial value of 0.1 for $\boldsymbol{m_m}$ and $\boldsymbol{u_m}$, 0.8 for $\boldsymbol{m}$ and 0.1 for $\boldsymbol{u}$.

Given two files, A and B, we began by iterating through all possible combinations of field comparisons between A and B. The count of each field state combination was tabulated (an example is shown in Table 1). There are, at most, 3^n possible field state combinations for n fields, assuming each field either agrees, disagrees or is missing. The 'missing' state occurs when a pairwise comparison involves a missing or empty value.

The first part of the EM algorithm is the expectation step. For each field state combination, we calculate *recall* and *false positive rate (fpr)*. For *recall*, each agreement in the table is replaced with $\boldsymbol{m}$, each

Table 1 Field state combinations

First Name	Last Name	Sex	Year of Birth	Count
Agree	Agree	Agree	Agree	1502
Agree	Agree	Missing	Disagree	2142
Agree	Disagree	Disagree	Missing	28,644
…	…	…	…	…

disagreement with ($1 - \boldsymbol{m_m} - \boldsymbol{m}$), and each missing with $\boldsymbol{m_m}$. The product of these is the *recall* for that field state combination. Similarly, for the *fpr*, each agreement in the table is replaced with $\boldsymbol{u}$, each disagreement with ($1 - \boldsymbol{u_m} - \boldsymbol{u}$) and each missing with $\boldsymbol{u_m}$. The product of these provides the *fpr*.

The *recall* and *fpr* allow us to calculate the proportion of true matches for each field state combination *j*:

$$p_j = \frac{p{\cdot}recall_j}{\left(p{\cdot}recall_j\right) + \left((1-p){\cdot}fpr_j\right)}$$

The maximisation step involves the calculation of $\boldsymbol{m}$, $\boldsymbol{u}$, $\boldsymbol{m_m}$, $\boldsymbol{u_m}$ and $\boldsymbol{p}$. The $\boldsymbol{m}$ value for each field is calculated as the ratio of true matches that 'agree' for that field to the total true matches. Likewise, the $\boldsymbol{u}$ value for each field is calculated as the ratio of false matches that 'agree' for that field to the total false matches. The $\boldsymbol{m_m}$ and $\boldsymbol{u_m}$ values use the ratio of matches that are 'missing'.

The output values of ($\boldsymbol{m}$, $\boldsymbol{u}$, $\boldsymbol{m_m}$, $\boldsymbol{u_m}$,$\boldsymbol{p}$) are then used as the input into the next iteration. Iterations are run until values converge. Convergence will occur when the output values differ only minimally from the input values.

Determining a threshold/cut-off setting

In addition to estimating probabilities for a probabilistic linkage, it is important to specify a threshold value that provides optimal resultant linkage quality.

Using the information generated during the EM step, we can estimate the quality of linkage for every combination of weights between a range of possible threshold values (i.e. using precision, recall and F-measure). However, the table of field state combinations used for the EM step only contains field state combinations that were present in the datasets A and B. The *full* set of possible combinations is required to calculate a suitable threshold setting. Field state combinations that are not present in the field state combination table were added with a count of zero, and *recall* and *fpr* were calculated.

Using the full field state combination set, we calculated the weight for each field state combination. Each agreement entry in the table was replaced with the corresponding agreement weight for that field using $\boldsymbol{m}$ and $\boldsymbol{u}$ calculated by the EM algorithm. Likewise, each disagreement entry was replaced with the disagreement weight for that field using the same $\boldsymbol{m}$ and $\boldsymbol{u}$. Each 'missing' entry was replaced with a weight of zero.

To estimate precision, recall and F-measure, we calculated the *True Positives* and *False Positives* for every field state combination. For these estimations, we required the total *True Matches* (true positives and false negatives) and *False Matches* (true negatives and false positives). The total *True Matches* was estimated as part of the EM algorithm, and thus we used the value calculated in the final iteration of the maximisation step. The total *False Matches* was re-estimated as the *total comparison space* less the *True Matches.*

For a single file de-duplication, the total comparison space is:

$$\text{total comparisons} = \left(\frac{\text{RecordCount} \times (\text{RecordCount}-1)}{2}\right)$$

To calculate the *True Positives* and *False Positives*, we multiplied the *recall* and *false positive rate* for each field state combination by the total *True Matches* and *False Matches* respectively.

$$True\ Positives_j = True\ Matches{\cdot}recall_j$$

$$False\ Positives_j = False\ Matches{\cdot}fpr_j$$

We calculated the *True Positives* and *False Positives* for each field state combination so that *precision* could be estimated. To calculate the *precision* for a particular threshold, each field state combination with a weight above that threshold value had their *True Positives* and *False Positives* summed before *precision* was estimated.

We did not calculate *False Negatives,* as this can be derived from the total *True Matches* (*True Positives* plus *False Negatives*) value calculated earlier to estimate *recall*. To calculate *recall* for a particular threshold, the *True Positives* were summed from values for each field state combination that have a weight above that threshold.

As the computation requirements for calculating precision, recall and F-measure are relatively low; we calculated these for all possible weight combinations. With a list of threshold values and corresponding *precision*, *recall* and *F-measure* values, we were able to determine an optimal threshold value for each linkage (i.e. the single threshold score with the highest estimated *F-measure*).

Evaluation of parameter and threshold estimation

For each version of the synthetic datasets, and additionally, for the administrative datasets, probabilities for $\boldsymbol{m}$ and $\boldsymbol{u}$ were estimated together with a threshold cut-off value. The EM algorithm was used to estimate $\boldsymbol{m}$ only for each de-duplication linkage. The frequencies used for our EM algorithm were calculated on blocks, and as such, the number of non-matches observed was greatly reduced, thereby introducing an undesirable bias into the EM algorithm's $\boldsymbol{u}$ estimates [21]. Consequently, we elected to use Jaro's u-probability estimate (on

unblocked data) $\boldsymbol{u}$, together with the EM algorithm's estimated $\boldsymbol{m}$ value.

As part of our simulation study, a de-duplication linkage was run on each synthetic dataset using this combination of values, and a linkage was also run using calculated m- and u- probabilities. Optimal threshold values were estimated for both sets of probabilities. The highest F-measure and estimated threshold F-measure were recorded and compared for all synthetic dataset de-duplication linkages. Similarly, in our test using real data, de-duplication linkages were run on the administrative data; calculated m- and u-probabilities were obtained using the administrative data 'truth sets'. The accuracy of the probability estimates on the administrative dataset samples was measured using the root-mean-square error (RMSE), comparing the F-measure obtained from the EM algorithm probabilities with that obtained from calculated probabilities. RMSE was also used to compare the F-measure obtained at the estimated threshold with that which would be obtained at the correctly chosen threshold. The formula used was as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{Dataset\ 1}^{Dataset\ n}\left(Fmeasure_{estimated} - Fmeasure_{actual}\right)^2}$$

Results

Synthetic data

The characteristics of the synthetic datasets are shown in Table 2. As the dataset error rates increase, the number of unique values for each field increases significantly because of the corruption introduced during dataset creation. The discriminating power for each field also increases with the simulated data corruption.

The results from de-duplication linkages of the synthetic datasets using calculated probabilities and EM probabilities are shown in Table 3. These results show that the use of EM for probability estimation, combined with our threshold estimation technique, provided linkage quality comparable to the best achievable using calculated probabilities, on data with up to 20% error.

As one would expect, de-duplication of the master dataset (without error) produced a perfect result with F-measure of 1.0 at a threshold of 49 (the sum of all agreement weights for each field). The use of EM estimated m-probabilities produced the same result. However, estimation of a threshold value for the master dataset was significantly lower, with a value of 8 for both calculated and estimated probabilities. Note, however, that although this threshold estimate is much lower, it results in just 60 false positives from the entire comparison space, giving an F-measure of 0.9999995.

While it is possible for the threshold to be estimated to one or two decimal places, the use of a whole number here was made for simplicity. It is possible that a better estimate could be made with a finer precision but the differences between thresholds shown here using whole numbers are already negligible.

As Table 3 shows, using our estimation technique, there is a slight decrease in linkage quality as error rates in the data increase (i.e. for 1% error, an F-measure of 0.9979 vs. 0.9979, compared to 20% error with an F-measure of 0.5217 vs. 0.4917). However, even at 10% error, the difference is very small with an F-measure of 0.8443 vs. 0.8436.

Administrative data

The characteristics of the fields in each administrative dataset, such as the number of unique values, missing

Table 2 Synthetic dataset characteristics

Field	0% Error		1% Error		5% Error		10% Error		20% Error	
	Unique Values	Discriminating Power	Unique Values	Discriminating Power	Unique Values	Discriminating Power	Unique Values	Discriminating Power	Unique Values	Discriminating Power
First Name	31,183	8.91	34,595	8.92	45,914	8.99	58,046	9.08	78,256	9.29
Middle Name	25,002	7.33	28,224	7.35	38,285	7.45	48,973	7.59	67,160	7.95
Last Name	56,507	10.87	61,198	10.88	77,088	10.96	94,925	11.07	125,483	11.35
Dob Year	112	6.49	114	6.49	116	6.50	117	6.51	119	6.53
Dob Month	12	3.58	12	3.58	12	3.58	12	3.58	12	3.58
Dob Day	31	4.94	31	4.94	31	4.94	31	4.94	31	4.93
Sex	2	1.00	2	1.00	2	1.00	2	1.00	2	1.00
Address	171,088	12.89	178,583	12.92	207,909	13.04	241,966	13.21	304,353	13.66
Suburb	1962	8.33	7390	8.36	19,664	8.48	31,054	8.65	49,929	9.10
Postcode	379	6.77	1755	6.80	2579	6.91	2981	7.06	3395	7.45

Table 3 Synthetic dataset linkage quality - estimated vs. calculated

Data Error Rate	Calculated Probabilities				EM m-probs and Estimated u-probs			
	Highest		Estimated		Highest		Estimated	
	Threshold	FMeasure	Threshold	FMeasure	Threshold	FMeasure	Threshold	FMeasure
0%	49	1.0000	8	0.9999	49	1.0000	8	0.9999
1%	9	0.9979	16	0.9978	13	0.9979	11	0.9979
5%	8	0.9549	16	0.9541	12	0.9549	11	0.9549
10%	8	0.8443	16	0.8399	12	0.8439	11	0.8436
20%	8	0.5217	16	0.4938	12	0.4999	11	0.4917

percentage, and discriminating power were recorded, shown in Table 4. The random samples generated for each administrative dataset were highly representative of the full dataset.

Linkage quality from EM estimates

The estimated m- and u-probabilities of the samples reflect the characteristics described above, with negligible differences observed between the samples for each dataset. The estimated probabilities for each dataset are shown in Table 5.

Comparisons of linkages using the calculated probabilities and the EM m-probabilities with estimated u-probabilities are shown in Table 6. The highest F-measure obtained using the estimated probabilities was slightly higher than that achieved using calculated probabilities in all cases.

Accuracy of threshold estimation

The quality of linkage using the F-measure at the estimated threshold is compared to the highest F-measure for each sample, as shown in Table 7. The RMSE values for each dataset were 0.0019 for NSW, 0.0001 for SA and 0.0046 for WA. The estimated threshold value was slightly below the best threshold for each dataset.

Discussion

In our simulation study, the use of the EM algorithm to estimate probabilities for a de-duplication linkage produced results comparable to those produced by calculated probabilities, even with synthetic datasets that contained 20% introduced error. Similarly, in our tests using administrative datasets, the probability and threshold estimation technique produced very high-quality linkage results. In comparison to the quality of linkage using calculated probabilities, the probabilities used from the EM algorithm produced linkage quality of the simulation datasets that was comparable to the best possible. However, we found better quality results using estimated probabilities on the real administrative datasets, at least in regards to F-measure. This is a somewhat surprising result, and why this occurred for all three administrative datasets is not entirely clear. A recent analysis of the popular F-measure metric suggests that it may not provide a fair comparison between linkage methods if the selected thresholds produce a different number of predicted matches [27]. This behaviour is one possible explanation for our results, and future work will consider additional metrics for measuring linkage quality. It should be noted that the differences between the linkage quality results were relatively small, and we

Table 4 Administrative dataset characteristics

Field	NSW(13,534,177 records)			SA(2,509,914 records)			WA(6,772,949 records)		
	Unique Values	Missing %	Discriminating Power	Unique Values	Missing %	Discriminating Power	Unique Values	Missing %	Discriminating Power
First Name	168,766	2.9%	8.61	124,849	5.5%	9.18	78,992	0.3%	8.54
Middle Name	114,686	54.2%	6.96	22,180	75.4%	7.19	61,241	40.8%	7.13
Last Name	291,595	0%	10.92	81,431	5.3%	10.81	123,481	0%	10.73
Dob Year	123	0%	6.47	115	0%	6.45	118	0%	6.39
Dob Month	12	0%	3.58	12	0%	3.58	12	0%	3.58
Dob Day	31	0%	4.94	31	0%	4.94	31	0%	4.94
Sex	2	0%	1.00	2	0%	1.00	2	0%	0.99
Address	3,084,889	1.5%	16.96	690,615	8.1%	14.92	1,350,796	0.2%	16.05
Suburb	49,843	0.5%	9.30	10,729	6.9%	7.85	5542	0.1%	7.73
Postcode	3947	0.8%	8.17	2238	8.5%	6.90	2319	0.2%	6.58

Table 5 Estimated probabilities

	NSW		SA		WA	
Field	EM m-prob	Est. u-prob	EM m-prob	Est. u-prob	EM m-prob	Est. u-prob
First Name	0.9817	0.0024	0.8707	0.0015	0.9732	0.0027
Middle Name	0.4686	0.0017	0.1846	0.0004	0.4385	0.0025
Last Name	0.9916	0.0005	0.8931	0.0005	0.9823	0.0006
Dob Year	0.9973	0.0113	0.9997	0.0114	0.9935	0.0119
Dob Month	0.9987	0.0834	0.9988	0.0834	0.9949	0.0835
Dob Day	0.9965	0.0325	0.9988	0.0325	0.9963	0.0326
Sex	0.9999	0.5008	1.0000	0.5010	0.9998	0.5018
Address	0.8325	7.99E-06	0.6486	2.8E-05	0.7338	1.7E-05
Suburb	0.9303	0.0016	0.7462	0.0038	0.8402	0.0047
Postcode	0.9540	0.0034	0.7574	0.0070	0.8640	0.0104

would not expect this to be the case for datasets of all sizes and quality.

The original unencrypted versions of these datasets had previously been linked by Boyd et al. using probabilities estimated with knowledge of previous linkages and refinement through pilot linkages [24]. The probabilities derived from the EM algorithm produced a higher F-measure for both the NSW (0.996 vs. 0.995) and WA (0.992 vs. 0.990) Bloom filter datasets; data for the unencrypted SA dataset was unavailable. On face value, at least, these results indicate that use of the EM algorithm for probability estimation is a viable option, especially where sampling techniques for estimation are not available due to the privacy-preserved nature of the data.

Our study found that the m-probabilities estimated via the EM algorithm did not necessarily match the calculated m-probabilities for each field; however, there was a general consistency of the m-probabilities across all fields. Both our synthetic datasets and the administrative datasets contained many matches and were thus good candidates for probabilities estimated through the EM algorithm. The EM algorithm is known to perform poorly with datasets that have too few matches [15]. Being able to identify and address this issue for privacy-preserved data will require further research.

Our threshold estimation technique also returned very good linkage quality, with a resulting F-measure that consistently approached the best F-measure achievable given the probabilities used. To our knowledge, no alternative method of estimating thresholds exists for use with privacy-preserved data. Without the ability to provide any manual review post-linkage, it is important to be able to estimate a single accurate threshold cut-off value. As such, this technique should be considered for use with Bloom filters for probabilistic linkage.

The threshold values estimated in our study were consistently higher than the optimum threshold when using the calculated probabilities, with fewer false positives and more false negatives returned in each of the linkages (with the exception of the 'perfect' synthetic dataset). Interestingly, we found the opposite to be true when using the estimated probabilities, with a consistently lower threshold. Additional simulation studies may help to understand this effect and improve the estimation accuracy. This effect may be a result of the blocking technique used to gather field state combinations and the similarities in the estimation methods for both probabilities and threshold. Although it may be possible to adjust for this underestimation, an advantage of using a lower threshold is that alternative approaches can be implemented which specifically target false positive matches. It may be possible to run automated clerical review procedures on the results, such as graph theory techniques, to find and correct false positive errors [28]. The effectiveness of these techniques on privacy-preserved data is unknown, however.

Table 6 Linkage quality (max F-measure) – EM vs. calculated

Dataset	Probabilities	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	RMSE
NSW	Calculated	0.9941	0.9943	0.9942	0.9941	0.9940	
	EM	0.9961	0.9965	0.9963	0.9963	0.9961	0.0021
SA	Calculated	0.9532	0.9521	0.9529	0.9553	0.9532	
	EM	0.9590	0.9567	0.9563	0.9582	0.9589	0.0046
WA	Calculated	0.9907	0.9904	0.9910	0.9905	0.9906	
	EM	0.9920	0.9916	0.9921	0.9917	0.9918	0.0012

Table 7 Linkage quality – max F-measure vs. F-measure at threshold estimate

Dataset	Threshold		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	RMSE
NSW	Best	14	0.9961	0.9965	0.9963	0.9963	0.9961	
	Estimated	12	0.9943	0.9946	0.9945	0.9944	0.9942	0.0019
SA	Best	13	0.9590	0.9567	0.9563	0.9582	0.9589	
	Estimated	12	0.9589	0.9566	0.9563	0.9581	0.9588	0.0001
WA	Best	13	0.9920	0.9916	0.9921	0.9917	0.9918	
	Estimated	11	0.9871	0.9870	0.9873	0.9871	0.9875	0.0046

Future research will examine the use of the EM algorithm on composite Bloom filters. While single field Bloom filters provide excellent quality with probabilistic linkage, they may not provide a sufficient level of privacy for some stakeholders. As such, the use of composite Bloom filters may be necessary. Row-level Bloom filters would not be viable; at least two fields are required for probabilistic record linkage. However, multiple Bloom filters comprising two or three fields may function sufficiently. The use of the EM algorithm and the threshold estimation technique on Bloom filters comprising two or more fields is untested, and more research into the performance of the EM algorithm on data containing composite fields is warranted.

Finally, it is worth noting that the EM algorithm and threshold estimation technique described in this paper have wider application and could be used for any probabilistic linkage (encrypted and unencrypted), not just Bloom filters for PPRL. Provided the datasets being linked have sufficient matches, the estimation technique will produce optimal m-probabilities and a suitable threshold cut-off for the linkage. The u-probabilities can be estimated using Jaro's estimation method. Unencrypted linkages would benefit from this technique as well, providing a strong empirical foundation from which to build a robust linkage strategy.

Conclusions

Previous evaluations have shown that privacy-preserving record linkage can be as accurate as traditional unencoded linkage. An important element in developing a practical probabilistic privacy-preserving approach is to determine how to appropriately set parameters without recourse to manual inspection or prior knowledge of data. As we have shown, use of the EM algorithm and our threshold estimation technique provides a robust method of estimating parameters for probabilistic linkage of Bloom filter datasets. This method appears highly accurate on datasets with varying error levels. Further testing is required on real-world datasets with poorer quality data and on datasets with fewer potential matches. The ability for these techniques to produce consistently accurate results on a variety of data will determine whether they are viable in an operational setting.

Endnotes

[1]Available from https://data.gov.au/dataset/geocoded-national-address-file-g-naf

Abbreviations

EM: Expectation-maximisation; FPR: False positive rate; NSW: New South Wales; PPRL: Privacy-preserving record linkage; RMSE: Root mean square error; SA: South Australia; WA: Western Australia

Acknowledgements

The project acknowledges the support of data custodians and data linkage units who provided access to the jurisdictional data.

Funding

Data for the project was provided as part of a Population Health Research Network (PHRN) 'Proof of Concept' collaboration which included the development and testing of linkage methodologies. The PHRN is supported by the Australian Government National Collaborative Research Infrastructure Strategy and Super Science Initiatives. AB has also been supported by an Australian Government Research Training Program Scholarship.

Availability of data and materials

The data that support the findings of this study are available from state data linkage units in NSW, SA and WA, but restrictions apply to the availability of these data, which were used under agreement with data custodians, and so are not publicly available.

Authors' contributions

AB, SR and JB designed the study. AB performed the evaluation and analysed the data. AB and SR wrote the first draft of the manuscript. SR, AF, JS and JB critically reviewed the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Ethical approval for developing and refining linkage methodology, which includes the parameter estimates for probabilistic linkage of privacy-preserved datasets, was obtained from Curtin University Human Research Ethics Committee (Reference: HR 15/2010) as well as approval from South Australia Department of Health and Ageing Human Research Ethics Committee (Reference: HREC 511/03/2015), New South Wales Cancer Institute Human Research Ethics Committee (HREC/10/CIPHS/37) and Western Australian Department of Health Human Research Ethics Committee (HREC/2009/54). Ethics approval included a waiver of consent based on the criteria in the national statement on ethical conduct in human research.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 19 December 2016 Accepted: 23 June 2017
Published online: 10 July 2017

References

1. Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. Inf Syst. 2013;38(6):946–69.
2. Brown AP, Ferrante AM, Randall SM, Boyd JH, Semmens JB. Ensuring privacy when integrating patient-based datasets: new methods and developments in record linkage. Front Pub Health. 2017;5:34.
3. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Making. 2009;9(1):41.
4. Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. J Biomed Inform. 2014;50:205–12.
5. Schnell R, Bachteler T, Reiher J. A Novel Error-Tolerant Anonymous Linking Code. In: Working Paper Series No WP-GRLC-2011-02. Nürnberg: German Record Linkage Center; 2011.
6. Basharin GP. On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables. Theory Probab Applic. 1959;4:333–6.
7. Wajda A, Roos LL. Simplifying Record Linkage: Software and Strategy. Comput Biol Med. 1987;17(4):239–48.
8. Fellegi I, Sunter A. A Theory for Record Linkage. J Am Stat Assoc. 1969;64:1183–210.
9. DuVall SL, Kerber RA, Thomas A. Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators. J Biomed Inform. 2010;43:24–30.
10. Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Berlin/Heidelberg: Springer Science & Business Media; 2012.
11. Winkler WE. Preprocessing of lists and string comparison. Rec Linkage Tech. 1985;985:181–7.
12. Thibaudeau Y. Fitting log-linear models when some dichotomous variables are unobservable. In: Proceedings of the Section on statistical computing: 1989; 1989. p. 283–8.
13. Winkler WE. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. Paper presented at the Annual ASA Meeting in Anaheim. Washington: Statistical Research Division, U.S. Bureau of the Census; 1990.
14. Ong TC, Mannino MV, Schilling LM, Kahn MG. Improving record linkage performance in the presence of missing linkage data. J Biomed Inform. 2014;52:43–54.
15. Herzog TN, Scheuren FJ, Winkler WE: Data quality and record linkage techniques. Springer Science & Business Media. 2007.
16. Winkler WE. Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. In: Proceedings of the Section on Survey Research Methods, American Statistical Association: 1988; 1988. p. 671.
17. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic Linkage of Vital Records. Science. 1959:954–9.
18. Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. Am Med Infom Assoc. 2003:259–63.
19. Bauman G John Jr: Computation of Weights for Probabilistic Record Linkage using the EM Algorithm. (Masters Thesis). Available from All Theses and Disserations (Paper 746): Brigham Young University; August 2006.
20. Inc IMaSL. User's manual: IMSL library: problem solving software system for mathematical and statistical FORTRAN programming, Ed. 9.2, rev edn: IMSL; 1984.
21. Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J Am Stat Assoc. 1989;84(406):414–20.
22. Gill L: Methods for automatic record matching and linkage and their use in national statistics. In: National Statistics Methodological Series No 25. Office for National Statistics. 2001.
23. Christen P, Pudjijono A. Accurate synthetic generation of realistic personal information. Adv Knowl Discov Data Min. 2009;5476:507–14.
24. Boyd JH, Randall SM, Ferrante AM, Bauer JK, McInneny K, Brown AP, Spilsbury K, Gillies M, Semmens JB. Accuracy and completeness of patient pathways–the benefits of national data linkage in Australia. BMC Health Serv Res. 2015;15(1):312.
25. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. J Biomed Inform. 2012;45(1):165–72.
26. Randall S, Ferrante A, Boyd J, Semmens J. The effect of data cleaning on data linkage quality. BMC Med Inform Decis Making. 2013;13(64):e1.
27. Hand D, Christen P. A note on using the F-measure for evaluating record linkage algorithms. Stat Comput. 2017:1–9.
28. Randall SM, Boyd JH, Ferrante AM, Bauer JK, Semmens JB. Use of graph theory measures to identify errors in record linkage. Comput Methods Prog Biomed. 2014;115(2):55–63.

2.4.2 Evaluation of approximate comparison methods on Bloom filters for probabilistic linkage

Brown AP, Randall, SM, Boyd, JH, Ferrante, AM (2019). *Evaluation of approximate comparison methods on Bloom filters for probabilistic linkage.* International Journal of Population Data Science, 4(1). https://doi.org/10.23889/ijpds.v4i1.1095

International Journal of Population Data Science (2018) 4:1:16

International Journal of Population Data Science

Journal Website: www.ijpds.org

Evaluation of approximate comparison methods on Bloom filters for probabilistic linkage

Brown, AP[1*], Randall, SM[1], Boyd, JH[1], and Ferrante, AM[1]

Submission History

Submitted:	26/10/2018
Accepted:	05/03/2019
Published:	23/05/2019

[1]Centre for Data Linkage, Curtin University, Western Australia, Perth, Australia

Abstract

Introduction
The need for increased privacy protection in data linkage has driven the development of privacy-preserving record linkage (PPRL) techniques. A popular technique using Bloom filters with cryptographic analyses, modifications, and hashing variations to optimise privacy has been the focus of much research in this area. With few applications of Bloom filters within a probabilistic framework, there is limited information on whether approximate matches between Bloom filtered fields can improve linkage quality.

Objectives
In this study, we evaluate the effectiveness of three approximate comparison methods for Bloom filters within the context of the Fellegi-Sunter model of recording linkage: Sørensen–Dice coefficient, Jaccard similarity and Hamming distance.

Methods
Using synthetic datasets with introduced errors to simulate datasets with a range of data quality and a large real-world administrative health dataset, the research estimated partial weight curves for converting similarity scores (for each approximate comparison method) to partial weights at both field and dataset level. Deduplication linkages were run on each dataset using these partial weight curves. This was to compare the resulting quality of the approximate comparison techniques with linkages using simple cut-off similarity values and only exact matching.

Results
Linkages using approximate comparisons produced significantly better quality results than those using exact comparisons only. Field level partial weight curves for a specific dataset produced the best quality results. The Sørensen-Dice coefficient and Jaccard similarity produced the most consistent results across a spectrum of synthetic and real-world datasets.

Conclusion
The use of Bloom filter similarity comparisons for probabilistic record linkage can produce linkage quality results which are comparable to Jaro-Winkler string similarities with unencrypted linkages. Probabilistic linkages using Bloom filters benefit significantly from the use of similarity comparisons, with partial weight curves producing the best results, even when not optimised for that particular dataset.

Introduction

In recent years, record linkage centres have adopted many different models and linkage methods to ensure the protection of individual privacy as part of their operational processes. With growing demand for linked data, it has been critical for record linkage centres to implement methods which protect privacy, yet maximise the benefits that can be derived from data assets. As a result, research around privacy-preserving record linkage (PPRL) methods has become a pressing area of inquiry, with much focus on the use of Bloom filters [1-7]. Much research has focussed on the security aspect of the Bloom filters, such as cryptographic analyses of encoding methods, modifications, and hashing variations [3, 7-12]. The resultant accuracy or 'quality' of these techniques has often been overlooked. To consider for operational use within large-scale linkage systems, accuracy must be sufficiently high [13].

A Bloom filter is a probabilistic data structure that is used to approximate the equality of two sets; these similarity comparisons are extremely useful in record linkage allowing for typographic errors and variations in spelling. Bloom filters are

*Corresponding Author:
Email Address: adrian.brown@curtin.edu.au (AP Brown)

https://doi.org/10.23889/ijpds.v4i1.1095

Brown, AP et. al. / International Journal of Population Data Science (2018) 4:1:16

implemented using an array of bits. Text values are first split into elements (typically bigrams); each element is added to the Bloom filter by applying one or more hash functions to it. The results of these hash functions determine which positions in the bit array are set to one.

Typically, PPRL techniques that use Bloom filters are applied at either the field or record level. Field level Bloom filters encode each identifier into a separate Bloom filter [14]. Record linkage techniques (deterministic and probabilistic) can then be used to link records in much the same way as with unencrypted identifiers [15-17]. Record level (or composite) Bloom filters encode two or more identifiers into a single Bloom filter [5, 18]. Composite Bloom filters may be useful in certain situations where a single linkage field is desirable or even mandated [19], but handling missing values and identifiers that change over time (such as address) remain issues [20].

Probabilistic record linkage is preferred by many data linkage centres due to its proven track record of producing high quality linkage results from unencrypted identifiers [21-23]. It has been shown to produce equally good results when applied to Bloom filters [1, 15, 16]. An extension to the basic probabilistic model of record linkage allows for approximate matches between fields. An approximate match is typically assigned a 'similarity score'. These scores are then converted into partial weights of agreement or partial disagreement weights (as distinct from full agreement or full disagreement [24, 25]). The use of partial agreement linkage models has been shown to greatly improve the linkage quality when compared to the use of exact comparisons [25-28].

There is little mention in the literature of Bloom filters being used in the context of probabilistic record linkage where the field similarity score is converted into a partial agreement weight during the calculation of a pair-wise score [1, 15, 28]. Several issues remain unclear: What is the effect of approximate matching on the linkage quality using Bloom filters? How does this quality vary as the level of error in datasets increases? How do different approximate comparison methods perform in this context? The commonly used approximate comparisons for Bloom filters include the Sørensen–Dice coefficient, Jaccard similarity and Hamming distance [4, 14]. In this paper, we evaluate the effectiveness of each of these comparisons within the approximate comparison extensions to the Fellegi-Sunter model of record linkage [24, 29].

Methods

Data Sources

Synthetic data was created using an amended version of the FEBRL data generator [30]. Datasets included some core identifiers for linkage: first name, middle name, last name, sex, date of birth, address, suburb, and postcode information. The population profile of the individual fields in the master dataset was based on the frequency distributions in the Western Australian population. Western Australian addresses were randomly allocated from records in the National Address File (a public dataset containing validated Australian addresses). An additional four 'corrupted' datasets were created by modifying the master dataset with varying levels of error (1%, 5%, 10% and 20% of fields containing errors, respectively). The number of records allocated to each individual was based on the admission/re-admission patterns found in the Western Australian hospital morbidity data collection.

Within the 'corrupted' datasets, the fields containing errors were restricted to those that typically use a similarity comparison during record linkage (the 'similarity fields'): first name, middle name, surname, address and suburb. The remaining fields were untouched. In the 1% error file, 1% of the designated fields were randomly selected to have their values corrupted, through the use of typographical errors, misspellings, truncation and replacement of values. The same procedure was used to generate a 5% error file, 10% error file and 20% error file.

Real data was also used in our evaluation. An extract from the New South Wales (NSW) Emergency Department Data Collection was used to demonstrate the effectiveness of the partial agreement methods on real-world data [31]. This dataset had previously been deduplicated to a very high standard, using full identifiers, by the Centre for Health Record Linkage (CHeReL) in NSW [32]. The results of these deduplications were used as our benchmark in determining linkage quality.

Application of Bloom filters

Privacy-preserved versions of each dataset were created using field level Bloom filters for the 'similarity fields'. These Bloom filters were constructed using the method first described by Schnell [14]. Fields were truncated to a maximum of twelve characters and split into bigrams that were hashed 40 times into Bloom filters 512 bits in length.

Linkage strategy

As per the Fellegi-Sunter approach, a single block, using the date of birth field value, was applied to reduce the comparison space. This field remained untouched during the corruption process and ensured full pairs completeness for our synthetic dataset linkages. The m- and u- probabilities for each linkage field within the datasets were estimated using known matches within the block. Known matches were identified using our generated key for the synthetic datasets, and the keys provided to us for the NSW administrative dataset (our 'truth sets'). These probabilities were used for all linkages of all datasets.

For the linkage of each dataset, the corresponding m- and u- probabilities were converted into agreement and disagreement weights as follows:

$$\textit{Agreement Weight} = \log\left(\frac{m}{u}\right)$$

$$\textit{Disagreement Weight} = \log\left(\frac{1-m}{1-u}\right)$$

Fields using exact comparisons used either the full agreement weight or the full disagreement weight. Fields using approximate comparisons used a value somewhere between these two weights. A missing field value on either side of the comparison resulted in a weight of zero. The weight values were summed across all fields to determine the total 'score' for each pairwise comparison.

All pairs above a score of zero were recorded. Using the 'truth set' for each dataset, the number of missed matches

Brown, AP et. al. / International Journal of Population Data Science (2018) 4:1:16

(false negatives) and incorrect matches (false positives) were calculated for each possible cut-off value above zero. False negatives and false positives were treated equally, the aim to minimise the sum of these misclassifications. Thus, the cut-off value with the smallest number of misclassifications was used as the best outcome (highest quality) for that linkage. Records were grouped using transitive closure ('merge' based) grouping, with all indirect links being honoured.

Similarity Comparators

For linkages of the privacy-preserved datasets, the Sørensen–Dice coefficient, Jaccard similarity, and Hamming distance comparators were used to compare the similarity between Bloom filtered fields. Sørensen–Dice coefficient and Jaccard similarity scores range from 0 to 1, where higher values represent greater similarity and a score of 1 represents identical values. Similarity is based on the set of bit positions set to one in each Bloom filter. Given two of these sets, A and B, similarities are calculated as follows:

$$S(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Hamming distance measures the difference between values. For Bloom filters, this is the number of bits that are different between each Bloom filter and resulting scores range from 0 to the length of the Bloom filter:

$$H(A,B) = |A \oplus B|$$

The raw Hamming score is normalised by dividing all raw scores by the maximum raw score giving us a value between 0 and 1; lower scores represent greater similarity, and a score of 0 represents identical values.

Modelling partial agreement

The method for modelling partial agreement required the distribution of matches and non-matches at defined similarity scores for each field in each dataset. This was achieved by performing a deduplication linkage on each dataset and recording matches and non-matches for each observed comparison. The study used the steps outlined by Winkler for estimating partial weights at specified similarity values [24]:

1. The similarity score range for all approximate comparison used is $0..1$. This range was partitioned into $i = 1, \ldots, N$ sub-intervals. We used $N = 20$ resulting in sub-intervals at 0.05 increments.

2. For each field j and each sub-interval $(k_i, k_{i+1}]$, the number of matches and non-matches were recorded.

3. For each sub-interval $(k_i, k_{i+1}]$, the match to non-match ratio τ_i was calculated as the probability of a match at interval i divided by the probability of a non-match at that interval:

$$\tau_i = \frac{P(\delta(\gamma^j(a,b)) \in \{(k_i, k_{i+1}]|M\})}{P(\delta(\gamma^j(a,b)) \in \{(k_i, k_{i+1}]|U\})}$$

$$\tau_i = \frac{matches_i / totalmatches}{nonmatches_i / totalnonmatches}$$

$$\tau_i = \frac{m_i}{u_i}$$

Here δ is the comparator function, γ^j is a comparison of the jth field, (a, b) is an arbitrary pair, M is the set of matches, and U the set of non-matches.

4. The ratio vector τ is then used to create the partial weight curve for the complete set of sub-intervals $i = 1, \ldots, N$, applying the normalised ratio vector to the field weight with the disagreement weight at 0 and the agreement weight at 1.

In addition to partial weight curves, we used a simple cut-off value for the field similarity score to determine where the full agreement or full disagreement is applied. Cut-off values between 0.6 and 0.95 (in 0.05 increments) were used for all similarity fields. The cut-off value with a linkage result having the lowest number of misclassified pairs (the sum of false positives and false negatives) was selected.

Measuring linkage quality

We used the number of misclassified pairs as a measure of linkage quality. Baselines were created for each dataset by performing deduplication linkages using exact comparisons only. Deduplication linkages using Bloom filters with each of the approximate comparisons (Sørensen–Dice, Jaccard and Hamming) were then compared to the baseline to measure the difference in linkage quality. Also, deduplication linkages using the Jaro-Winkler string comparison on unencrypted identifiers were undertaken to measure differences in linkage quality arising from the use of Bloom filters (the Jaro-Winkler comparator cannot be used directly on Bloom filtered data).

Results

Synthetic Data

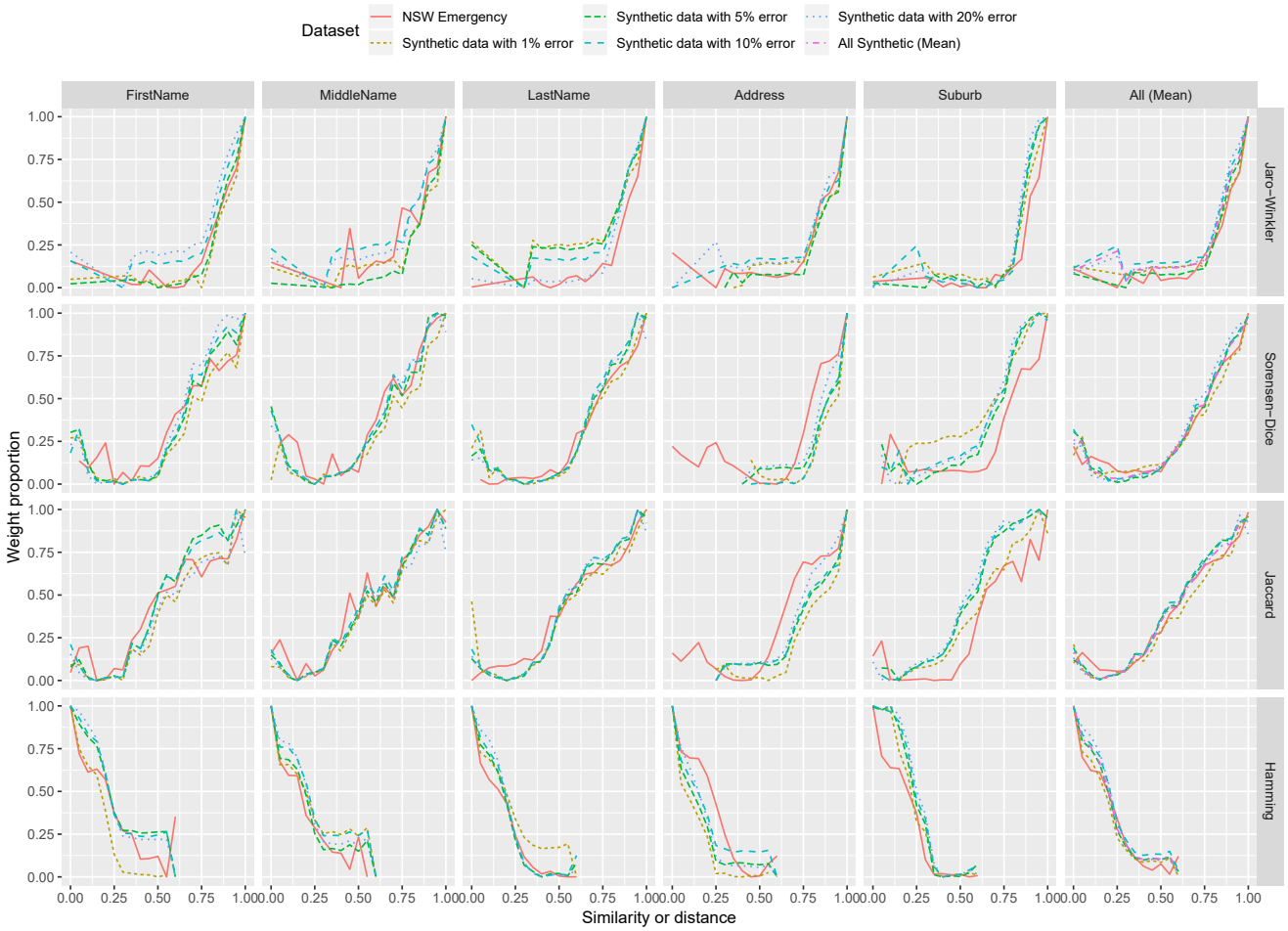
The 'master' dataset of 1 million records contained multiple records belonging to the same individual. From this master dataset, a series of new datasets were created by removing or degrading the quality of particular fields. The partial weight curves were created for each field in each synthetic dataset (shown in Figure 1). Dataset level weight curves were also created as an average of the weight curves of each field; the mean of the weight proportion at each interval is used. Deduplication linkages were then performed on each of the synthetic datasets using the field level weight curves and the dataset level weight curve.

The results of the deduplication linkages for each synthetic dataset are shown in Table 1, including the linkage using 'exact' comparisons with field level weight curves, dataset level weight curves and the simple cut-off value that produced the fewest errors.

The performance of the similarity comparisons, when compared to the 'exact' comparisons, shows only a small reduction in linkage errors with the dataset containing 1% error. The benefit derived from partial agreements in data linkage appears

Brown, AP et. al. / International Journal of Population Data Science (2018) 4:1:16

Figure 1: Estimated field and dataset weight curves



Weight proportion represents the proportion of a field match comparison weight (0 = full disagreement, 1 = full agreement)

Brown, AP et. al. / International Journal of Population Data Science (2018) 4:1:16

minimal when the quality of the data is this high. However, a significant reduction in linkage error can be seen for the datasets containing at least 5% error across almost all similarity comparisons. The reduction in misclassified pairs for the dataset with 20% error, while high, is less than both the datasets with 5% and 10% error.

While the results in Table 1 represent the lowest error achievable for each comparison, the trade-off between precision and recall for each comparison is shown in Figure 2. When compared to exact comparisons, the field level weight curves and the dataset level weight curves for all approximate comparison methods provide a consistent improvement in recall while maintaining a high level of precision. Some instances of the comparison methods with cut-off values also provide an improvement. However, there does not appear to be the same level of consistency across all synthetic datasets.

Evaluation on real data

The extract from the NSW Emergency Department Data Collection contained 4,304,458 records. Empty fields (missing values) were left as empty fields in the privacy-preserving version of the dataset.

Dataset level weight curves were created for the NSW Emergency dataset using the same method used with the synthetic datasets. As with the synthetic datasets, deduplication linkages were undertaken using field level, dataset level and simple cut-off values. Also, a deduplication linkage was performed using the dataset level weight curves derived from the synthetic datasets.

The results of all deduplication linkages on the NSW Emergency dataset are shown in Table 2, including the linkage using 'exact' comparisons. The trade-off between precision and recall is shown in Figure 3.

The field level weight curves produced the best results, followed by both dataset level weight curves and the use of simple cut-off values. Similarly to the synthetic datasets, the field level and dataset level weight curves demonstrate a consistent improvement to recall while maintaining a high level of precision.

Discussion

Our results show that the use of Bloom filter similarity comparisons for probabilistic record linkage can produce linkage quality results comparable to the use of the Jaro-Winkler string similarity on unencrypted identifiers. With synthetic datasets, we found that the highest linkage quality was achieved using Hamming distance, producing fewer linkage errors (on the 10% error and 20% error datasets) than the Jaro-Winkler similarity on unencrypted identifiers. Regardless of the comparator used, all approximate comparisons improved the quality of the linkage, particularly as the level of error in the dataset increased. While the dataset with 20% error did not show the same proportional reduction (%) in misclassified pairs (as compared to datasets with only 5% and 10% error), the total number of misclassified pairs was vastly reduced. This 'dip' in reduction may be an artefact of the artificial error generation within the synthetic datasets, or it may be due to a limit on how much error can be accounted for using partial agreements.

As expected, optimised partial weight curves for each field produced the best quality results. The dataset level weight curves, estimated as a single 'best-fit' curve for all fields, showed a well defined slope for each of the comparators, with only a small increase in the number of linkage errors for both the synthetic datasets and the NSW Emergency data. The synthetic datasets and the NSW Emergency datasets produced similar weight curves, so it was unsurprising the dataset level weight curves created from the synthetic data produced high quality results on the NSW Emergency data. The fact that these results were close suggests that it may be possible to estimate a generic curve (for each comparator) for use in the linkage of various types of data; however, further testing using a variety of real datasets is warranted.

The Sørensen-Dice and Jaccard similarity comparators produced very similar linkage quality results across the range of datasets. The Hamming distance comparison appeared to produce the fewest errors for the synthetic datasets overall; however, its performance against the other comparators on the NSW Emergency data was inconsistent. This may be explained by Hamming's observed improved performance under higher degrees of error with the synthetic datasets. If the NSW Emergency data has a similar error rate to the 1% error dataset, Hamming distance's relative performance may also be similar.

Field level and dataset level weight curves for all approximate comparators demonstrated improvement to recall while maintaining a high level of precision, a highly desirable outcome in many linkage settings. There is still a trade-off between missed matches and incorrect matches, however, and care must be taken in selecting an appropriate cut-off during linkage.

A single cut-off value was shown to perform well in the context of determining agreement or disagreement in probabilistic linkage. The linkage quality using a cut-off value is lower than the linkage quality from an approximate weight curve (at least, for the Bloom filter comparisons), and the precision/recall trade-off is less desirable. However, the reduced level of error from an exact linkage is significant, and there appears to be some level of stability in the cut-off values themselves across our datasets. These results suggest that in the absence of being able to estimate a weight curve for a new dataset, whether it is due to size or complexity or time constraints, the use of a standard cut-off value is a viable alternative.

There were several potential limitations to this study. This work uses previously linked real data as a benchmark. While this linked data is of very high quality, it may not be completely accurate. Bloom filtered comparisons on this particular linked data provided comparable results to Jaro-Winkler comparisons. However, this does not imply that these linkage methods are therefore equivalent in all aspects; specifically, ensuring high linkage quality with privacy preserving methods will always be far more difficult, given the limited ability to provide quality assurance or clerical review. Additionally, the synthetic datasets with introduced (manufactured) errors may not always capture the complexity of real datasets. Testing the performance of the Bloom filter comparisons against other kinds of datasets or 'gold standard' datasets would be a valuable exercise. However, such datasets are not always easy to find [33].

Brown, AP et. al. / International Journal of Population Data Science (2018) 4:1:16

Table 1: Linkage errors for each comparison (synthetic datasets)

	1% Error			5% Error			10% Error			20% Error		
	FP	FN	Total	FP	FN	Total	FP	FN	Total	FP	FN	Total
Exact	395	1,674	2,069	2,442	18,099	20,541	131,199	16,399	147,598	110,763	372,572	483,335
Field Level												
Jaro-Winkler	92	1,781	1,873	881	2,904	3,785	4,641	13,612	18,253	44,503	81,321	125,824
Sørensen–Dice	125	1,713	1,838	1,054	2,517	3,571	2,978	16,736	19,714	40,436	105,024	145,460
Jaccard	99	1,719	1,818	827	2,703	3,530	1,276	20,439	21,715	34,869	109,274	144,143
Hamming	132	1,732	1,864	830	2,691	3,521	5,033	10,526	15,559	39,301	76,619	115,920
Dataset Level												
Jaro-Winkler	74	1,752	1,862	1,034	2,799	3,840	3,427	15,199	17,343	47,449	84,134	135,452
Sørensen–Dice	109	1,742	1,848	1,401	3,652	4,612	3,540	25,521	27,343	53,702	120,408	166,761
Jaccard	83	1,744	1,819	1,205	3,708	4,563	10,691	19,047	28,179	66,948	109,909	169,002
Hamming	72	1,753	1,871	962	2,774	3,848	3,349	13,537	16,762	29,584	101,440	129,008
Cut-off value												
Jaro-Winkler	191	1,798	1,989 (0.85)	2,366	3,447	5,813 (0.90)	5,639	16,815	22,454 (0.85)	120,523	64,166	184,689 (0.85)
Sørensen–Dice	263	1,739	2,002 (0.90)	2,123	4,218	6,341 (0.85)	17,563	25,301	42,864 (0.80)	90,544	109,127	199,671 (0.80)
Jaccard	233	1,756	1,989 (0.80)	1,500	6,035	13,363 (0.75)	7,286	38,324	45,610 (0.70)	142,297	48,699	190,996 (0.70)
Hamming	155	1,806	1,961 (0.15)	1,710	3,677	5,387 (0.15)	6,799	15,687	22,486 (0.20)	25,428	158,455	183,883 (0.20)

FP = false positives, FN = false negatives, Cut-off values are shown in parentheses, Cut-off values are shown in parentheses

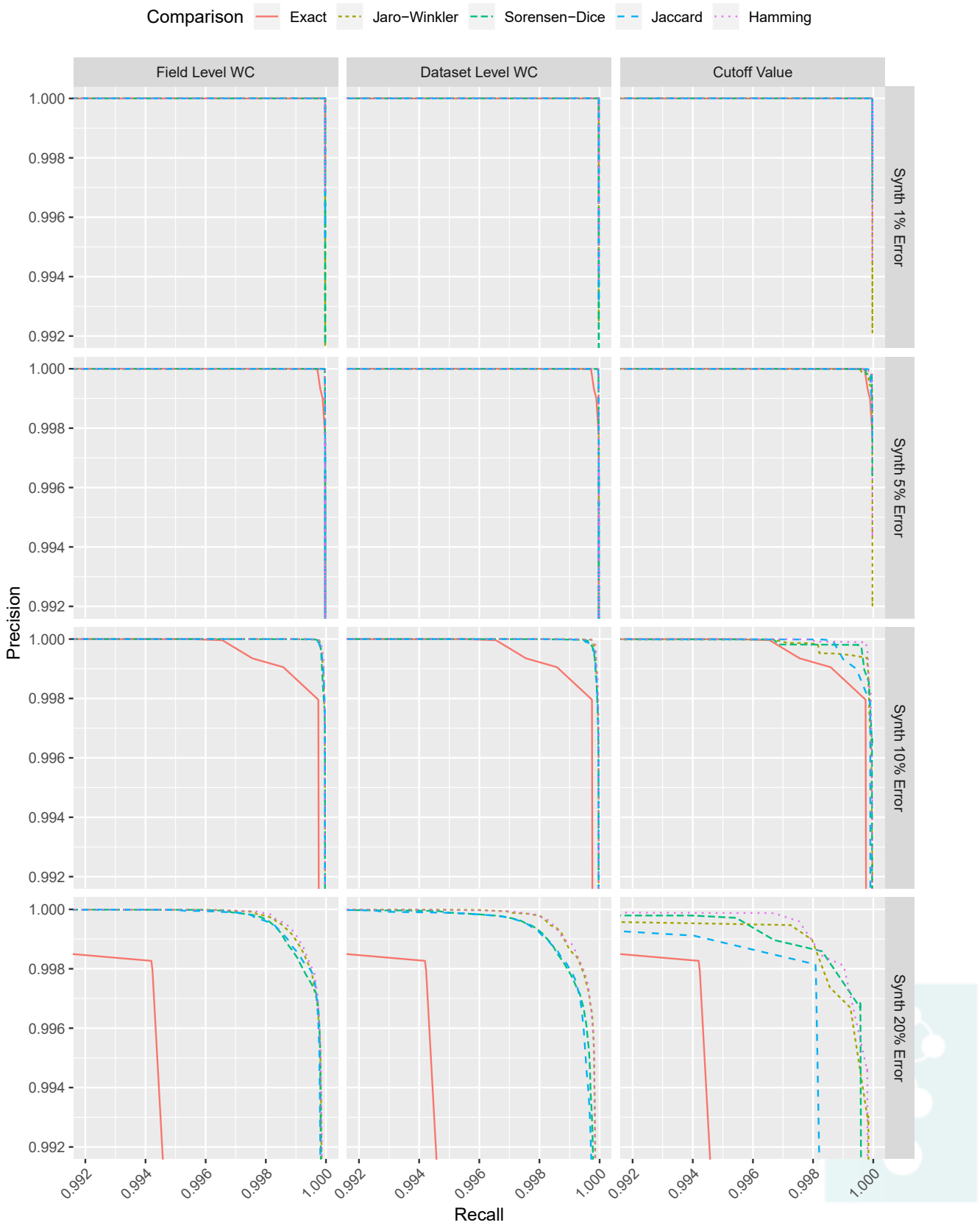
Table 2: Linkage errors for each comparison (NSW Emergency dataset)

	False Positives	False Negatives	Total
Exact	29,040	233,405	262,445
Field Level			
Jaro-Winkler	33,729	170,188	203,917
Sørensen–Dice	34,876	170,801	205,677
Jaccard	44,576	162,931	207,507
Hamming	46,905	166,138	213,043
Dataset Level			
Jaro-Winkler	34,513	170,298	204,811
Sørensen–Dice	41,929	176,513	218,442
Jaccard	35,172	181,066	216,238
Hamming	38,082	170,185	208,267
Cut-off value			
Jaro-Winkler (0.85)	44,038	169,633	213,671
Sørensen–Dice (0.75)	39,848	192,193	232,041
Jaccard (0.65)	42,598	193,117	235,715
Hamming (0.20)	39,750	191,080	230,830
Synthetic Dataset Level			
Jaro-Winkler	31,900	172,363	204,263
Sørensen–Dice	52,073	165,345	217,418
Jaccard	50,586	163,769	214,355
Hamming	40,112	170,642	210,754

Actual cut-off values are shown in parentheses

Brown, AP et. al. / International Journal of Population Data Science (2018) 4:1:16

Figure 2: Precision-recall for each comparison (synthetic datasets)



WC = weight curve

Brown, AP et. al. / International Journal of Population Data Science (2018) 4:1:16

Figure 3: Precision-recall for each comparison (NSW Emergency dataset)

WC = weight curve

Conclusion

Matching quality in probabilistic linkage benefits significantly from the use of similarity comparisons, with partial weight curves producing the best results. We have shown that this remains true even when the weight curve has not been optimised for the particular dataset being linked. This finding also applies to the comparison of Bloom filters within a probabilistic framework. Although determining the partial weight curves for producing optimal linkage quality typically requires the use of a truth set, our results show that adequate quality can be achieved through the use of weight curves derived from simulated datasets.

All similarity comparisons produce significantly better results than 'exact' comparisons. Despite some of the challenges of working with Bloom filters and the range of comparators available, there is not a great difference between these comparators when used within a probabilistic framework. On the basis of these findings, our recommendation to linkage units is to choose the comparator that you are most comfortable with but to use a weight curve estimated for that particular comparator.

Conversion of similarity scores to partial agreement weights is a quality optimisation available for all approximate comparisons (including Bloom filters) and is an essential element to maximising the pair-wise quality with the Fellegi-Sunter model of record linkage. Further work is required to determine how generalisable this option is, by analysing the weight curves with a broader variety of real-world datasets.

Acknowledgements

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy's Population Health Research Network and acknowledges the support of data custodians and data linkage units who provided access to the jurisdictional data. AB has also been supported by an Australian Government Research Training Program Scholarship.

Statement on conflicts of interest

All authors declare there are no conflicts of interest.

Ethics

Ethical approval for developing and refining linkage methodology was obtained from Curtin University Human Research Ethics Committee (HR 15/2010) as well as approval from New South Wales Cancer Institute Human Research Ethics Committee (HREC/10/CIPHS/37). Ethics approval included a waiver of consent based on the criteria in the national statement on ethical conduct in human research.

References

1. Brown AP, Randall SM, Ferrante AM, Semmens JB, Boyd JH. Estimating parameters for probabilistic linkage of privacy-preserved datasets. BMC Medical Research Methodology. 2017;17(1):95. https://doi.org/10.1186/s12874-017-0370-0
2. Pow C, Iron K, Boyd J, Brown A, Thompson S, Chong N, et al. Privacy-Preserving Record Linkage: An international collaboration between Canada, Australia and Wales. International Journal for Population Data Science. 2017;1(1). https://doi.org/10.23889/ijpds.v1i1.101
3. Schnell R, Borgs C. Secure Privacy Preserving Record Linkage of Large Databases by Modified Bloom Filter Encodings. 2016 International Population Data Linkage Conference. 2016. https://doi.org/10.23889/ijpds.v1i1.29

4. Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. Information Systems. 2013;38(6):946-69. http://dx.doi.org/10.1016/j.is.2012.11.005

5. Durham EA, Kantarcioglu M, Member S, Xue Y, Toth C, Kuzu M, et al. Composite Bloom Filters for Secure Record Linkage. IEEE Transactions on Knowledge and Data Engineering. 2014;26:2956-68. https://doi.org/10.1109/TKDE.2013.91

6. Ranbaduge T, Christen P, Vatsalan D. Tree based scalable indexing for multi-party privacy-preserving record linkage. AusDM, CRPIT. 2014;158.

7. Kroll M, Steinmetzer S. Who Is 1011011111...1110110010? Automated Cryptanalysis of Bloom Filter Encryptions of Databases with Several Personal Identifiers. Cham: Springer International Publishing; 2015. p. 341-56. https://doi.org/10.1007/978-3-319-27707-3_21

8. Schnell R, Borgs C. Randomized Response and Balanced Bloom Filters for Privacy Preserving Record Linkage. IEEE International Conference on Data Mining (ICDM'16)2016. https://doi.org/10.1109/ICDMW.2016.0038

9. Schnell R, Borgs C. XOR-Folding for Bloom Filter-based Encryptions for Privacy-preserving Record Linkage. Working Paper WP-GRLC-2016-03, German Record Linkage Center, Nuremberg. 2016.

10. Niedermeyer F, Steinmetzer S, Kroll M, Schnell R. Cryptanalysis of basic Bloom Filters used for Privacy Preserving Record Linkage. 2014. https://doi.org/10.29012/jpc.v6i2.640

11. Kroll M, Steinmetzer S. Automated Cryptanalysis of Bloom Filter Encryptions of Health Records. 2014.

12. Kuzu M, Kantarcioglu M, Durham E, Malin B, editors. A constraint satisfaction cryptanalysis of Bloom filters in private record linkage. Privacy Enhancing Technologies; 2011: Springer. https://doi.org/10.1007/978-3-642-22263-4_13

13. Brown AP, Ferrante AM, Randall SM, Boyd JH, Semmens JB. Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage. Frontiers in Public Health. 2017;5:34. https://doi.org/10.3389/fpubh.2017.00034

14. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Medical Informatics and Decision Making. 2009;9(1):41. https://doi.org/10.1186/1472-6947-9-41

15. Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. Journal of biomedical informatics. 2014;50:205-12. https://doi.org/10.1016/j.jbi.2013.12.003

16. Pinto C, Pita R, Barbosa G, Araújo B, Bertoldo J, Sena S, et al., editors. Probabilistic integration of large Brazilian socioeconomic and clinical databases. Computer-Based Medical Systems (CBMS), 2017 IEEE 30th International Symposium on; 2017: IEEE. https://doi.org/10.1109/CBMS.2017.64

17. Schmidlin K, Clough-Gorr KM, Spoerri A. Privacy Preserving Probabilistic Record Linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. BMC medical research methodology. 2015;15(1):46. https://doi.org/10.1186/s12874-015-0038-6

18. Schnell R, Bachteler T, Reiher J. A Novel Error-Tolerant Anonymous Linking Code. Working Paper Series No WP-GRLC-2011-02 Nürnberg, Germany: German Record Linkage Center. 2011.

19. Bundestag D. Gesetz ueber Krebsregister (Krebsregistergesetz KRG). Bundesgesetzblatt (in German). 1994;79:994.

20. Brown AP, Borgs C, Randall SM, Schnell R. Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets. BMC Medical Informatics and Decision Making. 2017;17(1):83. https://doi.org/10.1186/s12911-017-0478-5

21. Boyd JH, Randall SM, Ferrante AM, Bauer JK, McInneny K, Brown AP, et al. Accuracy and completeness of patient pathways–the benefits of national data linkage in Australia. BMC Health Services Research. 2015;15(1):312. https://doi.org/10.1186/s12913-015-0981-2

22. Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. American Medical Infomatics Association. 2003:259-63.

23. Aldridge RW, Shaji K, Hayward AC, Abubakar I. Accuracy of probabilistic linkage using the enhanced matching system for public health and epidemiological studies. PLoS One. 2015;10(8):e0136179. https://doi.org/10.1371/journal.pone.0136179

24. Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Paper presented at the Annual ASA Meeting in Anaheim. CA. Washington, DC: Statistical Research Division, U.S. Bureau of the Census; 1990.

25. DuVall SL, Kerber RA, Thomas A. Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators. Journal of Biomedical Informatics. 2010;43:24-30. https://doi.org/10.1016/j.jbi.2009.08.004

26. Porter EH, Winkler WE, editors. Approximate string comparison and its effect on an advanced record linkage system. Advanced record linkage system US Bureau of the Census, Research Report; 1997: Citeseer.

27. Winkler WE, Thibaudeau Y. An application of the Fellegi-Sunter model of record linkage to the 1990 US decennial census. US Bureau of the Census. 1991:1-22.

28. Durham E, Xue Y, Kantarcioglu M, Malin B, editors. Private medical record linkage with approximate matching. AMIA Annual Symposium Proceedings; 2010: American Medical Informatics Association.

29. Fellegi I, Sunter A. A Theory for Record Linkage. Journal of the American Statistical Association. 1969;64:1183-210. https://doi.org/10.1080/01621459.1969.10501049

30. Christen P, Churches T, Hegland M. Febrl–a parallel open source data linkage system. Advances in knowledge discovery and data mining: Springer; 2004. p. 638-47.https://doi.org/10.1007/978-3-540-24775-3_75

31. Mitchell RJ, Cameron CM, McClure R. Quantifying the hospitalised morbidity and mortality attributable to traumatic injury using a population-based matched cohort in Australia. BMJ open. 2016;6(12):e013266. https://doi.org/10.1136/bmjopen-2016-013266

32. Lawrence G, Dinh I, Taylor L. The Centre for Health Record Linkage: A New Resource for Health Services Research and Evaluation. Health Information Management Journal. 2008;37(2):60-2. https://doi.org/10.1177/183335830803700208

33. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. Journal of Biomedical Informatics. 2012;45(1):165-72. https://doi.org/10.1016/j.jbi.2011.10.006

Abbreviations

CHeReL	Centre for Health Record Linkage
FEBRL	Freely Extensible Biomedical Record Linkage
NSW	New South Wales
PPRL	Privacy Preserving Record Linkage

2.4.3 Grouping methods for ongoing record linkage

Randall SM, Ferrante AM, Boyd JH, **Brown AP**, Semmens JB (2015). *Grouping methods for ongoing record linkage* (2015) Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference.

Grouping methods for ongoing record linkage

Sean M. Randall
Centre for Data Linkage
Curtin University
Perth, Australia
sean.randall@curtin.edu.au

James H. Boyd
Centre for Data Linkage
Curtin University
Perth, Australia
j.boyd@curtin.edu.au

Anna M. Ferrante
Centre for Data Linkage
Curtin University
Perth, Australia
a.ferrante@curtin.edu.au

Adrian P. Brown
Centre for Data Linkage
Curtin University
Perth, Australia
adrian.brown@curtin.edu.au

James B. Semmens
Centre for Population Health Research
Curtin University
Perth, Australia
james.semmens@curtin.edu.au

ABSTRACT

The grouping of record-pairs to determine which records belong to the same individual is an important part of the record linkage process. While a *merge* grouping approach is commonly used, other methods may be more appropriate when linking to a repository of previously linked data.

In this paper, we applied a number of grouping strategies to three large scale hospital datasets (comprising around 27 million records), each with a known truth set. These datasets were linked against a created 'repository' whose quality was varied.

Experimental results show that alternate grouping methods can yield very large benefits in linkage quality, especially when the quality of the underlying repository is high. *Best link* methods can remove between 25-90% of matching errors, depending on the characteristics of the underlying datasets.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Record linkage, grouping

This paper was presented at the First International Workshop on Population Informatics for Big Data (PopInfo'15), Sydney, 10 August 2015.

1. INTRODUCTION

Widely utilised in health research, record linkage involves identifying records which belong to the same individual within and across administrative datasets. By linking together records from hospital and emergency collections, primary care facilities, and birth, death and disease registries, researchers can construct a chronological sequence of events for a particular individual. The linkage process provides researchers with an enriched, cost effective, longitudinal research dataset for the study of entire populations.

In the absence of a unique identifier, linkage involves matching records using personal identifiers (e.g. name, address, and date of birth). As this information changes, and/or can be in error, statistical techniques are used to ensure links of the highest quality [4]. Ensuring high quality is critical in record linkage, as research outcomes can be affected. Current methods used to maintain linkage quality [15, 3] are heavily manual which is both costly and time-consuming. Identifying methods to improve quality that do not rely on manual review is of high interest [12].

Specialised linkage units often provide the infrastructure and expertise required to carry out record linkage. These units carry out linkage on an on-going basis, creating a list of all records and the person identifier to whom they belong. Incoming datasets are linked to the repository which is updated with this new information.

During the linkage process, incoming data is first cleaned to ensure consistency and reliability. The files are then matched using a defined linkage strategy, resulting in pairs of records designated as belonging to the same person. A grouping or clustering process then amalgamates these record-pairs into groups to identify the full set of records belonging to the same individual.

The traditional grouping process uses transitive closure to merge all identified record-pairs, with all connected records being assigned to the same individual. Transitive or *indirect* links are formed where records which did not form a pair relationship nonetheless are assigned to the same individual, for instance because they form record-pairs with a

third record.

The merge based grouping process treats the repository as simply another set of records. However there is reason to believe that existing groups of records within the repository should rarely be merged together by incoming records - these groups have already been validated and are unlikely to be in error.

2. OBJECTIVES

We hypothesise that the use of grouping methods which reduce or remove the opportunity for groups within a repository to be joined together should result in higher linkage quality than the traditional merge based method. One such method has been suggested previously [9]; however this method (*best link* grouping) has never been evaluated against the traditional merge approach used in many operational linkage units across the world.

In this paper, we present an alternate best-link algorithm for grouping, and evaluate this algorithm against both the merge based and best link algorithms using real world datasets. We hypothesise that the appropriateness of these grouping techniques for on-going linkage will depend on the overall quality of the repository used. To test this, repositories of differing quality were used in the evaluation to allow us to determine the circumstances in which particular methods are appropriate.

3. METHODS

3.1 Grouping Methods

3.1.1 Merge Based Grouping

Merge grouping amalgamates all record pairs above the accepted threshold, with all connected records belonging to the same individual. Indirect or transitive links are formed where records which did not form a pair relationship nonetheless are marked as belonging to the same individual, for instance because they are both linked to a third record. If multiple groups in the repository are linked together in this way, these are merged. There is no limit to the length of indirect links accepted, although this can be used as a potential indicator of groups containing errors [12].

3.1.2 Best Link

In the approach presented by Kendrick [9], grouping is carried out in the order in which the records are matched. Each record from the incoming file is matched in turn against records in the repository. If the record from the incoming file matches to multiple records in the repository file, only the highest weighted match is accepted, and the record from the incoming file is added to this group. If the record does not link to any records in the repository, a new group is created, of which it is the sole member. The incoming record is then added to the repository, and subsequent records in the incoming file are able to match against this added record.

3.1.3 Weighted Best Link

Our modified grouping strategy which we will refer to as weighted best link, involves a linkage of records from the incoming file to the repository (along with a de-duplication of the incoming file) where all record pairs are created and evaluated. Once the linkage is completed, accepted record pairs are amalgamated in weight order. The pairs are examined in order from highest to lowest; a record-pair is accepted as valid provided it does not result in multiple groups from the repository merging together.

Algorithm 1 Best link

```
   Input: Incoming file, Repository
 1: for each record in Incoming File do
 2:     link record to Repository
 3:     if there is one pair found then
 4:         add record to that group
 5:     else if there are multiple pairs found then
 6:         choose the highest pair
 7:         add record to that group
 8:     else if there are no pairs found then
 9:         mark record as belonging to a new group
10:     add record to Repository
```

Algorithm 2 Weighted best link

```
   Input: Incoming file, Repository
 1: Link Incoming file to Repository
 2: Deduplicate Incoming file
 3: Concatenate pairs from (1) and (2)
 4: Sort output of (3) in weight descending order
 5: for each pair in sorted pairs do
 6:     if accepting will merge two repository groups then
 7:         ignore pair
 8:     else
 9:         accept pair
```

Both best link methods assume that record-pairs have some ordinal attribute which identifies how likely they are to belong to the same individual. In probabilistic linkage, this is the weight attached to each record-pair [11]. For deterministic linkage (another common method of record linkage), these grouping strategies can be used by ordering rules by strictness.

Both best-link algorithms are similar, and in many situations return the same results. An example of their difference is shown in Figure 1. Using the best link approach, the first record A is matched to record Z and joins this group. The second record B matches to both A and Y. Of these, A is the highest weighted, so record B will join the same group as A and Z. In the weighted best link method, the first accepted pair is that joining the incoming records A and B. The next pair joins B and Y; A, B and Y are now linked together. The final pair linking A to Z is ignored, as this would bring together two groups from the repository.

The advantage of the modified weighted best link methods is that it will consistently produce the same results irrespective of the order of records being processed. The best link method described by Kendrick [9] will produce different grouping results if the linkage of the incoming records is executed in a different order.

3.2 Evaluation Datasets

Three large hospital admissions datasets were used in this evaluation, for which we had pre-existing and accurate information about which records belonged to the same person.

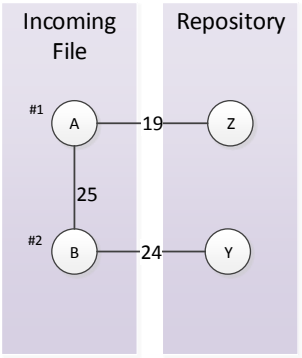


Figure 1: An example of the difference between best link algorithms. The number between records represents the weight of the record-pair comparison.

This information acted as the 'truth set' for each dataset and was used to compute differences in the performance of the three grouping algorithms. Ten years of Western Australian (WA) Hospital Admissions data, along with ten years of New South Wales (NSW) Admitted Patient Data and eight years of South Australian (SA) Hospital Admissions data were used in the evaluation. These datasets contained the typical data quality errors found in administrative data, including misspellings, name variations, missing data, changes in personal identifiers and incorrect values. Each dataset had been previously de-duplicated (by the WA Data Linkage Branch [8], the Centre for Health Record Linkage [10], and SA-NT DataLink respectively) utilising a variety of methods including exact matching, probabilistic linkage and intensive clerical review. All the linkage units employ rigorous manual reviews of created links, and a quality assurance program to analyse and review likely errors [3, 15] These links are further validated through use in a large number of research projects and published research articles [2]. Both WA and NSW have been operational for many years while in comparison SA data has only recently been linked, and has therefore been subject to less review by both clerical assessors and researchers. The data was made available as part of the Population Health Research Network Proof of Concept project [1]. A summary of the datasets is provided in Table 1.

3.3 Matching Strategy

A single matching strategy was used for all linkages in the study. This strategy utilised a probabilistic approach and was based on a previously published 'default' linkage strategy [7]. Two sets of blocks were used: Soundex of surname with first initial, and full date of birth. All variables were used in comparisons; string similarity measures were used for alphabetic variables (name, address and suburb) with exact matches used for all other variables. Agreement and disagreement weights were estimated.

3.4 Measuring Linkage Quality

Linkage quality was evaluated using saturated pairwise precision, recall and f-measure. Precision refers to the proportion of found links that were correct, and thus provides a measure of false positives. Recall is the proportion of all correct links found, and thus measures false negatives. The F-measure is the harmonic mean between precision and recall, giving a single figure from which we can compare results. These measures have been recommended for use in record linkage [5].

3.5 Repository Creation

To simulate linkage of an incoming file to a central repository, it was necessary to create *repositories* (datasets with coverage of close to the whole population). A repository for each of the original data sources was created by first randomly selecting one record per person from the hospital admissions file. This repository was 'complete' in the sense that it had coverage of the whole population being linked, and did not contain records for the same individual in more than one group.

Additional repositories of degraded quality were created by both removing records from the 'complete' repository, and by adding additional records belonging to a person already in the repository, as a separate person. Additional 'duplicate' records were specifically chosen so that differences existed in the personal identifiers between the records in the repository belonging to the same person.

Four repositories in total were created from each original dataset, differing in the number of errors they contained. These included a 'complete' repository, a repository with 1% of records missing and 1% of groups duplicated, a repository with 2.5% records missing and 2.5% groups duplicated, and a repository with 5% records missing and 5% of groups duplicated.

3.6 Evaluation Strategy

The linkage of the three datasets to their corresponding repositories was conducted separately; there was no linkage between hospital datasets.

'Incoming files' for linkage were constructed by breaking the hospital admissions records into batches containing admission records for a three month period. The batches were then linked to the repository in temporal order, to simulate on-going linkage. Records that were used to create repositories did not form part of the incoming files.

Each linkage of a batch of incoming records to the corresponding repository was grouped using three different methods - the traditional merge based method, best link and the new weighted best link approach.

Linkages were conducted using four different repositories, with three different grouping strategies, on the three state-based datasets, for a total of 36 linkage runs. The quality of each run was measured using the metrics described above.

4. RESULTS

The optimal F-measures of the overall linkage (after all batches were added) for each linkage run are shown in Figure 2. The figure displays the maximum F-measure achieved across a range of possible threshold settings.

Table 1: Dataset characteristics

Missing Values	NSW Morbidity	WA Morbidity	SA Morbidity
Surname	31.9%	<0.1%	5.3%
Given Names	33.9%	<1.0%	5.5%
Sex	<0.1%	<0.1%	<0.1%
DOB	<0.1%	<0.1%	0
Suburb	<1.0%	<1.0%	6.9%
Address	7.5%	<0.1%	8.1%
Postcode	<1.0%	<1.0%	8.5%
N	19,874,083 records	6,772,949 records	2,509,914 records

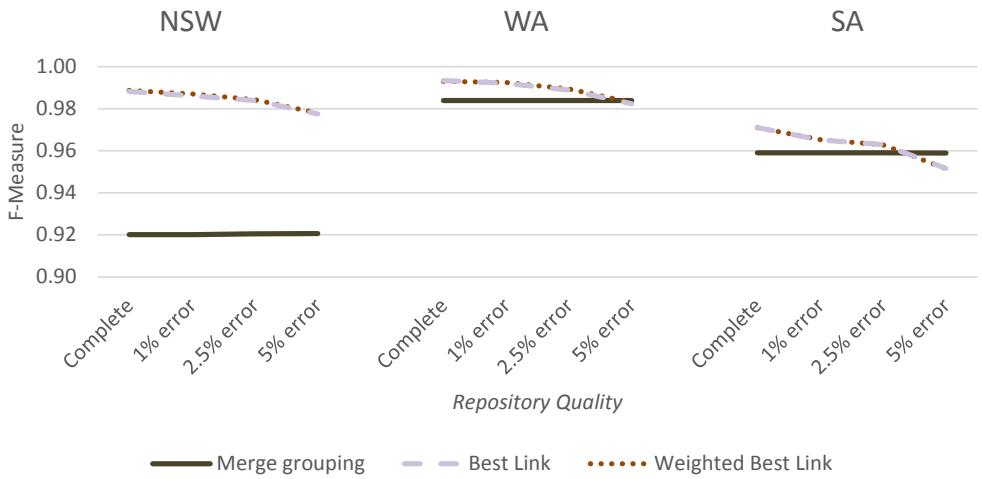


Figure 2: Results of grouping by repository quality

As can be seen, the effectiveness of merge-based grouping as compared with best link methods depended heavily on both the dataset used and the quality of the repository. For all datasets, the best link methods were superior when using a repository with an error rate of 2.5% or less. For an error rate of 5%, the most effective grouping strategy varied with the dataset.

Merge based grouping was not affected by repository quality, whereas the linkage quality of the best link methods decreased as the quality of the repository was degraded. This is unsurprising, as merge based grouping accepts all record-pairs above a certain threshold, without regard for the constitution of the repository, whereas best link methods will specifically reject certain record-pairs above the threshold based on records found in the repository.

Little difference was observed in the maximum F-measure between the two best link methods. This was a consistent finding across all datasets and all levels of repository quality.

Figure 3 shows the overall F-measure for each threshold value, for all grouping methods and for all repositories; displayed threshold are those found through probabilistic record linkage using the method of Fellegi-Sunter [6]. For higher valued thresholds, there was no difference between the merge based strategy and either of the best link strategies; however, for lower chosen thresholds the F-measures diverged, with merge based grouping scores rapidly decreasing, while best link scores improved.

As the threshold decreases, the number of false-positive pairs increase. The merge grouping method includes these false-positive pairs, resulting in lower linkage quality. Best link methods only accept these false-positives pairs if the incoming record has not already linked to a record in the repository. As this is nearly always the case, the vast majority of these false-positives are ignored, and so linkage quality remains relatively unchanged. For higher thresholds where there are fewer false-positives, there are smaller differences between these approaches.

A final notable difference is the much greater threshold range over which the F-measure for best link grouping is at a maximum.

5. DISCUSSION

The results of this study show that when optimising for linkage quality, the most appropriate grouping strategy depends on the underlying quality of the repository. If the repository is not representative of the study population or of poor quality with little confidence in the established groups, the merge based method can be considered as a possible grouping strategy. However, for better quality repositories, best link methods result in much higher linkage quality. It would be expected that most data repositories, or well-maintained

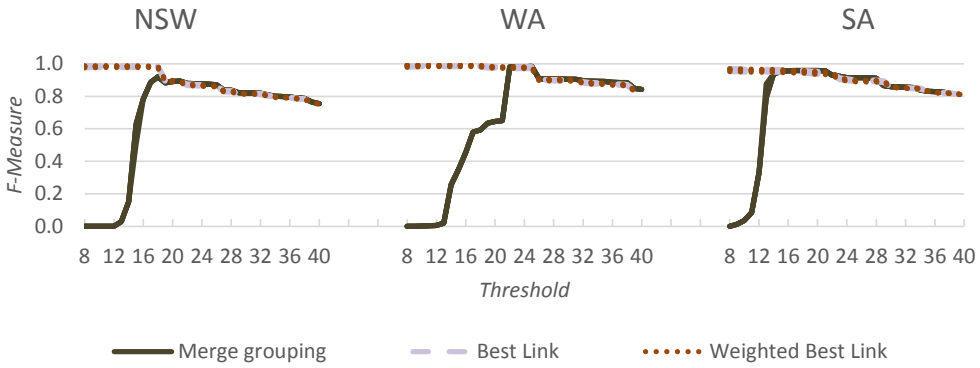


Figure 3: Results of grouping by threshold score

datasets with high population coverage, would contain only a small level of error, making best link the most appropriate grouping strategy to adopt. As the results indicate, best link methods have the added advantage of being insensitive to threshold changes. This increased tolerance reduces the likelihood of threshold estimation errors and suggests that these grouping methods could be useful in situations where determining thresholds is difficult, such as in privacy preserving linkage [13].

Our results were also highly dataset dependent, with best link methods proving superior on NSW data for all repositories. This is likely to be a reflection of the lower data quality (the NSW data has much higher rates of missing values; see Table 1).

Results showed little difference between the two best link methods. Factors other than linkage quality may be more appropriate in determining which of these methods should be used in ongoing linkage. The weighted best link method has the advantage that results are repeatable and not dependent on the order of incoming records. This means that it is possible to retrace and understand the sequence of links that were created over time without knowing the order in which records arrived. The weighted method also has the advantage that grouping decisions are made independently of matching decisions. This de-coupling of processes may be important in the design and development of linkage systems.

Given the dataset-specific nature of the results from this study, additional testing against other datasets may be required to gain a full understanding of the relationship between linkage quality, grouping strategy and population repository quality.

Our results show that the choice of grouping strategy can make a large difference to linkage quality. Within this evaluation, best link methods were able to remove between 25% (SA) to 90% (NSW) of matching errors using a high quality repository. This is an extremely large improvement in linkage accuracy, yielding far larger gains than other techniques in the literature [14, 12].

6. CONCLUSION

The effect of grouping methods on linkage quality is an understudied area of research. By adopting an appropriate grouping strategy, vast improvements in linkage quality can be achieved. The weighted best link strategy presented here shows large improvements against the merge strategy currently in operation, while providing practical benefits over the previous best link method.

Current methods of improving quality present as processing bottlenecks. Methods which improve the overall quality of linked data without impacting on performance will ultimately lead to more accurate and reliable research outcomes and increased utilisation of this resource by researchers.

7. ACKNOWLEDGMENTS

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy's Population Health Research Network.

8. REFERENCES

[1] J. H. Boyd, A. M. Ferrante, C. M. O'Keefe, A. J. Bass, S. M. Randall, and J. B. Semmens. Data linkage infrastructure for cross-jurisdictional health-related research in australia. *BMC health services research*, 12(1):480, 2012.

[2] E. L. Brook, D. L. Rosman, and C. J. Holman. Public good through data linkage: measuring research outputs from the western australian data linkage system. *Australian and New Zealand journal of public health*, 32(1):19–23, 2008.

[3] Centre for Health Record Linkage. Quality assurance, 2015. [Online; http://www.cherel.org.au/quality-assurance; accessed 3-June-2015].

[4] P. Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[5] P. Christen and K. Goiser. Quality and complexity measures for data linkage and deduplication. In *Quality Measures in Data Mining*, pages 127–151. Springer, 2007.

[6] I. P. Fellegi and A. B. Sunter. A theory for record

linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[7] A. Ferrante and J. Boyd. A transparent and transportable methodology for evaluating data linkage software. *Journal of biomedical informatics*, 45(1):165–172, 2012.

[8] C. D. J. Holman, J. A. Bass, D. L. Rosman, M. B. Smith, J. B. Semmens, E. J. Glasson, E. L. Brook, B. Trutwein, I. L. Rouse, C. R. Watson, et al. A decade of data linkage in western australia: strategic design, applications and benefits of the wa data linkage system. *Australian Health Review*, 32(4):766–777, 2008.

[9] S. Kendrick, M. Douglas, D. Gardner, and D. Hucker. Best-link matching of scottish health data sets. *Methods of information in medicine*, 37(1):64–68, 1998.

[10] G. Lawrence, I. Dinh, L. Taylor, et al. The Centre for Health Record Linkage: a new resource for health services research and evaluation. *Health Information Management Journal*, 37(2):60, 2008.

[11] H. B. Newcombe. *Handbook of record linkage: methods for health and statistical studies, administration, and business*. Oxford University Press, Inc., 1988.

[12] S. M. Randall, J. H. Boyd, A. M. Ferrante, J. K. Bauer, and J. B. Semmens. Use of graph theory measures to identify errors in record linkage. *Computer methods and programs in biomedicine*, 115(2):55–63, 2014.

[13] S. M. Randall, A. M. Ferrante, J. H. Boyd, J. K. Bauer, and J. B. Semmens. Privacy-preserving record linkage on large real world datasets. *Journal of biomedical informatics*, 50:205–212, 2014.

[14] S. M. Randall, A. M. Ferrante, J. H. Boyd, and J. B. Semmens. The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making*, 13(1):64, 2013.

[15] D. Rosman, C. Garfield, S. Fuller, A. Stoney, T. Owen, and G. Gawthorne. Measuring data and link quality in a dynamic multi-set linkage system. In *Proceedings of the Symposium on Health Data Linkage*, 2002.

2.5 Published letter(s)

2.5.1 Understanding the Origins of record linkage error and how they affect research outcomes

Boyd JH, Ferrante AM, Irvine K, Smith M, Moore E, **Brown AP**, Randall SM (2016). *Understanding the Origins of record linkage error and how they affect research outcomes* (2016) Australia and New Zealand Journal of Public Health.

doi: 10.1111/1753-6405.12597

Understanding the origins of record linkage errors and how they affect research outcomes

James H. Boyd,[1] Anna M. Ferrante,[1] Katie Irvine,[2] Michael Smith,[2] Elizabeth Moore,[2] Adrian Brown,[1] Sean M. Randall[1]

1. Centre for Population Health Research, Curtin University, Western Australia
2. Ministry of Health, New South Wales

Major investment in record linkage infrastructure in Australia and internationally reflects the strategic value of high-quality linked datasets. Dedicated record linkage units with secure environments and specialised linkage personnel have been established to support academic research, policy development and service design by government.[1] The challenge for units creating linked data is to maximise linkage quality using a variety of matching and management techniques. However, it is also important that researchers understand processes around both data collection and linkage to ensure that they are aware of strengths and limitations of the data and methods used to bring together records. In this way, research study design can be optimised and potential for misinterpretation is reduced.[2]

Knowing that linkage error can affect interpretation of research findings and introduce bias highlights a need for routinely measuring and reporting linkage quality.[3] Unfortunately, although there are a variety of standard matching and management techniques available, the assessment of linkage results varies.

The impact of linkage quality on research outcomes

While several papers have observed that reduced linkage quality can affect research results,[4] little is known about how errors (and different types of errors) directly affect particular methods of analysis.

In one-to-one linkage (involving only two datasets, where each has one record per person), the effect of linkage error is often simpler to assess. For instance, measures of prevalence (as identified by a link between the two datasets) will remain reliable where there is an equal number of false positives to false negatives. However, research suggests that in analysis of an association between exposure and disease, where exposure and disease are located in separate datasets, optimising for higher linkage specificity would achieve the most accurate rate ratio.[4]

More commonly, linkage involves many-to-many scenarios where the relationship between linkage quality and errors is often difficult to evaluate. Gaining a greater understanding of the impact of false positives and negatives on particular methods of analysis, including whether either of these error types plays a bigger role in biasing outcomes, would help interpretation and validity of research findings.

Evidence also suggests that record linkage errors are not distributed evenly throughout the population. Instead, these vary among particular subgroups. Sub-populations with greater levels of linkage error include women,[5] the elderly,[5] ethnic minorities,[5,6] indigenous people,[7] defined geographic areas (from recording differences in particular localities) and those from lower socioeconomic groups.[8] Analysis of both linked and unlinked records is an important step that allows assessment of potential variations within population subgroups, e.g. geographical, cultural, remoteness, etc. A growing body of Australasian research demonstrates lower linkage rates for indigenous people, and methods in both unlinked and linked analyses that correct for this.[7,9]

Mitigating the impact of linkage error

Given researchers' limited ability to detect incorrect links, and the infancy of statistical methods to control for linkage error, it is vital that linkage units work with researchers to develop sound statistical models and to provide accurate and detailed information about the quality of links provided. Information on linkage quality allows researchers to assess/address any bias in the study design (e.g. if data is coming from different systems, are the data and linkage results consistent)[1] or to allow adjustment to statistical confidence levels in the interpretation of results.

Greater transparency and improved reporting of linkage results will help researchers to improve study design, understand the impact of analytical techniques and strengthen the interpretation of results. Currently, there are no standard methods for assessing and reporting on the quality of linkage outputs.

Final remarks

Achieving high linkage quality is essential for ensuring and maintaining the quality and integrity of research based on linked data. It is important that researchers make time to understand both the data being used within a study (e.g. how it was collected, coding structures, completeness, etc.) and the linkage process use to create participant profiles for a research study. This may require additional information from data linkage units such as reports on the software, linkage strategy and on matching quality to ensure the appropriate analyses can be performed.[3,10]

The Population Health Research Network is a collaborative network of linkage units across Australia supporting research using linked data. The network understands the need to measure, monitor and improve linkage quality and is working to improve the measurement and reporting of linkage quality.

Acknowledgement

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy and Super Science Initiative's Population Health Research Network.

References

1. Boyd JH, et al. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC Health Serv Res*. 2012;12(1):480.
2. Harron K, et al. Evaluating bias due to data linkage error in electronic healthcare records. *BMC Med Res Methodol*. 2014;14(1):36.
3. Harron K, et al. Opening the black box of record linkage. *J Epidemiol Community Health*. 2012;66(12):1198.
4. Copeland KT, et al. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*. 1977;105(5):488-95.
5. Zingmond DS, et al. Linking hospital discharge and death records—accuracy and sources of bias. *J Clin Epidemiol*. 2004;57(1):21-9.
6. Lariscy JT. Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox. *J Aging Health*. 2011;23(8):1263-84.
7. Shaw C, Atkinson J, Blakely T. (Mis) classification of ethnicity on the New Zealand Cancer Registry: 1981-2004. *N Z Med J*. 2009;122(1294):10-22.
8. Bohensky MA, et al. Data linkage: A powerful research tool with potential problems. *BMC Health Serv Res*. 2010;10(1):346.
9. Boyd M, Atkinson J, Blakely T. Ethnic counts on mortality, New Zealand Cancer Registry and census data: 2006–2011. *N Z Med J*. 2015;129(1429):22-39.
10. Neter J, Maynes ES, Ramanathan R. The effect of mismatching on the measurement of response errors. *J Am Stat Assoc*. 1965;60(312):1005-27.

Correspondence to: Associate Professor James Boyd, Centre for Data Linkage, Curtin University, Kent Street, Bentley, WA 6102; e-mail: J.Boyd@curtin.edu.au

Chapter 3

Privacy and performance

Included Manuscript(s):

4. **Brown AP**, Borgs C, Randall SM, Schnell R (2017). *Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets.* BMC Medical Informatics and Decision Making, 17(1), 83. https://doi.org/10.1186/s12911

9. Randall SM, **Brown AP**, Ferrante AM, Boyd JH (2019). *Privacy preserving linkage using multiple match-keys* (2019) International Journal of Population Data Science, 4(1). https://doi.org/10.23889/ijpds.v4i1.1094

Conference proceeding(s):

6. Randall SM, **Brown AP**, Boyd JH, Ferrante AM, Semmens JB (2015). *Privacy preserving record linkage using homomorphic encryption* (2015) Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference.

Aim 3 of this thesis is to *identify methods for improving privacy and performance of privacy-preserving record linkage*. Privacy-preserving techniques are typically slower than traditional methods as they require considerable additional computation. This chapter addresses this aim by examining new ways to improve both the privacy and performance aspects of PPRL.

3.1 The trade-off between privacy, performance and quality

There are three main characteristics of privacy-preserving record linkage techniques that can be used to determine how well they perform: privacy, quality and efficiency. Often, there is a trade-off required between these three attributes; many techniques that increase privacy will impact negatively on quality and/or speed in some way. For example, there are several modifications to Bloom filters (composite Bloom filters [86, 251], XOR folding [255], Randomized Response [254] and salting [206]) that increase privacy, yet negatively impact upon quality. Statistical disclosure limitation (SDL) can also be used to deliberately introduce error into the data, improving the privacy and confidentiality aspect of the data at the expense of quality [156].

PPRL using Bloom filters exploits the ability for similarity comparisons between Bloom filters. This characteristic results in patterns within Bloom filters that are based on the characteristics of the original data. These patterns can also be exploited through cryptanalysis attacks. Modifications to Bloom filters for improving privacy aim to reduce these patterns sufficiently so that the cryptanalysis attacks are no longer successful. However, reducing these patterns also negatively impacts the effectiveness of the similarity comparisons, thereby reducing linkage quality.

An essential part of the data linkage process for determining the efficiency of the linkage is the indexing step. Indexing removes as many comparisons as possible that are likely to be false matches, leaving a significantly reduced set of record-pairs to match. Indexing solutions on unencrypted data typically seek high quality (pairs completeness) while striving for the best possible efficiency. Privacy-preserving techniques add privacy as an additional complexity to this, impacting both quality and efficiency.

There is currently no magic bullet for PPRL that singularly solves the privacy, quality and efficiency trade-off problem. However, PPRL methodologies are still relatively undeveloped, and there are opportunities to explore alternative and additional techniques in this space.

3.2 Encoding for privacy

Simple hash encoding of field values using MD5 and SHA were the first approaches to protecting the identity of individuals in data for linkage [88]. Today, this approach has been superseded; for example, the use of HMACs with strong hash functions (such as SHA256), providing a keyed hash. A secret key is used in addition to a field value during the HMAC computation.

Assuming the secret key is of a sufficient length, HMACs can provide a very high level of privacy protection. However, without providing some kind of record derived salt (an additional value that is used as input into a hash function) in the key, field-level hashing is susceptible to frequency attacks. The nature of hashing also prevents the use of any kind of similarity comparison, affecting the quality of the linkage, particularly with datasets of poor quality. This is an important limitation for data linkage.

The use of a single match key (often referred to as an anonymous linkage code) is a preprocessed subset of identifiers that represent an individual [42]. Too many identifiers and variations in data (for example, due to typographical errors) can result in records being identified as a different individual. Too few identifiers and separate individuals may be identified as the same person. The SLK-581 [242], used in Australia, and the Swiss Anonymous Linkage Code [32] are two examples used in practice. Match keys can provide excellent privacy protection through the use of HMAC and a strong hash function; their weakness, however, is reduced linkage quality [42].

The paper, *Privacy preserving linkage using multiple dynamic match keys*, included as a supporting manuscript for this thesis, presents a new protocol for privacy-preserving linkage that creates multiple match keys for each record, with the composition of each match key dependent on attributes of the underlying datasets. The multiple match key method aims to combine the privacy protection offered by the use of anonymous linkage codes with the linkage quality offered by the probabilistic approach. This new protocol is evaluated on synthetic and real-world datasets, and compared against unencrypted linkage and other privacy-preserving linkage techniques. On most datasets, the multiple match key method produces linkage quality only slightly below that produced by field-level Bloom filters on most datasets. This protocol uses the parameters of a probabilistic model to calculate agreement scores for each record's set of match keys to determine which match keys should be used. Therefore, the estimation of appropriate m and u probabilities and a threshold score is essential to achieving high quality. This is no different to standard probabilistic linkage; however, instead of comparing fields independently and summing the results, this approach simply looks for the same match key value between two records.

Field-level Bloom filters are establishing themselves as the benchmark for linkage quality amongst PPRL methods. Unfortunately, they are susceptible to the same frequency attacks as hash encoding, as well as a number of attacks that target the frequency of the q-grams used to construct the Bloom filter. Hardening methods, such as salting, can significantly reduce the chances of a successful frequency attack. However, deriving the right record-level salt is a challenge, and there have not been any published studies on this or its direct impact on linkage quality.

The paper, *Privacy preserving record linkage using homomorphic encryption*, included as a supporting manuscript for this thesis, extends the Bloom filter protocol to include a homomorphic

encryption step, which aims to completely remove the vulnerability to frequency attacks altogether; every field after encryption has entirely different ciphertext values. A homomorphic encryption scheme allows computations to be carried out on encrypted data producing encrypted results; when this encrypted data are finally decrypted, the decrypted results match those performed on an unencrypted version of the data. An evaluation of this extension showed that it achieved a degree of privacy of 0.0 (absolute privacy) using the privacy metrics of Vatsalan [278] while providing the same linkage quality as standard Bloom filters. While privacy is exceptionally high, the performance of homomorphic encryption is currently a critical limitation. Additional performance improvements could be made by using distributed computing techniques. Given the high security level of the encryption method, it may also be feasible to utilise public cloud computing resources to perform the inner product calculations, leveraging the elasticity of on-demand computation.

3.3 Indexing techniques

One of the simplest methods for indexing, known as blocking, involves partitioning datasets into mutually exclusive blocks, using values derived from one or more fields [126]. For example, the date of birth and sex values can be combined to produce a single blocking value; only records that have the same date of birth and sex values will be compared. One of the issues with this approach is that it does not account for any variation in the data. Even slight variations in data will prevent records from being compared. Thus, multiple blocks using different combinations of fields are often used to ensure that the highest possible coverage of true matches in the data is achieved. This technique is the most straightforward approach to indexing and has been effective in unencrypted linkage.

Blocking can be used in the context of PPRL with a strong HMAC hash of the block values. This technique will guarantee equivalent pairs completeness to unencrypted values. The selection of fields to use within the block is important from a privacy point of view as they may be susceptible to the same frequency attacks as individual fields. However, blocking in PPRL gives the opportunity for an ideal salt value to use for the records within each block. Other approaches to indexing encrypted identifiers, such as the Sorted Neighbourhood Method [125] and Canopy Clustering [191], have been developed, yet neither show optimal performance in all settings [60].

The use of composite Bloom filters, such as RBFs and CLKs, where a single Bloom filter represents each row, can prevent the ability for standard blocking to be used without creating separate external blocking values at encoding time. This negates some of the privacy characteristics that composite Bloom filters are trying to achieve through the use of a single value. There may also be constraints in some jurisdictions that prevent the use of more than one field [54].

The paper, *Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets*, included as part of this thesis, evaluates the linkage of

CLKs using multibit trees as an indexing technique on large datasets. The performance was measured against a gold-standard derived from clear-text probabilistic record linkage. Results showed that CLKs linked using multibit trees produced linkage quality less than that of field-level Bloom filters. This quality is highly dependent on the fields used to create the CLK, with missing values creating a large number of false positives. However, the use of multibit trees for indexing shows excellent potential, with one of the parameter sets achieving a higher recall than standard unencrypted linkage.

3.4 Conclusion

Maintaining high linkage accuracy while aiming to improve privacy and performance is a difficult task. Many techniques for improving privacy appear to have a direct trade-off on linkage quality or performance. Achieving the right balance between these properties may depend on the specific circumstances around a particular linkage of data. However, there is a lot of further work to be done, improving on homomorphic Bloom filters and the use of multibit trees for private indexing, which could alleviate the need for a trade-off of privacy or performance. The findings of this chapter will be discussed in the context of cloud computing in the final chapter.

3.5 Published manuscript(s)

3.5.1 Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets

Brown AP, Borgs C, Randall SM, Schnell R (2017). *Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets.* BMC Medical Informatics and Decision Making, 17(1), 83.

Brown *et al. BMC Medical Informatics and Decision Making* (2017) 17:83
DOI 10.1186/s12911-017-0478-5

BMC Medical Informatics and Decision Making

RESEARCH ARTICLE Open Access

Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets

Adrian P. Brown[1*], Christian Borgs[2], Sean M. Randall[1] and Rainer Schnell[2]

Abstract

Background: Integrating medical data using databases from different sources by record linkage is a powerful technique increasingly used in medical research. Under many jurisdictions, unique personal identifiers needed for linking the records are unavailable. Since sensitive attributes, such as names, have to be used instead, privacy regulations usually demand encrypting these identifiers. The corresponding set of techniques for privacy-preserving record linkage (PPRL) has received widespread attention. One recent method is based on Bloom filters. Due to superior resilience against cryptographic attacks, composite Bloom filters (cryptographic long-term keys, CLKs) are considered best practice for privacy in PPRL. Real-world performance of these techniques using large-scale data is unknown up to now.

Methods: Using a large subset of Australian hospital admission data, we tested the performance of an innovative PPRL technique (CLKs using multibit trees) against a gold-standard derived from clear-text probabilistic record linkage. Linkage time and linkage quality (recall, precision and F-measure) were evaluated.

Results: Clear text probabilistic linkage resulted in marginally higher precision and recall than CLKs. PPRL required more computing time but 5 million records could still be de-duplicated within one day. However, the PPRL approach required fine tuning of parameters.

Conclusions: We argue that increased privacy of PPRL comes with the price of small losses in precision and recall and a large increase in computational burden and setup time. These costs seem to be acceptable in most applied settings, but they have to be considered in the decision to apply PPRL. Further research on the optimal automatic choice of parameters is needed.

Keywords: Medical record linkage, Blocking, Indexing, Private record linkage

Background

In medical research, information on patients is often scattered across different databases of several data holders. The task of finding records referring to the same person across one or more datasets is, in medical contexts, denoted as *record linkage*. Linking databases is a valuable and cost-effective technique, increasingly used in public health [1, 2], official statistics [3, 4], medical service research [1, 5], pharmacovigilance [6] and demographic research [7]. Applications of record linkage in medical informatics enabled new research on topics such as increased mortality risk after imprisonment [8], increased risk of road traffic accidents after treatments for drug overdoses [9] or mortality for hepatitis C and HIV vs. non-HIV patients [10].

For many research endeavors, linking the information needed would be trivial if a unique personal identifier (PID) is available. However, in many settings, legal and administrative issues prevent the use of PIDs, restricting data linkage to personal identifiers such as names. Since this requires the release of personally identifying information to trusted third parties [11], privacy regulations, such as the HIPAA Privacy Rules [12] or current EU regulations

*Correspondence: adrian.brown@curtin.edu.au
[1]Centre for Population Health Research, Curtin University, Western Australia, Kent Street, Bentley, Western Australia, 6102 Perth, Australia
Full list of author information is available at the end of the article

[13], often mandate using encrypted personal information. Standard probabilistic record linkage methods [3] are sometimes unsuitable for methods based on encrypted identifiers.

A number of new record linkage methods have been developed to overcome this problem at a technical level. These methods, known collectively as *privacy-preserving record linkage*, allow linkages using encrypted identifiers. Although no personal identifying information is released by data custodians, record linkage is still possible.

A summary of privacy-preserving record linkage techniques notes that each method differs in its accuracy, maturity, practicality and suitability for large-scale linkages [14]. Few of the available privacy-preserving linkage techniques are suitable for operational linkage units [15].

One notable method for privacy-preserving record linkage utilises *Bloom filters* to enable linkage [16]. The Bloom filters main advantage over many other approaches is that it incorporates uncertainty into matching, allowing the similarity between two fields to be measured (for instance, between two surnames) – a method regularly used in traditional unencrypted record linkage that typically yields high quality. The original Bloom filter approach encodes each field into a separate Bloom filter (a binary vector) which is then compared for similarity using a measure such as the Sørensen-Dice coefficient or Jaccard index. The Dice coefficient of Bloom filter-encrypted identifiers seems to be comparable to the similarity of a Jaro-Winkler comparison on unencrypted identifiers [17]. As encryption occurs on individual fields, standard record linkage procedures can still be used such as blocking (to reduce the comparison space and allow timely linkage to occur) and the assignment of weights to particular fields. Real-world evaluations show similar linkage quality when comparing Bloom filter-based methods with clear-text probabilistic record linkage [15].

Alternate methods of privacy-preserving record linkage using Bloom filters have been developed, with a single Bloom filter composed from many identifiers. Reasons for using only a single Bloom filter for linkage include legal constraints in some jurisdictions [18] and attempts at improving the privacy of the data [19, 20]. A record-level Bloom filter (RBF) combines all fields into a single Bloom filter using the discriminatory power of each field [20]. Fields with a higher discriminatory power are allocated a larger proportion of bits within the RBF, with some bits excluded completely to maximise privacy. Another composite Bloom filter approach uses a basic set of identifiers to produce a cryptographic long-term key or CLK [19]. This was developed as an irreversibly encrypted, anonymous linkage code, that allowed for small typographical errors in the identifiers.

Both of these composite Bloom filter methods have been shown to increase privacy by reducing the chance of a successful, malicious attack [21, 22]. However, the ability of composite Bloom filters to perform highly accurately and efficiently on large real-world data is unknown. As there are no individual fields, indexing (or blocking) methods such as standard Blocking [3] cannot be used without blocking externally on a separate, encrypted identifier. Other approaches to indexing encrypted identifiers, such as the Sorted Neighbourhood Method [23] and Canopy Clustering [24], have been developed, yet neither show optimal performance in all settings [25]. Another recently introduced method using multibit trees has been shown to be very suitable for CLKs, with potential for good quality linkage, and with performance at least as good as other methods on synthetic data [26].

In this paper, we test the accuracy and efficiency of the multibit tree technique on CLKs generated from large real-world medical data, for which the true links (which records belong to the same person) are already known. Testing multibit trees on real-world data is an important step in verifying its viability for linking record-level Bloom filters in public health settings.

Methods

Datasets

Ten years of Western Australian (WA) Hospital Admissions data, along with ten years of New South Wales (NSW) Admitted Patient Data were used in this evaluation. For each of these datasets, we had pre-existing and accurate information about which records belonged to which person.

The datasets had been de-duplicated previously (by the WA Data Linkage Branch (WADLB) [27] and the Centre for Health Record Linkage (CHeReL) [28] respectively). De-duplication was undertaken using a variety of methods including exact matching, probabilistic linkage, and intensive clerical review. WADLB and CHeReL employed rigorous manual reviews of created links and a quality assurance program to analyse and review likely errors. These links have been further validated through use in a large number of research projects and published research articles [29], and are used as a 'truth set' for linkage quality estimations.

A summary of these datasets can be found in Table 1. The NSW Morbidity data has been separated into public and private hospital data. The private hospital data contains no name information.

Linkage quality metrics

Linkage quality was evaluated using pairwise precision, recall, and F-measure. Precision refers to the proportion of incorrect links found from all the found links and thus provides a measure of false positives. Recall is the proportion of all correct links found, and thus measures false negatives. The F-measure is the harmonic mean between

Table 1 Missing value percentages

Identifier	NSW morbidity (public hospital)	NSW morbidity (private hospital)	WA morbidity
First Name	3%	100%	< 1%
Middle Name	54%	100%	41%
Last Name	< 1%	100%	< 1%
Date of Birth	0%	0%	< 1%
Sex	< 1%	< 1%	< 1%
Suburb	< 1%	3%	< 1%
Address	2%	22%	< 1%
Postcode	< 1%	3%	< 1%
# Records	13810088	6498579	6772949

precision and recall, giving a single figure from which we can compare results. These measures are widely used in the record linkage literature [16, 30].

CLK method

The CLK encryption method is based on the idea of hashing all available personal identifiers into a single structure called a Bloom filter (a binary vector). Each Bloom filters is used as an encrypted linkage key and can then be compared with other keys, resulting in a score which describes how similar the Bloom filters (and thus the personally identifying information) are.

Four different parameter sets were tested, which corresponded to different choices of personal identifiers to combine into each CLK, and are outlined in Table 2. These parameter sets replicate typical blocking and linkage options in traditional record linkage.

Consistent with the CLK construction method suggested by Schnell et al. [19], each dataset was transformed into four CLK files, one for each parameter set. All CLKs were 1000 bits in length. Each identifier in the parameter set used to make up the CLK (i.e. first name, date of birth, etc.) was converted into unigrams (individual characters) or bigrams (sets of two overlapping characters) with each unigram or bigram hashed 10 times. The modulus of each hash with respect to the Bloom filter was taken, and this position in the Bloom filter set to 1.

Table 2 Identifiers used for each parameter set

Identifier	Parameter sets				Average length
	Set 1	Set 2	Set 3	Set 4	
First Name	✓	✓	✓	✓	5
Middle Name	✗	✓	✗	✓	5
Last Name	✓	✓	✓	✓	6
Date of birth	✓	✓	✓	✓	8
Sex	✓	✓	✓	✓	1
Suburb	✗	✓	✓	✗	8
Address	✗	✓	✗	✗	17
Postcode	✗	✓	✓	✗	4

Pairs of Bloom filters are compared using the Jaccard, or Tanimoto, similarity. The intersection of the bit positions set to one in both Bloom filters is divided by the union of the bit positions set to one in the two Bloom filters. This results in a similarity score between 0 and 1, where a higher score reflects a greater similarity measure:

$$J(A,B) = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|}$$

Security of CLKs and Bloom filters

The desirable property of all Bloom filter-based encryptions is that they are similarity-preserving. This presents security considerations, as this property can be exploited to attack the encryption and potentially reveal personal identifiers. In recent years, several attacks have been published. The first attack, proposed by Kuzu et al. [21], revealed personal identifiers by performing a frequency analysis of individual fields. A discussion on the scope and limitations of the attack is given by Schnell and Borgs [31].

A second attack was devised by Niedermeyer et al. [32] and extended by Kroll and Steinmetzer [33], which focuses on the frequency distributions of the bit patterns of Bloom filters, as well as CLKs. The attack was very successful in decoding CLKs using the double-hashing scheme as proposed in the original publication [16]. However, replacing the double-hashing scheme with full random hashing prevents the attack [31]. Several other hardening techniques have been proposed to make CLKs more resilient against bit-pattern based attacks [31, 34]. For example, using a stable identifier as an additional part of the secret (password) used for encryption is suggested by [32] as a hardening method (salting). Currently, there are no published attacks on such variants of the CLK construction.

Multibit trees

Searching for similar pairs is computationally expensive. To reduce the search space and thus improve linkage speeds, tree-based structures can be used for blocking. One prominent method is the use of multibit trees, as suggested by Kristensen et al. [35] and suggested for PPRL by Bachteler et al. [36]. Multibit trees show better performance in terms of quality and linkage speed than most current methods, like Canopy Clustering [26], LSH-based blocking [37] or PPJoin [38]. A tree structure is constructed for one record file by finding multiple match bit positions in all Bloom filters where approximately half the records have their bit position set to one, while the other half exhibits a value of zero. Each of these halves are called leaves. This *split-half technique* is repeated until a user-defined minimum number of records in each leaf

is reached (usually one to eight records). For our experiments, a leaf limit of one was used.

To find similar pairs in terms of Tanimoto-similarity, every record in the second dataset is queried sequentially. For each record, an upper bound of the Tanimoto-similarity can be estimated before the actual similarity calculation, by comparing the values at the bit positions of each leaf in the tree. Leaves with a similarity under a user-defined Tanimoto threshold are disregarded in the calculation of the similarities. This way, the search space can be reduced drastically.

For our de-duplication linkages, the same dataset was used for the multibit tree and for the sequential queries. We applied a construction method for multibit trees similar to Bachteler et al. [36], testing multiple Tanimoto thresholds for each parameter set.

Evaluation strategy

All NSW and WA datasets were encrypted into CLKs for each parameter set as described above. For testing of linkage quality and blocking ability on data with few missing values, the WA CLK dataset was then de-duplicated, using multibit trees as the blocking method, at a range of Tanimoto thresholds. For testing of linkage quality on data with many missing values, a random sample of 5 million records was taken from the combined NSW CLK datasets using parameter set 1 (first name, last name, date of birth and sex). This represents a reasonable sample size for a real-world operation, the name identifiers resulting in approximately 30% missing values. The pair-wise precision, recall and F-measure scores were calculated by comparing results to the 'truth set.'

For testing of performance, the NSW (Public Hospital) and WA CLK datasets were combined for a dataset with a total of approximately 20 million records. From this combined dataset, random samples were taken to create datasets of 5, 10 and 15 million records. All of these datasets were then de-duplicated, using multibit trees with a single Tanimoto threshold of 0.85, as this has previously been shown to be a reasonable value for most applications [26]. The execution time of the multibit tree search was recorded.

All de-duplication linkages used multibit trees with a leaf limit value of one. The multibit tree outputs all candidate pairs, where the criterion for a pair is that it exceeds the given Tanimoto threshold value.

The evaluation was run on a Windows Server 2012 R2 Virtual Machine, running under ESXi on a Cisco UCSC-C240-M3S Server with Intel Xeon CPU E5-2609@2.40GHz. The VM was assigned 48GB RAM and 6 vCPUs. The evaluation code was assigned 4 vCPUs.

Results

Linkage quality

Results for the de-duplication of the WA CLK dataset can be found in Fig. 1. The highest recall value across all threshold levels was achieved using parameter set 1 (first name, last name, date of birth, sex), with the best value of 0.986 at a threshold of 0.8. The next highest recall was achieved using parameter set 4 (first name, middle name, last name, date of birth, sex). The two lowest recall values came from the use of parameter sets 2 and 3. Values for parameter set 3 at Tanimoto thresholds 0.8 and 0.85 are not provided as these runs failed to complete successfully.

Maximum F-measure varied considerably across the different parameter sets. Highest F-measure was 0.978 from parameter set 1 while lowest F-measure was 0.781 for parameter set 3. The inclusion of address information (parameter sets 2 and 3) tended to reduce overall scores. This can be explained by the varying recall: including addresses introduces unstable identifiers, which either differ in the datasets (e.g. because individuals have moved to a different address) or are missing. This will lead to a

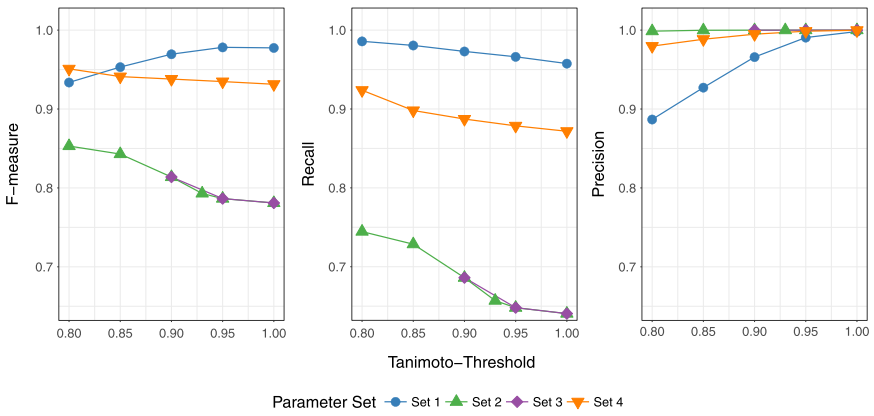


Fig. 1 Linkage quality results for WA CLK de-duplication

reduction in the amount of true pairs found, which is why sets 1 and 4 show superior linkage quality with respect to recall.

All parameter sets but set 1 show high precision scores. Since adding middle names allows for better discrimination between records that would otherwise exhibit the same values across all identifiers, the amount of false positive classifications will decrease, leading to increased precision values for these parameter sets.

The de-duplication linkage of the 5 million sample CLK dataset of the combined NSW Public and Private Hospital datasets (30% of all rows had missing name identifiers) was abandoned after 2 weeks of elapsed execution time. Analysis of the pairs created to that point showed that the number of missing identifiers in the CLKs was leading to the creation of an inordinately large number of false positives; a large portion of rows with only values for date of birth and sex appeared to be linking to each other. The anticipated poor linkage results and excessive processing time led to the decision to abandon all linkage quality tests with this particular dataset.

Performance

The time taken to complete the de-duplication of the samples of our combined dataset was a monotone function of the sample size (see Fig. 2). The smallest sample of 5 million records took just under a day to complete. For the large dataset sizes, the run time slowed considerably, taking one month to complete the 20 million de-duplication linkage.

The results in Fig. 2 include the time taken to run the de-duplication of the WA CLK dataset (6.8 million records) was 2,445 minutes. When the same dataset was split into ten roughly equal parts with blocking on year of birth, the total time taken to de-duplicate was 1,828 minutes.

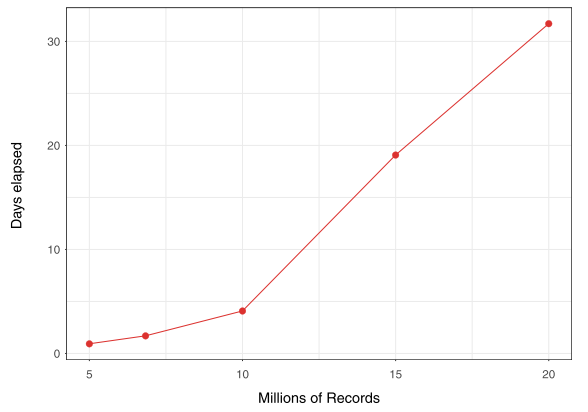


Fig. 2 Performance results for combined CLK dataset de-duplication

Discussion

Overall, the use of CLK with multibit trees for a full linkage was not as high quality as could be achieved using either unencrypted linkage or with field level Bloom filters [16]. Using the same dataset (WA Hospital), both unencrypted and field level Bloom filters had achieved an F-measure of 0.99 [15], while this measure achieved a maximum F-measure of 0.978 in our current evaluation. Overall, this difference is small, and this may be acceptable, particularly in cases where the use of a single data item for anonymous linkage is prescribed by law [18].

Our results show that the use of multibit trees for indexing/blocking of CLK data has great potential. The best recall was achieved using parameter set 1, with a value of 0.9858 at a threshold of 0.8. The unencrypted linkage on the same dataset, mentioned previously, had a recall of just 0.981, using standard blocking. The worst results for recall were for parameter sets 2 and 3, with values at all thresholds below 0.75. This is unacceptably low for any linkage, but the inclusion of all identifiers, especially with volatile address information, precludes the ability to match individuals that have changed their address. This shows, that while including more identifiers in the CLKs will usually increase the discriminative power, leading to higher precision, stable identifiers without missing data fields are needed in order to avoid sacrificing recall. While using multibit trees for indexing of CLK data has the ability for a very high coverage of possible links, its quality is ultimately determined by the identifiers used to create the CLK and the quality of the data.

In terms of performance, the linkages were reasonably slow. While operational linkages are commonly performed on an ad-hoc basis, and there are tight processing deadlines to meet, linkages which take more than a few days processing time are probably not feasible. As such, the multibit tree method, as it is currently implemented, could not be recommended for large-scale linkages. As a comparison, an unencrypted linkage of the same 20 million records can be completed within a day.

An alternate approach to using the multibit tree method may be to create a set of hashed blocking variables alongside the CLK, referred to as external blocking [26]. Our simple external blocking of the WA CLK dataset into just ten blocks based on year of birth was enough to reduce the execution time by 25%. In practice, the external blocking required to maintain linkage quality is likely to be more complex, requiring additional information alongside the CLK and may provide an additional attack vector for a malicious individual. However, external blocking provides a considerably faster method for linkage with CLKs, and at this time is a practical way for large-scale private record linkage.

Conclusion

Further testing is required to improve the CLK linkage results. One factor which is likely to improve results is the use of methods of weighting different personal identifiers based on how likely they are to identify an individual. The impact that a field has within a Bloom filter is directly proportional to how many bits that field encodes. However, in this paper, we used the baseline approach, where the number of bits was solely based on the number of bigrams in the identifier. For example, addresses usually contain many bigrams but are far less useful in identifying an individual over time when compared to date of birth or name. Testing Bloom filters which weight individual fields (by hashing bigrams more or less often) according to their usefulness in identifying individuals (discriminating power) may be an important avenue of further research.

The results reported here are heavily dependent on parameter settings. For these methods to be useful in practice, where 'truth sets' are usually not available, tried and tested parameter settings that are robust across different kinds of datasets are required. Missing values were also shown to be a major factor affecting the quality of the indexing and linkage. Since CLKs do not account for the number of identifiers for which valid information is present, calculation of similarities based on CLKs will be attenuated by asymetrically missing identifiers. However, handling missing identifiers in PPRL is a largely unexplored field of research.

Demand for privacy-preserving record linkage is increasing [39]. Security of PPRL solutions against cryptographic attacks is therefore of utmost importance in medical settings.

However, very few techniques for PPRL suitable for large data sets are available. One of these few techniques are Bloom filter-based methods for PPRL. These methods are increasingly used for a wide variety of medical research projects, such as linking mammography data [40] or building a national perinatal database [41]. State of the art variants of Bloom filter-based methods have been shown to be more resilient than competing approaches [31]. Successful attacks on these variants seem to be harder than the effort which can be expected willingly to be provided by a rational attacker [42]. Further hardening Bloom filters is subject of ongoing research by our group.

Abbreviations

CLK: Cryptographic long-term key; CHeReL: Centre for Health Record Linkage; EU: European Union; HIPAA: Health Insurance Portability and Accountability Act; NSW: New South Wales; PPRL: Privacy-preserving record linkage; PID: personal identifier; RBF: Record-level Bloom filter; WA: Western Australia; WADLB: WA Data Linkage Branch

Acknowledgements

The project acknowledges the support of data custodians and data linkage units who provided access to the jurisdictional data.

Funding

Data for the project was provided as part of a Population Health Research Network (PHRN) "Proof of Concept" collaboration which included the development and testing of linkage methodologies. The PHRN is supported by the Australian Government National Collaborative Research Infrastructure Strategy and Super Science Initiatives. AB has also been supported by an Australian Government Research Training Program Scholarship.

Availability of data and materials

The data that support the findings of this study are available from state data linkage units in NSW and WA, but restrictions apply to the availability of these data, which were used under agreement with data custodians, and, consequentially, are not publicly available.

Authors' contributions

AB and SR wrote the first draft, evaluated the clear text linkage, provided the data and ran the simulations, CB programmed the encryption and evaluation programs for PPRL and wrote parts of the final manuscript, RS devised the parameter sets and wrote parts of the final manuscript. All authors approved the final paper version.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Ethical approval for developing and refining linkage methodology, which includes blocking and linking techniques of privacy-preserved datasets, was obtained from Curtin University Human Research Ethics Committee (Reference: HR 15/2010) as well as approval from New South Wales Cancer Institute Human Research Ethics Committee (HREC/10/CIPHS/37) and Western Australian Department of Health Human Research Ethics Committee (HREC/2009/54). Ethics approval included a waiver of consent based on the criteria in the national statement on ethical conduct in human research.

Publisher's Note

Author details

[1]Centre for Population Health Research, Curtin University, Western Australia, Kent Street, Bentley, Western Australia, 6102 Perth, Australia. [2]University of Duisburg-Essen, German Record Linkage Center, Lotharstr. 65, 47057 Duisburg, Germany.

Received: 13 March 2017 Accepted: 25 May 2017
Published online: 08 June 2017

References

1. Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. Annu Rev Public Health. 2011;32(1):91–108.
2. Shah GH, Lertwachara K, Ayanso A. Record linkage in healthcare: Applications, opportunities, and challenges for public health. Int J Healthcare Delivery Reform Initiatives. 2010;2(3):29–47.
3. Herzog TN, Scheuren FJ, Winkler WE. Data Quality and Record Linkage Techniques, 1st ed. New York: Springer; 2007.
4. Smith J. The History and Future of Record Linkage in the ONS Longitudinal Study. Stat J U N Econ Comm Eur. 1999;16(3):197–205.
5. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health services research and data linkages: Issues, methods, and directions for the future. Health Serv Res. 2010;45(5 Pt. 2):1468–88.
6. Evans JMM, MacDonald TM. Record-linkage for pharmacovigilance in scotland. Br J Clin Pharmacol. 1999;47(1):105–10.
7. Maxfield MG, Weiler BL, Widom CS. Comparing self-reports and official records of arrests. J Quant Criminol. 2000;16(1):87–110.
8. Binswanger IA, Stern MF, Deyo RA, Heagerty PJ, Cheadle A, Elmore JG, Koepsell TD. Release from prison – a high risk of death for former inmates. N Engl J Med. 2007;356(2):157–65.

9. Dassanayake TL, Jones AL, Michie PT, Carter GL, McElduff P, Stokes BJ, Whyte IM. Risk of road traffic accidents in patients discharged following treatment for psychotropic drug overdose: a self-controlled case series study in australia. CNS Drugs. 2012;26(3):269–76.
10. McDonald SA, Hutchinson SJ, Bird SM, Mills PR, Dillon J, Bloor M, Robertson C, Donaghy M, Hayes P, Graham L. A population-based record linkage study of mortality in hepatitis c-diagnosed persons with or without hiv coinfection in scotland. Stat Methods Med Res. 2009;18(3): 271–83.
11. Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB. Data linkage infrastructure for cross-jurisdictional health-related research in australia. BMC Health Serv Res. 2012;12(1):480.
12. Trinckes JJ. The Definitive Guide to Complying with the HIPAA/HITECH Privacy and Security Rules. Boca Raton: CRC Press; 2013.
13. Council of European Union. Council regulation (EU) no 679/2016. 2016.
14. Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. Inform Syst. 2013;38(6):946–69.
15. Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. J Biomed Inform. 2014;50:205–12.
16. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using bloom filters. BMC Med Inform Decision Making. 2009;9(1):41.
17. Durham E, Xue Y, Kantarcioglu M, Malin B. Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. Information Fusion. 2012;13(4):245–59.
18. Deutscher Bundestag. Gesetz über Krebsregister (Krebsregistergesetz KRG). 1994. Bundesgesetzblatt Nr. 79, 11.11.1994, 3351–3355.
19. Schnell R, Bachteler T, Reiher J. A novel error-tolerant anonymous linking code. 2011. German RLC Working Paper, German Record Linkage Center.
20. Durham EA, Kantarcioglu M, Member S, Xue Y, Toth C, Kuzu M, Malin B. Composite bloom filters for secure record linkage. IEEE Trans Knowl Data Eng. 2014;26(12):2956–68.
21. Kuzu M, Durham E, Kantarcioglu M, Malin B. A constraint satisfaction cryptanalysis of bloom filters in private record linkage In: Fischer-Huebner S, Hopper N, editors. Privacy Enhancing Technologies 11th International Symposium, PETS 2011 Waterloo, ON, Canada, July 27-29, 2011, vol. 6794. Heidelberg: Springer; 2011. p. 226–45.
22. Kuzu M, Kantarcioglu M, Durham EA, Toth C, Malin B. A practical approach to achieve private medical record linkage in light of public resources. J Am Med Inform Assoc. 2013;20(2):285–92.
23. Hernández MA, Stolfo SJ. Real-world data is dirty: data cleansing and the merge/purge problem. Data Mining Knowl Discov. 1998;2(1):9–37.
24. McCallum A, Nigam K, Ungar LH. Efficient clustering of high-dimensional data sets with application to reference matching. In: Proceedings of the Sixth ACM SIGDD International Conference on Knowledge Discovery and Data Mining – KDD 2000. New York: ACM; 2000. p. 169–78.
25. Christen P. A survey of indexing techniques for scalable record linkage and deduplication. IEEE Trans Knowl Data Eng. 2012;24(9):1537–55.
26. Schnell R. An efficient privacy-preserving record linkage technique for administrative data and censuses. Stat J IAOS. 2014;30(3):263–70.
27. Rosman D, Garfield C, Fuller S, Stoney A, Owen T, Gawthorne G. Measuring data and link quality in a dynamic multi-set linkage system. In: Symposium on Health Data Linkage Proceedings 20-21 March 2002, Potts Point, Sydney, New South Wales. Adelaide: Public Health Information Development Unit; 2003. p. 184–7.
28. Lawrence G, Dinh I, Taylor L. The centre for health record linkage: a new resource for health services research and evaluation. Health Inform Manag J. 2008;37(2):60–2.
29. Brook EL, Rosman D, Holman CDJ. Public good through data linkage: Measuring research outputs from the western australian data linkage system. Aust N Z J Public Health. 2008;32(1):19–23.
30. Randall SM, Ferrante AM, Boyd JH, Semmens JB. The effect of data cleaning on record linkage quality. BMC Med Inf Decis Making. 2013; 13(1):64.
31. Schnell R, Borgs C. Randomized response and balanced bloom filters for privacy preserving record linkage. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDM 2016). Dec 12, 2016 - Dec 15, 2016. Barcelona: IEEE Publishing; 2016.
32. Niedermeyer F, Steinmetzer S, Kroll M, Schnell R. Cryptanalysis of basic bloom filters used for privacy preserving record linkage. J Privacy Confidentiality. 2014;6(2):59–69.
33. Kroll M, Steinmetzer S. Who Is 1011011111...1110110010? Automated Cryptanalysis of Bloom Filter Encryptions of Databases with Several Personal Identifiers. In: Biomedical Engineering Systems and Technologies 2015. Cham: Springer; 2015. p. 341–56.
34. Schnell R. Privacy preserving record linkage In: Harron K, Goldstein H, Dibben C, editors. Methodological Developments in Data Linkage. Chichester: Wiley; 2015. p. 201–25.
35. Kristensen TG, Nielsen J, Pedersen CN. A tree-based method for the rapid screening of chemical fingerprints. Algorithm Mol Biol. 2010;5(9):1–10.
36. Bachteler T, Reiher J, Schnell R. Similarity filtering with multibit trees for record linkage. 2013. Technical Report 1, German Record Linkage Center.
37. Karapiperis D, Verykios VS. An lsh-based blocking approach with a homomorphic matching technique for privacy-preserving record linkage. IEEE Trans Knowl Data Eng. 2015;27(4):909–21.
38. Sehili Z, Kolb L, Borgs C, Schnell R, Rahm E. Privacy preserving record linkage with PPJoin. In: Proceedings 16. GI-Konferenz Für Datenbanksysteme in Business, Technologie und Web (BTW), LNI; 2015. p. 85–104.
39. In: Gkoulalas-Divanis A, Loukides G, editors. Medical Data Privacy Handbook. Cham: Springer; 2015.
40. Schnell R, Richter A, Borgs C. A comparison of statistical linkage keys with bloom filter-based encryptions for privacy-preserving record linkage using real-world mammography data. In: 10th International Joint Conference on Biomedical Engineering Systems and Technologies (HEALTHINF 2017); Porto, 22.02.2017. Setubal: SCITEPRESS; 2017.
41. Schnell R, Borgs C. Building a national perinatal database without the use of unique personal identifiers. In: Proceedings of the 2015 IEEE 15th International Conference on Data Mining Workshop. Los Alamitos: IEEE Computer Society Press; 2015. p. 232–9.
42. Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Ganta R, Heatherly R, Malin BA. A game theoretic framework for analyzing re-identification risk. PloS ONE. 2015;10(3):1–24.

3.5.2 Privacy preserving linkage using multiple match-keys

Randall SM, **Brown AP**, Ferrante AM, Boyd JH (2019). *Privacy preserving linkage using multiple match-keys.* International Journal of Population Data Science, 4(1). https://doi.org/10.23889/ijpds.v4i1.1094

International Journal of Population Data Science (2018) 4:1:15

International Journal of Population Data Science

Journal Website: www.ijpds.org

Privacy preserving linkage using multiple match-keys

Randall, SM[1*], Brown, AP[1], Ferrante, AM[1], and Boyd, JH[1]

Submission History

Submitted:	25/10/2018
Accepted:	15/02/2019
Published:	23/05/2019

[1]Curtin University, Western Australia, Perth, Australia

Abstract

Introduction
Available and practical methods for privacy preserving linkage have shortcomings: methods utilising anonymous linkage codes provide limited accuracy while methods based on Bloom filters have proven vulnerable to frequency-based attacks.

Objectives
In this paper, we present and evaluate a novel protocol that aims to meld both the accuracy of the Bloom filter method with the privacy achievable through the anonymous linkage code methodology.

Methods
The protocol involves creating multiple match-keys for each record, with the composition of each match-key depending on attributes of the underlying datasets being compared. The protocol was evaluated through de-duplication of four administrative datasets and two synthetic datasets; the 'answers' outlining which records belonged to the same individual were known for each dataset. The results were compared against results achieved with un-encoded linkage and other privacy preserving techniques on the same datasets.

Results
The multiple match-key protocol presented here achieved high quality across all datasets, performing better than record-level Bloom filters and the SLK, but worse than field-level Bloom filters.

Conclusion
The presented method provides high linkage quality while avoiding the frequency based attacks that have been demonstrated against the Bloom filter approach. The method appears promising for real world use.

Introduction

Privacy preserving record linkage (PPRL) protocols involve determining which records from data collections describe the same individual where these records are encrypted or encoded so as to protect privacy. The challenge for these protocols is to allow for variations in data arising from missing, changed or incorrect identifiers (vital for ensuring a high level of matching accuracy) while at the same time ensuring that no information about the individuals within the dataset is revealed.

PPRL techniques typically adopt a semi-honest (also known as an *honest-but-curious*) model of security [1]. It is assumed that individual parties in the protocol will encode data as instructed and will not collude to leak information. However, parties can record and infer any available information, perform statistical frequency attacks on the data, use brute force attack techniques (such as dictionary attacks) to guess possible encoded values, or utilise other publicly available data to discover information about the encoded datasets [2].

A range of techniques for PPRL have been proposed, utilizing different methodologies, and providing different levels of privacy [1]. An important distinction lies between those protocols which utilise a party independent of the data owners (three party protocols) to conduct the linkage and those which rely solely on communications between data owners for linkage to occur (two party protocols) [1]. In protocols which utilise an independent third party, data is first encoded by the data custodians before being passed to the linkage unit, who determine which records belong to the same individual (see

*Corresponding Author:
Email Address: sean.randall@curtin.edu.au (SM Randall)

https://doi.org/10.23889/ijpds.v4i1.1094

Randall, SM et. al. / International Journal of Population Data Science (2018) 4:1:15

Figure 1). This differs from two-party protocols, where data is transferred repeatedly between the two parties (the situation is more complex when involving more than two parties). Two-party protocols are significantly more complex and require greater expertise from data custodians.

Of the proposed protocols which utilise an independent third party, two approaches are prominent in the literature, each having several variants. The first involves combining particular subsets of identifiers into a hashed key which is then used in matching (referred to as a match-key [3], linkage key [4], an anonymous linkage code [5] or the minimum linkage information approach [6]). A second approach uses a structure known as a Bloom filter to store encoded information, which allows string similarity techniques to be used across encoded data.

Anonymous linkage codes

The anonymous linkage code approach involves conducting an exact match on a pre-processed subset of personal identifiers. These identifiers are concatenated and encoded into a 'key' by which to identify an individual. Importantly, these methods use only a subset of identifiers. By creating a key using all available identifiers, any variation in records belonging to the same person (such as typographical errors) would result in those records being identified as belonging to different individuals. However, using too few identifiers can have the opposite effect, namely that separate individuals would be identified as the same person. This approach tries to use the optimum level of identifying information, allowing some error tolerance while correctly distinguishing between individuals [6].

Cryptographic hash functions are used to convert the concatenated identifiers into a fixed length encoded form. These hash functions have several important properties that make them suitable for this purpose. They are deterministic, meaning the same input will produce the same encoded output. They have the property that a small change in the data input will change the hash value extensively so that the new hash value does not appear correlated with the old hash value. They are also one-way functions, meaning that it is not feasible to determine the original input data when given only the hash value, other than by hashing guesses of the possible input and checking these against the original hash value [6]. To ensure adequate security, it is important that the hash function is used in combination with a secret cryptographic key which is sufficiently hard to guess [2]. The construction recommended for this purpose is known as a keyed-hash message authentication code, or HMAC (in this paper we generally use the term hash as shorthand for HMAC). This construction provides a secure way to combine a hash function with a secret cryptographic key [7]. This key should be shared amongst all data custodians and kept hidden from the linkage unit (see Figure 1). The use of a secret key prevents brute force attack techniques where an individual can guess values of concatenated identifiers, hash them, and check to see if they exist within the dataset.

There are several variants of the anonymous linkage code approach. In Australia, the Statistical Linkage Key-581 (SLK) [8] involves concatenating the second and third letters of an individual's first name, the second third and fifth letters of their surname, and their full date of birth and sex, into a single field. This method is regularly used to link a number of national datasets. However, in practice the SLK is typically not hashed, greatly reducing its privacy protection. The Swiss Anonymous Linkage Code involves creating a hash from phonetically encoded first and last names, along with full date of birth and sex [9]. Another variant proposed by Weber [10] concatenates and then hashes the first two letters of first and last names with date of birth and sex. A method proposed by the Office for National Statistics UK (ONS) [3] extends the anonymous linkage code through the use of multiple match-keys, each made up of different combinations of personal identifiers; a match on any match-key identifies two records as belonging to the same individual.

Hashed anonymous linkage keys combined with a secret key can provide strong privacy protection. Their weakness lies in ensuring a high level of linkage quality [6]. Single linkage keys cannot tolerate differences in the identifiers selected for matching, nor can they handle missing identifiers or utilise additional available information. A number of studies have documented the lower linkage quality found through this method [11, 12], in particular the reduced sensitivity of these methods. Procedures to improve sensitivity when using the SLK have been used in practical applications; these include the use of additional variables such as address information and the splitting of the SLK back into its component fields to carry out more fine-grained matching [13]. Such procedures reduce or remove the privacy protections provided by the method.

Privacy preserving linkage using Bloom filters

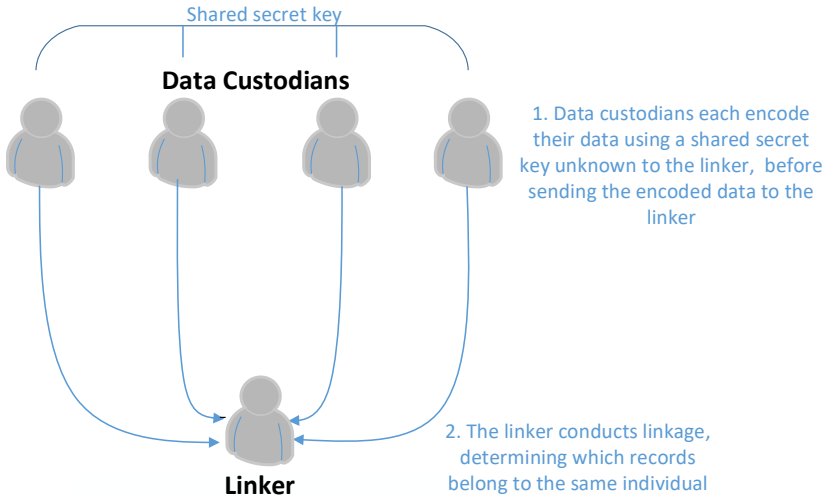
This approach involves adding encoded personal identifiers into structures known as Bloom filters (a binary array); these Bloom filters are then compared. The encoding process uses a series of hash functions (again using the HMAC construction with a secret key) to map elements of the data field to positions within the Bloom filter. The encoding process allows string similarity metrics to be used, so that variations in spelling or typographical errors can be accommodated [14].

There are two main variants of privacy preserving linkage using Bloom filters. The first is field based Bloom filters where every identifier (first name, surname etc.) is encoded into its own Bloom filter. This allows the use of techniques typical of un-encoded record linkage, such as the use of field specific weights, and the ability to appropriately handle missing values, along with previously mentioned string similarity measures. Very high linkage quality has been achieved using this method [15]. The second variant utilises a record-level Bloom filter [16] where all fields are added to a single Bloom filter which is then compared using string similarity measures. This method does not satisfactorily handle missing values [17] and uses a less sophisticated weighting approach. As such, it is likely this method yields poorer linkage quality than the field based approach, although comprehensive testing has yet to appear in print.

While Bloom filter methods have a greater tolerance for differences between records as compared to anonymous linkage codes, they also have weaker privacy protection. Both the field and record-level Bloom filter approaches have been shown to be vulnerable to frequency attacks [18-21]. While potential solutions to these attacks have been proposed [19, 22, 23], the nature of the Bloom filter method makes it difficult to ensure

Randall, SM et. al. / International Journal of Population Data Science (2018) 4:1:15

Figure 1: Privacy preserving linkage using an independent third party



that further attacks will not be found. Frequency based patterns must exist within Bloom filters; these patterns are what enables the approximate matching techniques used by these methods. It is these frequency based patterns that also make it vulnerable to attacks. While solutions to these attacks remove or hide some of these frequency patterns, such patterns must always exist, and as such, it is difficult to provide any surety regarding the existence of further attacks.

Other methods: hashing individual identifiers

Some of the earliest examples of privacy preserving record linkage involve encoding individual identifiers separately using hash functions. These encoded identifiers are then sent to the linkage unit who perform the linkage. Linkage can be carried out using standard deterministic or probabilistic methods; however, the encoding process does not allow string similarity comparisons to occur. To account for possible misspellings, techniques such as phonetic encoding can be used prior to the encoding process [24, 25].

This method of linkage has been used in Germany and France for the linkage of cancer registries. The method appears to provide high linkage quality [24, 25]; however, its weakness lies in its vulnerability to frequency attacks. For instance, it is trivial to determine the most common hash value for the 'surname' field, which will correspond to the most common surname in the population. As such, the privacy protections provided by these techniques are minimal.

An alternate approach: privacy preserving linkage using multiple match-keys

In this paper we present an alternate methodology to the approaches described above. The proposed method seeks to combine the best features of both approaches: namely, the privacy protection offered by the use of anonymous linkage code along with the linkage quality offered by the Bloom filter approach. An approach that could achieve linkage quality similar to the Bloom filter method without the associated privacy risks would be highly desirable [26].

Unlike the Bloom filter approach, our method does not make use of approximate string matching. Rather it aims to achieve high quality linkage through utilising other important techniques from traditional (un-encoded) probabilistic linkage including the use of weights and methods for managing missing values.

Methods

Overview of the protocol

Overview

The proposed method extends the anonymous linkage code approach. For each record, a number of hashes are created, each from different sets of concatenated personal identifiers; we refer to these hashes as 'match-keys'. Any pair of records with the same value for any particular match-key are identified as belonging to the same individual; that is, each match-key will directly identify an individual. Rather than use a predetermined set of match-keys, match-keys are generated based on parameters which describe the underlying characteristics of the data. These parameters are shared between the data custodians, and are used as input to the encoding process. Once encoded, data is sent to the linkage unit for linkage.

The number of match-keys and the composition of each match-key are all determined as part of this privacy preserving approach. Importantly, this composition differs depending on the characteristics of the dataset(s) in question. These parameters are identified through utilizing methods from probabilistic record linkage.

Probabilistic linkage methods

Probabilistic linkage uses conditional probabilities to compute the likelihood of two records belonging to the same person [27]. Records are compared on a pairwise basis. A comparison of two records involves comparing each field. Each field

Randall, SM et. al. / International Journal of Population Data Science (2018) 4:1:15

comparison results in a score based on whether the fields do or do not match (known as the agreement and disagreement weight, respectively). The field scores are then summed; if the summed score exceeds a specific threshold, the two records are deemed a match [27]. Field scores are calculated using two conditional probabilities, known as m and u probabilities. The *m-probability* is the likelihood that two records belonging to the same person have the same value for a particular field. The *u-probability* is the likelihood that two records belonging to different people have the same value for a particular field [27]. These are converted into agreement and disagreement weights using the following formulas:

$$\textit{Agreement Weight} = \log\left(\frac{m}{u}\right)$$

$$\textit{Disagreement Weight} = \log\left(\frac{1-m}{1-u}\right)$$

Numerous techniques exist for estimating m and u probabilities for a particular dataset and for estimating the designated threshold [28]. These include Jaro's method for estimating u-probabilities, the expectation-maximisation estimation algorithm [29] and the iterative refinement procedure first described by Newcombe [30].

Methodology

From the basic probabilistic model, it is possible to iterate through all possible combination of field state comparisons for a pair of records [28]. We will consider a simplified model, whereby a field comparison can either agree or disagree. The total number of different combinations of field state comparisons is then two to the power of the number of fields (i.e. the total number of field state comparisons doubles with the introduction of another field).

Using estimated m and u probabilities and an estimated threshold score, we can calculate the exact total score each combination of field comparisons would receive, and determine which would score above the threshold (an example is shown in Table 1) [28].

The proposed method replicates the combinations that occur above the threshold in a privacy preserved manner. The encoding process is simple; for each field state comparison, a match-key is created from hashing a concatenation of each field comparison in agreement. These hashes use the HMAC construction with a secret key shared between data providers, as in Figure 1. If one of the component fields to be concatenated is missing, the match-key is left blank. An example of this process, using example data and the combinations from Table 1, is shown in Table 2. Any two records with the same value for a particular match-key are designated a match.

Reducing the number of match-keys

The method as described creates match-keys for every field combination above the threshold. This can result in a large number of match-keys per record, creating large encoded datasets and increasing computational load. However, a great number of these created match-keys are redundant. For instance, if a match-key made up of encoded first name, surname and date of birth is considered identifying, then there is no need to also compute a match-key for first name, surname, date of birth and sex; no additional matches could possibly be found. In this way we can remove a large number of field combinations without affecting results. To identify the redundant field combinations is straightforward; given a field combination with a set of x fields marked as 'Agree' (see Table 1), any other field combinations that also contains that same set of x fields marked as 'Agree' are not required. This procedure can be applied iteratively over all field combinations to remove all redundant field comparisons (example code to encode match-keys from raw data is provided as supplementary material).

Preliminary testing suggests this method can greatly reduce the number of match-keys required. A typical example of a linkage involving nine fields produced 402 field state combinations over the given threshold; after removal of unnecessary combinations, only 41 match-keys were required.

In the next sections, we evaluate this simple method on a range of synthetic and real administrative datasets. We compare the results against those achieved with un-encoded linkage and against other privacy preserving techniques.

Evaluation methodology

Evaluation strategy

De-duplication linkages were undertaken on a range of synthetic and real administrative datasets. Each dataset had either a truth-set available (for the synthetic datasets) or a gold standard benchmark with which to compare results (for real datasets). A range of different linkage methods were compared, including both un-encoded and privacy preserving methods. The un-encoded methods included probabilistic record linkage using approximate string matching and probabilistic linkage using exact matching only. Privacy preserving methods tested comprised field-level Bloom filters, record-level Bloom filters, the SLK-581 and the multiple match-key methodology. Parameters for each linkage method were calculated using the available truth-sets and gold standard benchmarks, with results reported at the threshold providing the optimal linkage quality (where F-measure was maximised). Parameters were shared across methodologies where possible. Results were compared using the precision and recall measures described below. Algorithms were implemented in Python 2.7 [31]; linkage was conducted using the LinXmart linkage engine [32], which implements the standard probabilistic linkage methodology.

Datasets

Six separate datasets were included in the evaluation; two synthetic datasets and four real-world administrative datasets.

The two synthetic datasets 'FER12' and 'BRO17' have been used in previously published research, and detailed information on the data generation process is available [28, 33]. The FER12 dataset contained 400,000 records, of which an individual could at most have 6 duplicate records; fields included first and last name, date of birth, sex, and postcode. Each field had its own rate of errors and distribution of types of errors. The BRO17 dataset contained 1,000,000 records; the distribution of records per person was modelled on a hospital morbidity data collection with a 'long tail' where a small

Randall, SM et. al. / International Journal of Population Data Science (2018) 4:1:15

Table 1: A list of field state combinations (16 different states are possible as there are four fields)

	First Name	Surname	Sex	Year of Birth	Summed Score
1	Agree	Agree	Agree	Agree	17
2	Agree	Agree	Disagree	Agree	15.5
3	Disagree	Agree	Disagree	Agree	10
	...	...	...	...	...

Table 2: An example of the encoding process: two un-encoded records (top) are encoded (bottom) using the field state combinations from Table 1

Original Data

Record ID	First Name	Surname	Sex	Year of Birth
Record1	Sean	Randall	M	1986
Record2	John	Doe		1957

Encoded Data

Record ID	Match-Key1	Match-Key2	Match-Key3
Record1	HMAC(SeanRandallM1986)	HMAC(SeanRandall1986)	HMAC(Randall1986)
Record2		HMAC(JohnDoe1957)	HMAC(Doe1957)

number of individuals had hundreds of records per person. The BRO17 dataset had exactly 10% of fields randomly set to missing, and another 10% of fields modified in some way (by truncation, misspellings, replacement of values, etc). Fields included first name, middle name, last name, date of birth, street address, and postcode. Each of these datasets contained the 'answers', identifying which records did belong to the same individual. Both datasets are available from the authors on request.

Four large-scale Australian health datasets were also used in this evaluation; these were hospital admission records from New South Wales (NSW) and South Australia (SA), and emergency department presentations from NSW and SA. Each dataset contained all records from the three years 2008-2010; only public hospital data was available in the South Australian datasets. Each dataset had previously been de-duplicated to a high quality by jurisdictional linkage units (the Centre for Health Record Linkage and SANT Data Link for NSW and SA datasets respectively); the links created by these units were used as the gold-standard benchmark against which our de-duplication results were compared. These linkage units utilised a variety of deduplication methods including intensive manual review of created links along with quality assurance procedures to analyse and review potential errors [34]. The links created by these linkage units have been further validated through their regular use in academic and government research. The data (personal identifiers only) was made available as part of a Proof of Concept project for the Population Health Research Network [35]; ethics approvals were obtained from SA Health, the Cancer Institute NSW and Curtin University.

Each dataset contained name information (first name, middle name and surname), sex, date of birth, and address information (street address and postcode). Fields used for linkage and the percentage of missing values within each dataset are described in Table 3.

Linkage methods

Each dataset was de-duplicated using a range of linkage techniques; no linkages were conducted between any of the datasets. The same weights and blocking methods were used across linkage techniques, and multiple threshold scores were tested for each method (not all techniques required blocking, weights or thresholds). Agreement and disagreement weights were calculated directly from the available gold standard benchmark/truth-set. Two sets of blocks were used; Soundex of surname concatenated with sex, and full date of birth. This linkage strategy was based on a previously published 'default' strategy that has been regularly used in linkage evaluations [36, 37].

Probabilistic linkage was carried out using un-encoded identifiers. All available variables were used in comparisons. Two probabilistic linkages were carried out; the first used the Jaro-Winkler string similarity metric [38] for alphabetic variables (names and address) and exact matching for other variables; the second used exact matching for all variables.

Field based Bloom filters were created according to a previously published methodology [14, 15]. Bloom filters were 100 bits in length, with each variable split into bigrams that were hashed and added to the Bloom filter; three hashes were created for each bigram. The Sorenson-Dice coefficient [39] was used to compare Bloom filters. Weights and blocking fields were used as described above.

Record based Bloom filters were created based on the cryptographic long-term key construction by Schnell [16], using the weighting method described by Durham [40]. For each record, a Bloom filter of 1000 bits was created. The number of hashes computed for each bigram in each field depended

Randall, SM et. al. / International Journal of Population Data Science (2018) 4:1:15

Table 3: Number of records and percentages of missing values for each dataset

	FER12	BRO17	SA Emergency	NSW Emergency	SA Hospital	NSW Hospital
No. Records	400,000	1,000,000	813,839	4,304,459	1,007,242	6,658,380
Proportion of missing values						
First Name	2.4%	10.0%	2.2%	0.1%	3.1%	33.2%
Middle Name	-	10.0%	74.4%	83.4%	79.3%	66.9%
Surname	2.6%	10.0%	1.3%	0.0%	2.4%	33.3%
Date of Birth	11.8%	10.0%	0.0%	0.0%	0.0%	0.0%
Sex	5.2%	10.0%	0.0%	0.0%	0.0%	0.0%
Address	-	10.0%	4.6%	4.2%	7.8%	10.4%
Postcode	1.1%	10.0%	7.5%	1.2%	9.4%	0.6%

on the weight of the field, as well as the average length of the field. Address information was not added to record-level Bloom filters as preliminary testing indicated reduced linkage quality when these fields were included; previous research has also noted this issue [17]. The middle name field was also excluded due to its high proportion of missing values. Separate blocking fields were also created as described above.

The standard SLK-581 was also evaluated, created from the second and third letters of the individual's first name, the second third and fifth letters of their surname, along with full date of birth and sex [4].

For the multiple match-key algorithm, weights were used to generate field state combinations. Linkage quality was calculated on all generated match-keys over the chosen threshold. The SHA-1 hash algorithm was used with output truncated to 90 bits per hash; this provided adequate security against collisions (for 100 million unique hash values there was approximately a 1 in a trillion chance of two hashes having the same value) while reducing file sizes.

Measuring linkage quality

Linkage quality was measured using pairwise precision and recall, with the F-measure used as an overall metric of linkage quality. Results were reported at the threshold which maximised the F-measure.

Results

Linkage quality results for each tested linkage method across all six datasets are shown in Table 4; results are shown at the threshold which optimised linkage quality.

As expected, un-encoded probabilistic record linkage using approximate string matching achieved the highest linkage quality across all datasets. Generally, the use of approximate string matching as compared to exact matching resulted in minor decreases in linkage quality; this decrease was larger for the synthetic datasets, likely due to their higher rates of error.

In regards to privacy preserving techniques, field-level Bloom filters provided the highest linkage quality on all but one of the tested datasets. The multiple match-key method was the next best in terms of quality, with results only slightly below those for the field-level Bloom filters on most datasets. The record-level Bloom filters typically performed below that of the multiple match-key method, except for the SA Hospital dataset, where all three of these PPRL methods performed equally. The SLK method performed adequately on three of the four administrative datasets, however, results were lower than for all other tested methods. This method performed notably poorer for the NSW hospital dataset and both synthetic datasets due to the preponderance of missing values in these files.

One notable outlier was the results from the BRO17 dataset, where the multiple match-key method outperformed all compared methods, including un-encoded methods. We attributed this to the fact that the multiple match-key method does not require blocking; the BRO17 dataset had high levels of missing values in all fields and the standard blocking strategy was likely not appropriate here.

The number of hashes created in the multiple match-key method for each dataset varied from 14 (FER12 synthetic data) to 83 (NSW hospital data). Time taken for data encoding and linkage, and encoded file sizes (not reported but available from authors) were comparable to other evaluated methods.

Discussion

In general, the privacy preserving linkage methods evaluated here showed high linkage quality, providing continuing evidence of the viability of this method of record linkage. This was particularly apparent in datasets with few missing values or errors in identifiers, where all tested methods provided very high linkage quality.

Based on these results, field-level Bloom filters are the privacy preserving method which provides the greatest linkage quality. The high quality returned from our linkages were consistent with those achieved previously [15]. However as previously mentioned this method is vulnerable to frequency attacks [18-21] and so may not be suitable in situations where privacy protection is paramount. As expected, record-level Bloom filters performed poorly when compared against their field-level equivalents, and also performed poorly relative to the multiple match-key method introduced here.

In contrast, while the SLK method is simple to implement and can provide strong privacy protection if used appropriately (i.e. using the HMAC algorithm with a strong password), it

Table 4: Results from linkage quality evaluation

Dataset 1: FER09		Precision	Recall	F-measure
New PPRL[1] method	Multiple match-key PPRL	0.928	0.788	0.856
PPRL	SLK[2]	0.871	0.570	0.689
PPRL	Record-level bloom filter	0.937	0.778	0.850
PPRL	Field-level bloom filter	0.941	0.793	0.860
Un-encoded	Probabilistic linkage using approximate string matching	0.986	0.805	0.886
Un-encoded	Probabilistic linkage using exact matching only	0.940	0.777	0.851
Dataset 2: BRO17		Precision	Recall	F-measure
New PPRL method	Multiple match-key PPRL	0.992	0.943	0.967
PPRL	SLK	0.960	0.239	0.383
PPRL	Record-level bloom filter	0.934	0.691	0.794
PPRL	Field-level bloom filter	0.997	0.813	0.896
Un-encoded	Probabilistic linkage using approximate string matching	0.996	0.815	0.897
Un-encoded	Probabilistic linkage using exact matching only	0.993	0.810	0.892
Dataset 3: SA Emergency		Precision	Recall	F-measure
New PPRL method	Multiple match-key PPRL	0.967	0.990	0.978
PPRL	SLK	0.995	0.945	0.969
PPRL	Record-level bloom filter	0.992	0.956	0.974
PPRL	Field-level bloom filter	0.984	0.978	0.981
Un-encoded	Probabilistic linkage using approximate string matching	0.985	0.980	0.982
Un-encoded	Probabilistic linkage using exact matching only	0.969	0.990	0.979
Dataset 4: NSW Emergency		Precision	Recall	F-measure
New PPRL method	Multiple match-key PPRL	0.997	0.983	0.990
PPRL	SLK	0.999	0.966	0.982
PPRL	Record-level bloom filter	0.989	0.978	0.983
PPRL	Field-level bloom filter	0.995	0.987	0.991
Un-encoded	Probabilistic linkage using approximate string matching	0.995	0.990	0.993
Un-encoded	Probabilistic linkage using exact matching only	0.995	0.985	0.990
Dataset 5: SA Hospital		Precision	Recall	F-measure
New PPRL method	Multiple match-key PPRL	0.993	0.991	0.992
PPRL	SLK	0.975	0.988	0.981
PPRL	Record-level bloom filter	0.991	0.992	0.992
PPRL	Field-level bloom filter	0.995	0.989	0.992
Un-encoded	Probabilistic linkage using approximate string matching	0.996	0.987	0.992
Un-encoded	Probabilistic linkage using exact matching only	0.995	0.988	0.991
Dataset 6: NSW Hospital		Precision	Recall	F-measure
New PPRL method	Multiple match-key PPRL	0.983	0.991	0.987
PPRL	SLK	0.072	0.920	0.134
PPRL	Record-level bloom filter	0.754	0.921	0.829
PPRL	Field-level bloom filter	0.992	0.989	0.990
Un-encoded	Probabilistic linkage using approximate string matching	0.992	0.989	0.991
Un-encoded	Probabilistic linkage using exact matching only	0.988	0.991	0.990

[1] Privacy preserving record linkage
[2] Statistical linkage key

Randall, SM et. al. / International Journal of Population Data Science (2018) 4:1:15

does not appear suitable as an all-purpose privacy preserving linkage method, given the very poor linkage quality seen with some of the datasets. Although not tested here, we expect other anonymous linkage code methods to perform similarly to the SLK.

The multiple match-key method introduced in this paper provided admirably high linkage quality. It was superior to the SLK method, which was the only evaluated privacy preserving method with similar privacy protections. The field level Bloom filter was the only privacy preserving method to produce higher linkage quality; this method has known deficits in terms of its privacy protections [18-21]. It was not unexpected that field-level Bloom filters provided higher quality results, given their additional use of approximate string matching to identify matches; however, the associated increase was typically small in magnitude.

A key consideration in assessing the viability of the multiple match-key privacy preserving method was determining the extent to which string similarity matching (which this method does not use) contributes to high linkage quality. Previous studies comparing results using string similarity matching to those without have found large decreases in error rates for some datasets [38]. A number of publications (including those of the authors [6]) have stressed the importance of approximate matching methods for ensuring accurate privacy preserving record linkage. However, this study has found the difference between un-encoded linkages utilising approximate matching and those using only exact matching to be small, suggesting the importance of string similarity matching in ensuring quality may be overstated. The extent to which string similarity metrics improve results will clearly depend on the characteristics of the dataset in question; in an extreme example, Winkler reports a linkage in which among true-matches 20% of last-names and 25% of first names contained spelling differences [41]. Such a dataset clearly would require approximate matching techniques, and we would expect our multiple match-key method to perform poorly here. It is an open question as to what proportion of administrative datasets fall into this category.

Privacy of the multiple match-key PPRL method

The multiple match-key method presented here appears highly resistant to both dictionary and frequency attacks. Dictionary attacks are not possible through the use of a secret key in hashing (the HMAC construction) which is shared amongst data custodians and kept from the linkage unit. Frequency attacks also do not appear possible. Each particular match-key generated by this protocol is made up of a combination of fields that directly identifies an individual. If the same value of a match-key exists in two or more records, this means these records belong to the same person. As such any frequency analysis of match-keys will simply provide a list of which individuals who are found in the datasets most often, rather than provide any information on their identifiers.

The hash-based encoding process used in this protocol means that similar input values do not result in similar match-keys, a feature of the Bloom filter approach which has allowed frequency attacks to occur. As the protocol does not create match-keys if one of their component fields is a missing value, it is also not possible to perform frequency attacks of match-keys on the subset of records where particular fields are missing. The use of inappropriate match-keys (for instance, the use of the single surname field as a match-key) would allow frequency attacks to occur. This could potentially occur through human or other error. Such a match-key is not advisable not just on privacy grounds but also on quality grounds, as it would of course also result in extremely poor linkage quality (all records with the same surname would be matched together). In practice, this type of error would be easy to identify before data is encoded and sent to the linkage unit, and so is unlikely to occur.

The use of more than one match-key provides one vector by which information about the individuals can be learnt. Information is leaked when comparing two records with some match-keys matching and others not-matching. For instance, if two records have the same match-keys for combinations that do not include surname, but different match-keys for combinations that do include surname, it is likely that the surnames differ between these matching records. This can reveal information about the record in question; for instance, as it is more common for women than men to change surname in their lifetime, we could guess that this record is more likely to be female than male. While the use of multiple match-keys can leak information, it does not appear able to re-identify an individual; rather, it suggests broad demographic groupings of which a record may be part. This privacy issue is not unique to the multiple match-key method but is inherent in all privacy preserving methods which use multiple encoded values. In situations where greater privacy considerations are required such that no information about an individual can be inferred, a single match-key (i.e. the SLK approach but using the HMAC construction) is the most viable option, despite its associated reduction in linkage quality.

Strengths and limitations

The privacy preserving method presented here achieves both high accuracy and appears to provide strong privacy protection. While the absence of approximate string matching in the method may present as a limitation, our results suggest that, in general, approximate string matching provides limited quality improvement. However, for certain datasets, approximate matching will be of greater importance, such as those with very few identifiers or large numbers of typographical errors, and in these situations, we expect the multiple match-key protocol will likely perform worse than other techniques such as record-level Bloom filters.

The method proposed here is an extension of the anonymous linkage code concept to utilise more than one match-key. A similar method has been proposed by the ONS [3], although it has yet to be evaluated. A key difference is that in our method, the generation of match-keys is based on underlying characteristics of the datasets while the ONS approach uses a set of predetermined match-keys for all datasets. By generating match-keys in this way, our method will be applicable to a wider range of datasets, including those containing fields with large proportions of missing values and those with additional or alternate fields to the ones specified in the hard-coded method.

Further research is needed to investigate the performance

of the multiple match-key method (as well as the other methods) in a real-world setting, where parameters must be estimated rather than calculated. Techniques for estimating weights and thresholds necessary for the multiple match-key methodology exist and have received evaluation in privacy preserving contexts [28]. It should be noted that such parameters are normally estimated by the linkage unit at time of linkage; however, the proposed protocol requires estimation prior to data transfer, as estimated parameters are used in data encoding. A simple method to generate these parameters would be for each data custodian to compute parameters for their datasets and provide these to the linkage unit, who can then calculate a set of global parameters to be used for encoding all datasets, based on these local parameters. Additional work is required to validate such a procedure.

Conclusion

In this paper we describe and evaluate a new approach to PPRL. The results of our evaluation suggest this method can achieve very high quality results, while at the same time providing strong privacy protection.

The differing privacy preserving protocols evaluated in this paper each have their own strengths and weaknesses, and will each be suitable in particular scenarios. The multiple match-key protocol does not achieve as high a quality as field-level Bloom filters but offers greater privacy protection. It provides slightly better linkage quality in most scenarios as compared with the record-level Bloom filter approach, while providing greater certainty regarding privacy. Finally, it provides greater linkage quality than that offered by a solitary match-key such as the SLK method. As such, we feel this protocol is an important and timely contribution to the current state of the art.

Acknowledgements

We would like to thank the Ministry of Health NSW and SA Health for allowing the use of their data for this project, as well as the Centre for Health Record Linkage and SA-NT DataLink for use of their linkage keys as a benchmark in this study. This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy's Population Health Research Network.

Statement on conflicts of interest

The authors declare they have no conflict of interest.

Ethics

This study received ethical approval under the Population Health Research Network's Proof of Concept project, which included approval for developing and refining linkage methodology. Approval was obtained from Curtin University Human Research Ethics Committee as well as from New South Wales Cancer Institute Human Research Ethics Committee and South Australian Department of Health and Aging Human Research Ethics Committee.

References

1. Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. Information Systems. 2013;38(6):946-69, https://doi.org/10.1016/j.is.2012.11.005.
2. Culnane C, Rubinstein BI, Teague V. Options for encoding names for data linking at the Australian Bureau of Statistics. arXiv preprint arXiv:180207975. 2018
3. Office for National Statistics. Beyond 2011: Matching Anonymous Data. 2013. Available from: https://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-p beyond-2011-matching-anonymous-data--m9-.pdf.
4. Karmel R. Data linkage protocols using a statistical linkage key. Canberra: Australian Institute of Health and Welfare; 2005.
5. Schnell R, Richter A, Borgs C, editors. A Comparison of Statistical Linkage Keys with Bloom Filter-based Encryptions for Privacy-preserving Record Linkage using Real-world Mammography Data. Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017); 2017, https://doi.org/10.5220/0006140302760283.
6. Boyd JH, Randall SM, Ferrante AM. Application of privacy-preserving techniques in operational record linkage centres. Medical Data Privacy Handbook: Springer International Publishing; 2015. p. 267-87, https://doi.org/10.1007/978-3-319-23633-9_11.
7. Bellare M, Canetti R, Krawczyk H, editors. Keying hash functions for message authentication. Annual International Cryptology Conference; 1996: Springer, https://doi.org/10.1007/3-540-68697-5_1
8. Ryan T, Holmes B, Gibson D. A national minimum data set for home and community care. Canberra: Australian Institute of Health and Welfare; 1999.
9. Borst F, Allaert F-A, Quantin C. The Swiss solution for anonymously chaining patient files. Studies in Health Technology and Informatics. 2001(2):1239-41, https://doi.org/10.3233/978-1-60750-928-8-1239.
10. Weber SC, Lowe H, Das A, Ferris T. A simple heuristic for blindfolded record linkage. Journal of the American Medical Informatics Association. 2012;19(e1):e157-e61, https://doi.org/10.1136/amiajnl-2011-000329.
11. Randall SM, Ferrante AM, Boyd JH, Brown AP, Semmens JB. Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581? Health Information Management Journal. 2016;45(2):71-9, https://doi.org/10.1177/1833358316647587.

12. Bass J, Garfield C. Statistical linkage keys: How effective are they? Symposium on Health Data Linkage, Sydney 2002: Available online at: http://www.phidu.torrens.edu.au/pdf/1999-2004/symposium-proceedings-2003/bass.pdf; 2002. p. 40-5.

13. Karmel R, Anderson P, Gibson D, Peut A, Duckett S, Wells Y. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. BMC Health Services Research. 2010;10(41) https://doi.org/10.1186/1472-6963-10-41

14. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Medical Informatics and Decision Making. 2009;9(41) https://doi.org/10.1186/1472-6947-9-41.

15. Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. Journal of biomedical informatics. 2014;50:205-12, https://doi.org/10.1016/j.jbi.2013.12.003.

16. Schnell R, Bachteler T, Reiher J. A Novel Error-Tolerant Anonymous Linking Code. Working Paper Series No. WP-GRLC-2011-02. Nürnberg, Germany: German Record Linkage Center, 2011.

17. Brown AP, Borgs C, Schnell R, Randall SM. Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets. BMC medical informatics and decision making. 2017;17(1):83, https://doi.org/10.1186/s12911-017-0478-5.

18. Kuzu M, Kantarcioglu M, Durham E, Malin B, editors. A constraint satisfaction cryptanalysis of Bloom filters in private record linkage. Privacy Enhancing Technologies; 2011: Springer, https://doi.org/10.1007/978-3-642-22263-4_13.

19. Niedermeyer F, Steinmetzer S, Kroll M, Schnell R. Cryptanalysis of basic Bloom filters used for privacy preserving record linkage. Journal of Privacy and Confidentiality. 2014;6(2):3, https://doi.org/10.29012/jpc.v6i2.640.

20. Kroll M, Steinmetzer S. Automated Cryptanalysis of Bloom Filter Encryptions of Health Records. arXiv preprint arXiv:14106739. 2014

21. Christen P, Ranbaduge T, Vatsalan D, Schnell R. Precise and Fast Cryptanalysis for Bloom Filter Based Privacy-Preserving Record Linkage. IEEE Transactions on Knowledge and Data Engineering. 2018 https://doi.org/10.1109/TKDE.2018.2874004.

22. Schnell R, Borgs C. XOR-Folding for Bloom Filter-based Encryptions for Privacy-preserving Record Linkage. Working Paper Series No. WP-GRLC-2016-03. Nürnberg, Germany: German Record Linkage Center.

23. Schnell R, Borgs C. Randomized response and balanced bloom filters for privacy preserving record linkage. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW); 2016: IEEE, https://doi.org/10.1109/ICDMW.2016.0038.

24. Schmidtmann I, Sariyar M, Borg A, Gerold-Ay A, Heidinger O, Hense H-W, et al. Quality of record linkage in a highly automated cancer registry that relies on encrypted identity data. GMS Medizinische Informatik, Biometrie und Epidemiologie. 2016;12(1)

25. Quantin C, Bouzelat H, Allaert F, Benhamiche A-M, Faivre J, Dusserre L. How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure. International journal of medical informatics. 1998;49(1):117-22

26. Randall SM, Brown AP, Ferrante AM, Boyd JH, Semmens JB. Privacy preserving record linkage using homomorphic encryption. Population Informatics for Big Data, Sydney, Australia. 2015 https://doi.org/10.13140/RG.2.1.3052.4887

27. Newcombe HB. Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business. New York: Oxford University Press; 1988.

28. Brown AP, Ferrante AM, Semmens JB, Boyd JH, Randall SM. Estimating parameters for probabilistic linkage of privacy-preserved datasets. BMC medical research methodology. 2017;17(1):95, https://doi.org/10.1186/s12874-017-0370-0.

29. Winkler WE, editor Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. Proceedings of the Section on Survey Research Methods, American Statistical Association; 1988

30. Newcombe HB, Kennedy JM, Axford S, James AP. Automatic linkage of vital records. Science. 1959;130(3381):954-9, https://doi.org/10.1126/science.130.3381.954

31. Python Software Foundation. Python language reference, version 2.7. Python Software Foundation Wilmington, DE; 2010.

32. Curtin Data Linkage. LinXmart 2018 [Available from: www.linxmart.com.au]

33. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. Journal of Biomedical Informatics. 2012;45(1):165-72, https://doi.org/10.1016/j.jbi.2011.10.006.

34. CHeReL. Quality Assurance 2013 [Available from: http://www.cherel.org.au/quality-assurance]

35. Mitchell RJ, Cameron CM, McClure RJ, Williamson AM. Data linkage capabilities in Australia: practical issues identified by a Population Health Research Network 'Proof of Concept project'. Australian and New Zealand journal of public health. 2015;39(4):319-25, https://doi.org/10.1111/1753-6405.12310.

36. Randall SM, Boyd JH, Ferrante AM, Brown AP, Semmens JB. Grouping methods for ongoing record linkage. Population Informatics for Big Data, Sydney, Australia. 2015 https://doi.org/10.13140/RG.2.1.2003.9120

37. Randall SM, Ferrante AM, Boyd JH, Semmens JB. The effect of data cleaning on record linkage quality. BMC Medical Informatics and Decision Making. 2013;13(1):64, https://doi.org/10.1186/1472-6947-13-64.

38. Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. US Bureau of the Census. 1990

39. Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945;26(3):297-302, https://doi.org/10.2307/1932409

40. Durham EA, Kantarcioglu M, Xue Y, Toth C, Kuzu M, Malin B. Composite bloom filters for secure record linkage. IEEE transactions on knowledge and data engineering. 2014;26(12):2956-68, https://doi.org/10.1109/TKDE.2013.91.

41. Winkler WE, Thibaudeau Y. An application of the Fellegi-Sunter model of record linkage to the 1990 US decennial census. US Bureau of the Census. 1991:1-22

Abbreviations

NSW	New South Wales
ONS	Office for National Statistics UK
PPRL	Privacy preserving record linkage
SA	South Australia
SLK	Statistical linkage key

3.5.3 Privacy preserving record linkage using homomorphic encryption

Randall SM, **Brown AP**, Boyd JH, Ferrante AM, Semmens JB (2015). *Privacy preserving record linkage using homomorphic encryption* (2015) Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference.

Privacy preserving record linkage using homomorphic encryption

Sean M. Randall
Centre for Data Linkage
Curtin University
Perth, Australia
sean.randall@curtin.edu.au

Adrian P. Brown
Centre for Data Linkage
Curtin University
Perth, Australia
adrian.brown@curtin.edu.au

Anna M. Ferrante
Centre for Data Linkage
Curtin University
Perth, Australia
a.ferrante@curtin.edu.au

James H. Boyd
Centre for Data Linkage
Curtin University
Perth, Australia
j.boyd@curtin.edu.au

James B. Semmens
Centre for Population Health Research
Curtin University
Perth, Australia
james.semmens@curtin.edu.au

ABSTRACT

The bloom filter method for privacy preserving record linkage [24] has been shown to be both efficient, and provide equivalent linkage quality to that achievable with unencoded identifiers [23]. However in some situations, the bloom filter method may be vulnerable to frequency attacks, which could potentially leak identifying information [18]. In this paper we extend the bloom filter protocol to include a homomorphic encryption step which removes the vulnerability to frequency attacks. We evaluate our method by conducting a de-duplication of emergency presentation data.

Categories and Subject Descriptors

H.2.7 [**Database Management**]: Database Administration - Security, integrity, and protection

General Terms

Algorithms, Security

Keywords

Record linkage, privacy preserving record linkage, homomorphic encryption

1. INTRODUCTION

Record linkage is the process of identifying which person-based records from disparate data collections belong to the same individual. Throughout Australia, numerous operational record linkage units carry out this process, providing linked datasets to researchers, administrators and planners. Traditionally, linkage for research purposes has predominantly focused on the health sector, where it has had a significant impact on medical knowledge, and led to changes in health policy [5].

Administrative health data is highly sensitive, containing both medical and personal information collected about an individual during contact with health services and systems. The use of record linkage methods which implement privacy preserving techniques aims to satisfy privacy concerns regarding the release of named information, while allowing record linkage to take place.

Privacy preserving record linkage involves conducting record linkage on 'scrambled data', whereby records are identified as belonging to the same individual without the disclosure of personally identifying information. While these techniques provide safeguards around spontaneous recognition, they do not completely remove the privacy risk associated with large and complex datasets which are still susceptible to disclosure through unique combinations of the 'content' data.

Privacy preserving record linkage has recently become a popular area of research, with an array of protocols emerging. These protocols differ in their methods, maturity, practicality and suitability for large scale linkages. Comprehensive reviews of these methods exist in the literature [29].

1.1 Privacy preserving protocols - differences and requirements

Privacy preserving protocols can be divided into which utilise the data owners only (often known as two-party protocols) and those which include one or more independent third parties, who do not own data (often known as three-party protocols). Under a two-party protocol, only the organisations that hold data are involved in the linkage process. Under a three party model, data custodians provide encoded or encrypted data to an independent third party, which perform a specialised linkage of this data.

In Australia, when linking administrative data, the usefulness of two-party protocols appears limited. Two-party protocols require data custodians to take a substantial and ac-

This paper was presented at the First International Workshop on Population Informatics for Big Data (PopInfo'15), Sydney, 10 August 2015.

tive part in the linkage process. However, data custodians exist to manage the quality and security of their collections and linking data is not part of their core business. While custodians are often happy for their datasets to be used for linked research, they typically do not have the resources to undertake linkage themselves, and in many cases conducting linkage does not offer them any direct benefit. At the same time, there are already a number of dedicated 'third party' linkage centres around Australia with significant expertise, and the resources to undertake record linkage [13, 1, 4].

Privacy preserving protocols also differ in the level of privacy they provide. The lowest level of privacy are provided by techniques such as the statistical linkage key (SLK) [16], which simply amalgamate personally identifying attributes (like name, date of birth and gender) into one variable in clear text. The next level of privacy techniques encodes data using hash functions so that those with access cannot learn any information directly from the encoded values; however these encoded values are vulnerable to frequency attacks, which can leak personally identifying information. A final class of privacy techniques encrypts data in such a way that it is not possible to learn any information about individuals. Such methods utilise cryptographic techniques similar to those used in modern computing. Few methods such as these exist, and those that do typically require data custodians to carry out multiple computations and communication steps [29, 7, 31].

For a privacy preserving record linkage protocol to be practical, it needs to be secure, efficient and provide high linkage quality; ideally both linkage efficiency and quality would be comparable to what can be achieved with un-encoded personal identifiers. Record linkage is computationally expensive, and while tight turnaround times are not always required for record linkage processing, slower algorithms can result in impractical processing times and unworkable solutions [10]. In addition to responsive linkage services, researcher expectations also include high quality matching to ensure they can draw the correct conclusions from their research [12].

1.2 Privacy preserving record linkage using Bloom filters

A protocol for privacy preserving linkage that appears most promising utilises Bloom filters to encode data in a way that is both efficient, and allows string similarity measures (important for ensuring high linkage quality) to be computed. The use of Bloom filters for privacy preserving record linkage was first proposed by Schnell in 2009 [24]. Since then, there have been numerous variants, extensions and evaluations of this protocol [23, 25, 19, 8, 30, 15]. The method has been shown to provide similar linkage quality to that found in probabilistic record linkage with un-encoded identifiers, and to be efficient enough for large scale linkages [23].

However recent evaluations have shown this method may be vulnerable to frequency attacks; first in its original field level form [22, 19], and then later for record level Bloom filters [18]. As such, in situations where very high levels of privacy are required, this method may not be sufficient.

1.3 Objectives of this paper

In this paper we outline an extension to the generic Bloom filter protocol, which utilises a somewhat homomorphic encryption scheme that allows us to calculate a similarity metric on fully encrypted identifiers. We implement and evaluate this method on a sample of real data sourced from hospital emergency departments.

2. PROTOCOL

2.1 Overview

Our proposed protocol is a 'four party' protocol; it utilises two independent parties to conduct linkage. One has responsibility for conducting the actual linkage (the *linker*), while the second has responsibility for decrypting the similarity score of the resulting record-pairs (the *decrypter*). In our protocol, data is first encoded into Bloom filters using the methods developed by Schnell [24]. We utilise record level Bloom filters [25] (where all fields from a record are placed within a single Bloom filter) although our method would also work with field level Bloom filters. These Bloom filters are then encrypted using the system described below, again at an individual record level. This encryption will use as input a public key supplied by the decrypting third party. This two-stage encryption process (personal identifiers encoded into Bloom filters which are then encrypted) is carried out by the data custodians. It should be noted that our protocol does not limit the number of data custodians to two; any number of data custodians can be involved in the linkage.

The encrypted data is then sent to the *linker*, who conducts the required linkage. The output of this linkage (a list of the record-pairs which have been compared along with their encrypted similarity score) is then sent to the *decrypter*, who, with possession of the private key, can decrypt the similarity score. The role of the decrypter must be separate from the linker, as giving the linker access to the private key to decrypt the encrypted similarity score would also allow them to decrypt the encrypted Bloom filters. An outline of these data movements is shown in Figure 1.

2.2 Bloom filter method

A Bloom filter is a binary vector of a set length with all values initially set to zero. Using the method outlined by Schnell [24], bigrams (overlapping sets of two letters) of personal identifiers are hashed, with their modulus taken with respect to the length of the Bloom filter. The corresponding position in the Bloom filter is then set to 1. There are several variations to this method; in our implementation all personally identifying fields (i.e. first name, surname, date of birth, sex, and address) are placed within a single large Bloom filter.

Bloom filters can be compared using typical set similarity comparisons. In this implementation we focus on the dice coefficient metric, outlined in section 2.6.

2.3 Homomorphic encryption

A homomorphic encryption scheme allows computations to be carried out on encrypted data producing encrypted results; when this encrypted data is finally decrypted, the decrypted results match the results of those same operations performed on an unencrypted version of the data. While

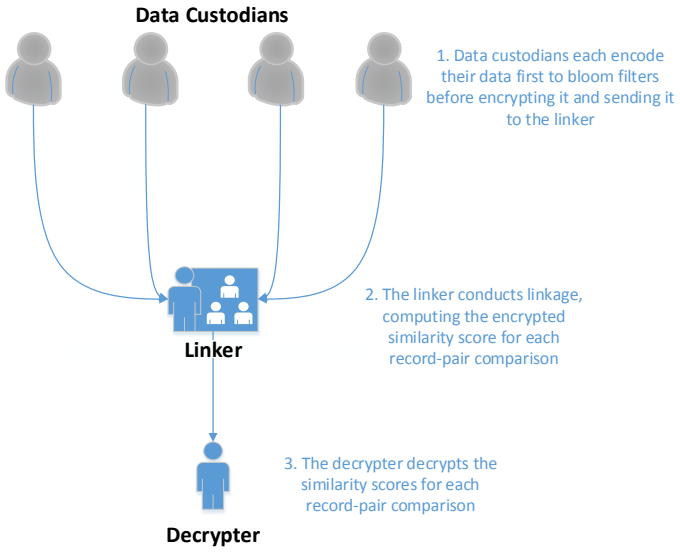


Figure 1: Data movements for the proposed protocol

homomorphic encryption protocols have existed for many years, protocols prior to 2000 only supported simple operations of either addition or multiplication. In 2009, Gentry developed the first fully homomorphic encryption system which allowed arbitrary calculations [11], and since then a large number of advances in this area have been made. However fully homomorphic systems are still too slow to be practical for most purposes [20].

Somewhat homomorphic encryption schemes only support a limited number of operations on encrypted data; however they are much faster and thus far more practical. In this paper we utilise a somewhat homomorphic encryption scheme developed by Lauter, Naehrig and Vaikuntanathan [20], along with a packing method for encrypting data developed by Yasuda [32] which allows us to compute similarity measures.

2.4 Encryption method

This scheme of Lauter, Naehrig and Vaikuntanathan [20] bases its security on the *ring learning with errors* problem. In colloquial terms, this problem is based on the difficulty of distinguishing a true signal (in this case, the secret) from noisy data. The problem, while relatively recent, is believed to be exponentially hard [20], and forms the basis for numerous modern cryptosystems [2, 21].

The scheme used in this paper allows an arbitrary number of additions of encrypted values, along with a set number of multiplications.

The system utilises several parameters. These include;

- The dimension n, which is a multiple of 2, and the corresponding cyclotomic polynomial $f(x) = x^n + 1$.

- The modulus q, a prime. Together, q, n and $f(x)$ define the rings $R \coloneqq \mathbb{Z}[x]/f(x)$ and $R_q \coloneqq R/qR = \mathbb{Z}_q[x]/f(x)$.

- The standard deviation σ of a discrete Gaussian error distribution χ.

- An integer $t < q$, which defines the message space.

Description of the algorithms key generation, encryption and decryption are given below. These are taken verbatim from Yasuda et al [32].

Key Generation We choose an element $R \ni s \leftarrow \chi$ and sample a random element $a_1 \in R_q$ along with an error $R \ni e \leftarrow \chi$. We define the public key pk as (a_0, a_1), where $a_0 \coloneqq -(a_1 \cdot s + t \cdot e)$, and we define the secret key sk as s.

Encryption For a plaintext message $m \in R_t$, with public key (a_0, a_1), the encryption samples $R \ni u, f, g \leftarrow \chi$ and computes $Enc(m, pk) = (c_0, c_1) = (a_0u + tg + m, a_1u + tf) \in (R_q)^2$, where $m \in R_t$ is considered an element of R_q.

Decryption For a ciphertext $ct = (c_0, ..., c_\xi) \in (R_q)^{\xi+1}$ (homomorphic multiplication will increase ciphertext size), with private key s, decryption is computed by $Dec(sk, ct) = [\widetilde{m}]_q mod\, t \in R_t$ where $\widetilde{m} = \sum_{i=0}^{\xi} c_i s^i \in R_q$.

2.5 Packing method

The homomorphic encryption scheme described above will allow us to encrypt individual numbers, and perform operations on these encrypted numbers. It is possible then to use the scheme to compute the dice coefficient of two Bloom filters, by first encrypting each element in the two Bloom filters individually, multiplying the elements of each position together, and summing these results. However such a scheme would be extremely slow, requiring a large number of encryptions and computations for every comparison.

Packing methods provide an alternative, allowing a vector of values to be encrypted in a single operation. Operations can then be homomorphically computed on this vector. In this work we utilise a packing method developed by Yasuda [32]. This method allows us to encrypt an entire Bloom filter (essentially a binary vector) at once, and compute its inner product using a single multiplication operation.

For a Bloom filter A of length n with elements $A_0, \ldots, A_{n-1}$ we define two packed ciphertexts.

$$ForwardPack(A) = \sum_{i=0}^{n-1} A_i x^i$$

$$BackwardPack(A) = -\sum_{i=0}^{n-1} A_i x^{n-i}$$

where Σ refers to the regular summation operator. Both of these polynomials are then encrypted as described in 2.4. Each Bloom filter is both forward and backward packed; that is, there are two encrypted values for each Bloom filter.

We can compute the inner product of two Bloom filters by multiplying one Bloom filter's forward packing by the others backward packing, as shown below.

$$ForwardPack(A) \times BackwardPack(B)$$

$$= (\sum_{i=0}^{n-1} A_i x^i) \times (-\sum_{i=0}^{n-1} B_i x^{n-i})$$

$$= \cdots - (\sum_{i=0}^{n-1} A_i B_i x^n) + \ldots$$

$$= \cdots + A \cdot B + \ldots$$

in R_t, since $x^n = -1$ with all other terms non-constant. Thus after a multiplication, upon decryption, the value of the constant term in the resulting polynomial will be our inner product.

2.6 Computing similarity measures

The most common metric used in Bloom filter similarity calculations is the dice coefficient, typically expressed as

$$Dice\ Coefficent_{A,B} = \frac{2h}{a+b}$$

where h refers to the number of positions in both bloom filters set to 1, and a and b refer to the number of positions set to 1 in bloom filters A and B respectively.

This equation can be re-written as

$$Dice\ Coefficent_{A,B} = \frac{2A \cdot B}{A \cdot A + B \cdot B}$$

where $\cdot$ refers to the inner product operation. This allows us to compute the dice coefficient using the packing method described above.

The cryptosystem employed does not allow integer division; instead, we calculate the encrypted values of the numerator and denominator separately. Both of these values are provided (encrypted) to the *decrypter* for each record pair. Once decrypted, the *decrypter* can calculate the dice coefficient from these two provided values.

2.7 Related work

Our protocol aims to allow linkage to be conducted with only the minimum participation of data custodians, and to a level of security where frequency based information is not available to the independent third parties.

There have been a number of related works published in the literature. A range of secure set intersection protocols have been proposed [26, 27, 17], many of which adopt homomorphic encryption methods to ensure security. While these methods have strong security equivalent to our protocol, they operate without the use of an independent third party, and instead require multiple communication steps from data custodians.

The closest protocol to the one described in this paper is by Kantaricioglu et al. [14], who provides a method for privacy-preserving joins utilising homomorphic encryption and two independent third parties. Similar to our work, in this protocol data custodians are only required to encrypt and transfer their data, taking no further part in the protocol. A uniquely identifying key is used to determine whether two records should be joined. A homomorphic subtraction operation is then performed when comparing individual records; where this subtraction (when decrypted) equals to 0, the two records have the same unique identifier, and so are joined.

The main difference between our method and Kantaricioglu's is that ours is aimed at the problem of record linkage, where we do not have keys which uniquely identify individuals across distinct datasets. Our proposed method tolerates the full range of 'noisy' data, utilising approximately matching techniques to handle missing values, misspellings, incorrect values and changing values over time. Previous evaluations of the approximate matching method used in our protocol have shown it to perform as well as probabilistic linkage on un-encoded identifying information [23].

3. EVALUATION

3.1 Evaluation details

We evaluated this system by performing a deduplication of 275,626 event records (one years' worth) from an emergency presentation data collection. First name, surname, date of birth, sex, address and postcode fields were used in linkage. These fields were mapped into a single 512 bit bloom filter, using weighting methods developed by Durham et al [9]. A standard blocking method was used to enable timely linkage; the date of birth field was used as the sole block.

Bloom filters were then encrypted using the encryption scheme described above. Our system utilised the parameters $n = 1024$, $\sigma = 8$, $t = 512$, and q, a 54 bit prime. These parameters were chosen to be the most efficient possible, while both ensuring correctness of results, and a security level equivalent to 128 bits; the detail of determining ac-

Table 1: Results from de-duplication of emergency presentation data

Linkage Type	Precision	Recall	F-Measure
Linkage on un-encoded identifiers	0.985	0.978	0.981
Linkage with unencrypted bloom filters	0.985	0.977	0.981
Linkage with encrypted bloom filters	0.985	0.977	0.981

curate and secure parameters is described in Lauter et al [20].

Our linkage quality results were evaluated using precision and recall measures, as recommended in the record linkage literature [6]. Efficiency and privacy were also evaluated with reference to measures described within the privacy preserving literature [28]. The emergency presentation dataset had been previously independently linked by a data linkage unit with their results made available to us. The results were used as the 'truth set' with which we compared our results.

Encryption, linkage and decryption were performed on a 64-bit Windows Server virtual machine with an Intel Xeon E5-2609 CPU at 2.4GHz, with 32GB of memory. Our implementation utilised a single core.

3.2 Results

The results for the linkage of emergency presentation data using encrypted Bloom filters, unencrypted Bloom filters, and un-encoded personal identifiers are shown in Table 1. As expected, there was no difference in quality between encrypted Bloom filters and unencrypted Bloom filters. The Bloom filter methods result in linkage quality equal to that achieved by linkage with un-encoded identifiers.

The encrypted Bloom filter linkage took slightly over 12 hours to complete, while the encryption step took 4 hours and 20 minutes, and the decryption of the answer file took almost 17 hours. A total of 1,164,305 record comparisons were performed.

In terms of individual operations, a single inner product calculation took, on average, 31 milliseconds, while encryption of a single record took 58 milliseconds, and decryption of a single record-pair took 52 milliseconds.

Our implementation was significantly slower than the more optimised implementation reported on by Yasuda et al [32]. Using equivalent parameters, our inner product calculation (i.e. our linkage) was 27 times slower, while our encryption and decryption of data was 23 and 14 times slower, respectively. While their CPU was slightly faster (Intel Xeon X3480 at 3.07GHz), the majority of this difference appears to be due to code optimisations.

In terms of privacy, using the privacy metrics of Vatsalan [28], our protocol on its own has a degree of privacy of 0.0 (absolute privacy), as all records have completely different ciphertext values. However our protocol is not complete; for efficiency, it requires a blocking component to be used in conjunction which itself may decrease privacy.

4. DISCUSSION

As expected, the linkage quality achieved through our protocol was the same as that achieved using the regular Bloom filter method, and the same as that achieved through probabilistic linkage. The advantage of the presented methodology is a far higher level of security over the Bloom filter method. This method provides a level of security equivalent to that provided by regular encryption algorithms, and removes the possibility of frequency attacks; the same plaintext value can encrypt to a very large number of ciphertext values.

By building upon the Bloom filter methods previously published, our methodology can be expected to achieve the same level of linkage quality as other Bloom filter methods. It can also leverage off the significant work already conducted to improve and refine the Bloom filter methodology, such as Durham's weighting method (used in this paper) [9].

A key limitation to our proposed method is speed. As currently implemented, our method is only suitable for small linkages. However, our naive implementation is approximately 14 to 27 times slower than the more optimised version developed by Yasuda [32]. By optimising the code used in our implementation, our method would be suitable for larger dataset sizes. Additional performance improvements could be made by using distributed computing techniques. Given the high security level of our encryption method, it may also be feasible to utilise public cloud computing resources to perform our inner product calculations, which would provide substantial potential for scalability. The blocking method used (comparing only records with the same date of birth) is relatively strict, and similarly strict blocks may be a requirement to ensure the efficiency of this method.

5. CONCLUSIONS

As far as we are aware, this is the first record linkage protocol which provides a demonstrably high level of security, without requiring numerous communication steps by data custodians. Future developments will focus on improving performance to a comparable level with that achieved by Yasuda et al [32].

This paper presents a protocol for record comparison, and does not provide any recommendations for private blocking systems. However, a private blocking scheme is necessary for a complete private linkage system. Future work will explore the use of more secure blocking methods.

Our protocol provides protection against attacks by the third or fourth party; however it does not protect against collusion by these two parties. Should these parties collude, the security of our system reduces to that of the regular privacy preserving linkage using Bloom filters (which has been evaluated previously [18]).

6. ACKNOWLEDGMENTS

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy's Population Health Research Network.

The authors would also like to thank Wenjie Lu, whose publically available code formed an initial reference point for our implementation of this cryptosystem.

7. REFERENCES

[1] J. H. Boyd, A. M. Ferrante, C. M. O'Keefe, A. J. Bass, S. M. Randall, and J. B. Semmens. Data linkage infrastructure for cross-jurisdictional health-related research in australia. *BMC health services research*, 12(1):480, 2012.

[2] Z. Brakerski, C. Gentry, and V. Vaikuntanathan. (leveled) fully homomorphic encryption without bootstrapping. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 309–325. ACM, 2012.

[3] Z. Brakerski and V. Vaikuntanathan. *Fully homomorphic encryption from ring-LWE and security for key dependent messages*, pages 505–524. Springer, 2011.

[4] E. Brook, D. Rosman, C. Holman, and B. Trutwein. Summary report: research outputs project, WA data linkage unit (1995–2003). Perth: WA data linkage unit, 2005.

[5] E. L. Brook, D. L. Rosman, and C. D. J. Holman. Public good through data linkage: measuring research outputs from the western australian data linkage system. *Australian and New Zealand Journal of Public Health*, 32(1):19–23, 2008.

[6] P. Christen and K. Goiser. Quality and complexity measures for data linkage and deduplication. In *Quality Measures in Data Mining*, pages 127–151. Springer, 2007.

[7] W. Du and M. J. Atallah. Secure multi-party computation problems and their applications: a review and open problems. In *Proceedings of the 2001 workshop on New security paradigms*, pages 13–22. ACM, 2001.

[8] E. Durham, Y. Xue, M. Kantarcioglu, and B. Malin. Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. *Information Fusion*, 13(4):245–259, 2012.

[9] E. A. Durham. *A framework for accurate, efficient private record linkage*. Thesis, 2012.

[10] A. Ferrante and J. Boyd. A transparent and transportable methodology for evaluating data linkage software. *Journal of Biomedical Informatics*, 45(1):165–172, 2012.

[11] C. Gentry. *A fully homomorphic encryption scheme*. Thesis, 2009.

[12] K. Harron, A. Wade, R. Gilbert, B. Muller-Pebody, and H. Goldstein. Evaluating bias due to data linkage error in electronic healthcare records. *BMC medical research methodology*, 14(1):36, 2014.

[13] K. A. Irvine and L. K. Taylor. The centre for health record linkage: fostering population health research in NSW. *New South Wales public health bulletin*, 22(2):17–18, 2011.

[14] M. Kantarcioglu, A. Inan, W. Jiang, and B. Malin. Formal anonymity models for efficient privacy-preserving joins. *Data & Knowledge Engineering*, 68(11):1206–1223, 2009.

[15] A. Karakasidis and V. S. Verykios. Secure blocking+ secure matching= secure record linkage. *JCSE*, 5(3):223–235, 2011.

[16] R. Karmel. *Data linkage protocols using a statistical linkage key*. Australian Institute of Health and Welfare, 2005.

[17] L. Kissner and D. Song. Privacy-preserving set operations. In *Advances in Cryptology–CRYPTO 2005*, pages 241–257. Springer, 2005.

[18] M. Kroll and S. Steinmetzer. Automated cryptanalysis of bloom filter encryptions of health records. *arXiv preprint arXiv:1410.6739*, 2014.

[19] M. Kuzu, M. Kantarcioglu, E. Durham, and B. Malin. A constraint satisfaction cryptanalysis of bloom filters in private record linkage. In *Privacy Enhancing Technologies*, pages 226–245. Springer, 2011.

[20] K. Lauter, M. Naehrig, and V. Vaikuntanathan. Can homomorphic encryption be practical? In *Proceedings of the 3rd ACM workshop on Cloud computing security workshop*, pages 113–124. ACM, 2011.

[21] V. Lyubashevsky, C. Peikert, and O. Regev. On ideal lattices and learning with errors over rings. *Journal of the ACM (JACM)*, 60(6):43, 2013.

[22] F. Niedermeyer, S. Steinmetzer, M. Kroll, and R. Schnell. Cryptanalysis of basic bloom filters used for privacy preserving record linkage. *Journal of Privacy and Confidentiality*, 6(2):3, 2014.

[23] S. M. Randall, A. M. Ferrante, J. H. Boyd, J. K. Bauer, and J. B. Semmens. Privacy-preserving record linkage on large real world datasets. *Journal of biomedical informatics*, 50:205–212, 2014.

[24] R. Schnell, T. Bachteler, and J. Reiher. Privacy-preserving record linkage using bloom filters. *BMC Medical Informatics and Decision Making*, 9(41), 2009.

[25] R. Schnell, T. Bachteler, and J. Reiher. A novel error-tolerant anonymous linking code. Report, Working Paper Series No. WP-GRLC-2011-02. Nürnberg, Germany: German Record Linkage Center, 2011.

[26] J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 639–644. ACM, 2002.

[27] J. Vaidya and C. Clifton. Secure set intersection cardinality with application to association rule mining. *Journal of Computer Security*, 13(4):593–622, 2005.

[28] D. Vatsalan, P. Christen, C. M. O'Keefe, and V. S. Verykios. An evaluation framework for privacy-preserving record linkage. *Journal of Privacy and Confidentiality*, 6(1):3, 2014.

[29] D. Vatsalan, P. Christen, and V. S. Verykios. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969, 2013.

[30] D. Vatsalan, P. Christen, and V. S. Verykios. An

efficient two-party protocol for approximate matching in private record linkage. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, pages 125–136. Australian Computer Society, Inc., 2014.

[31] M. Yakout, M. J. Atallah, and A. Elmagarmid. Efficient private record linkage. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*, pages 1283–1286. IEEE, 2009.

[32] M. Yasuda, T. Shimoyama, J. Kogure, K. Yokoyama, and T. Koshiba. *Practical packing method in somewhat homomorphic encryption*, pages 34–50. Springer, 2014.

Chapter 4

Cloud models for record linkage

Included Manuscript(s):

5. **Brown AP**, Randall SM (2020). *Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model.* JMIR Medical Informatics, 8(9), e18920. https://doi.org/10.2196/18920

Aim 4 of this thesis is to *develop a model for record linkage that retains the privacy of data and utilises the scalability of cloud computing*. This chapter addresses this aim by examining how cloud services can be harnessed to tackle the 'big data' issue within record linkage without the additional risks associated with the release of raw identifiers to cloud infrastructure.

4.1 Cloud models for data linkage

Privacy-preserving record linkage protocols can be divided into two main categories: those where the only participants are the data custodians (known as two-party protocols) and those which include one or more independent third parties, who do not own data (known as three-party protocols) but perform other functions in the data linkage process. The three-party protocol requires the data owners to send encoded or encrypted data to an independent third party to perform the linkage. Three-party protocols are more common in practice than two-party protocols. The utility of two-party protocols appears limited to jurisdictions where specialised data linkage units do not exist.

The typical three-party protocol using a trusted third-party (TTP) is shown in Figure 4.1. Data custodians send datasets containing linkage variables to the TTP. Using this information, the TTP links the datasets together (often in an enduring fashion) and creates a linkage map across the data from all custodians. This is a typical approach for data linkage in Australia, Canada and the UK. However, some datasets cannot be released due to legal, administrative and other reasons, and consequently, these are not able to participate in this process [36, 98]. The capabilities of the TTP also depend on the availability of on-premises resources. Ever-increasing volumes of data are requiring TTPs to enhance their infrastructure capabilities [3, 134].

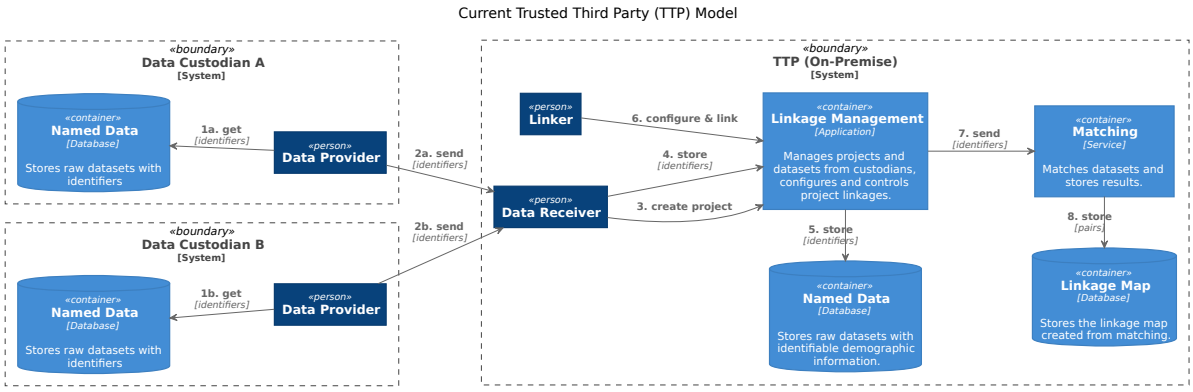


FIGURE 4.1: A trusted third-party model for record linkage

There are three main types of models for record linkage that could utilise cloud computing resources and capabilities:

- a trusted third-party hybrid cloud model,
- a trusted third-party full cloud model, and
- a two-party self-service cloud model.

These models are designed to solve specific issues. The following section describes these different models in more detail. Privacy-preserving record linkage is used on scalable cloud infrastructure for all models described below.

4.2 Trusted third-party hybrid cloud

Specialised data linkage units with insufficient computing resources and growing data sizes would benefit from a hybrid cloud model. In this model, data custodians send their data to the TTP in the usual way. The TTP stores the named identifiers locally but encodes the data into a privacy-preserved state before sending it to a cloud service for matching (see Figure 4.2). All data matching occurs on privacy-preserved datasets, and the linkage map is also stored on cloud infrastructure. This approach moves almost all computation to scalable and elastic cloud services. For linkage units, this considerably reduces the need to maintain expensive hardware on-premises. Computation is paid for, as required, and can scale with demand.

There are two direct benefits to the TTP having access to all of the named identifiers. Firstly, the data encoding process can be optimised for quality; placeholder and default values can be removed, missing values such as sex could be imputed, and data cleaning and standardisation can be consistently applied to all datasets and thoroughly validated. Secondly, once the matching process has completed, the linkage map can be validated manually with quality assurance processes already used by the TTP. Automated and batch processing for quality assurance is highly desirable; however, the ability for human inspection to validate linkage results should not be understated.

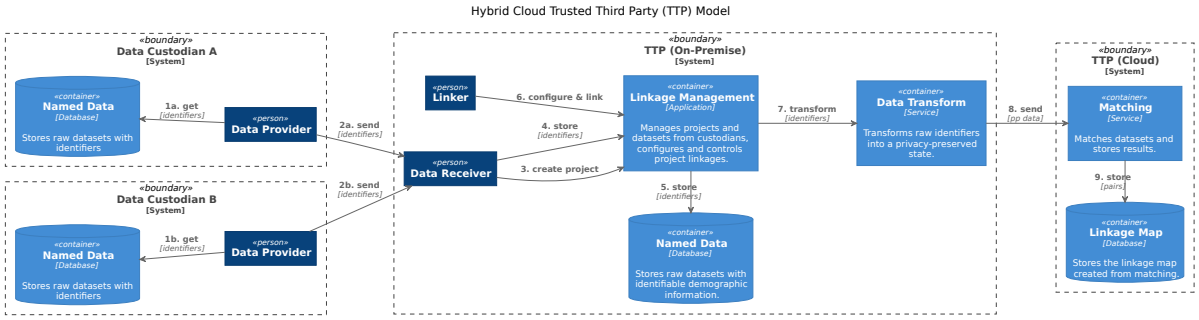


FIGURE 4.2: A hybrid cloud trusted third-party model for record linkage

The paper, *Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model*, included as part of this thesis, presents a model for record linkage that utilises cloud computing capabilities while assuring data custodians that identifiable datasets remain secure and local. This new hybrid cloud model includes privacy-preserving record linkage techniques and container-based batch processing to satisfy stated tenets. An evaluation of the model was conducted with a prototype implementation using large synthetic datasets. The results showed that an effective hybrid cloud model could be devised which extends linkage capacity. The cloud model uses PPRL techniques and moves computation to scalable cloud infrastructure.

This approach provides linkage units with the ability to process increasingly larger datasets without impacting on data disclosure issues.

4.3 Trusted third-party full cloud

A TTP full cloud model requires data custodians to encode their datasets before transmission to the trusted third party. This transmission involves uploading the encoded data directly to a cloud-based system that is managed by the specialised linkage unit, as shown in Figure 4.3. The TTP still manages the linkage system and performs the linkage activities. However, all required infrastructure is hosted in the cloud.

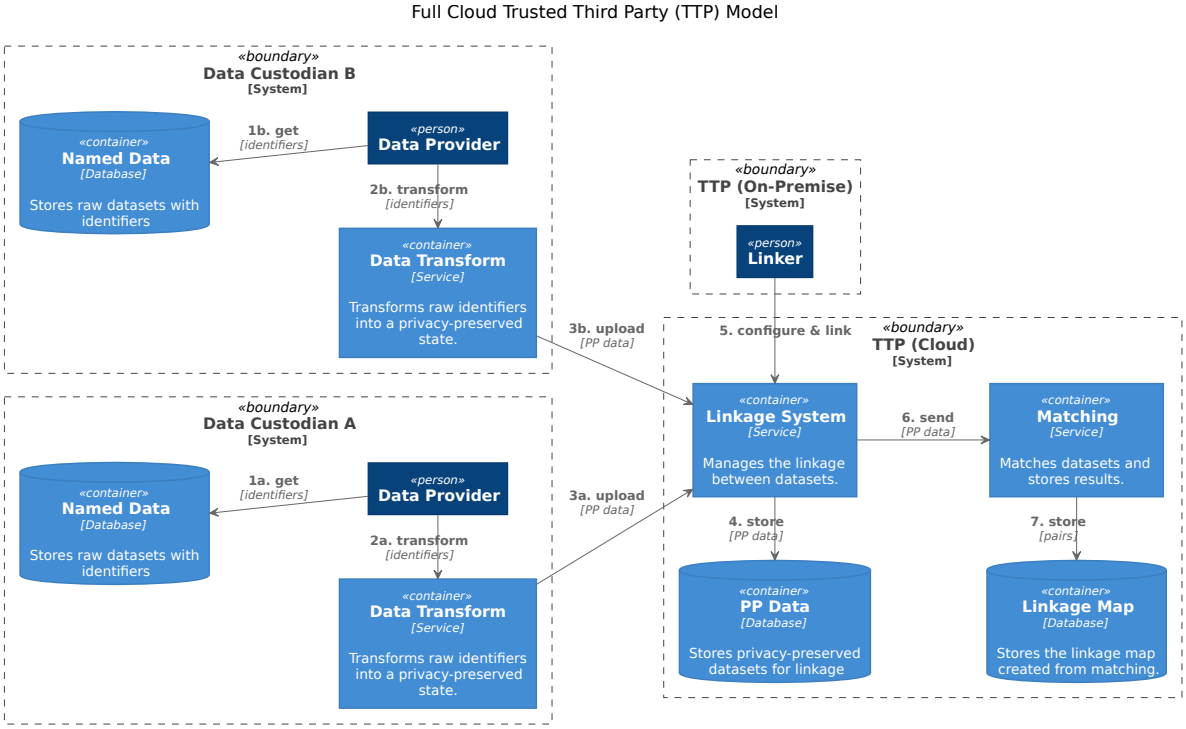


FIGURE 4.3: A full cloud trusted third-party model for record linkage

The advantage for the data custodians under this model is that all of their data is encoded prior to release, considerably reducing risks to privacy. The data that is stored and processed within the cloud environment is entirely encoded, providing an extra level of security. This model also allows those hard-to-get datasets to participate in linkage activities. One of the main disadvantages of this model, however, is the pre-processing and encoding of data which must be performed at the data custodian site. Optimising the encoding for quality at this point is much more challenging.

4.4 Self-service full cloud

A self-service model is similar to the TTP full cloud model, except one or more of the data custodians takes responsibility for the linkage activities. The linkage system is hosted within a

secure cloud environment, providing custodians with the ability to manage and participate in their own linkages. This model is shown in Figure 4.4. The cloud linkage system is effectively acting as a TTP, providing linkage as a service to data custodians.

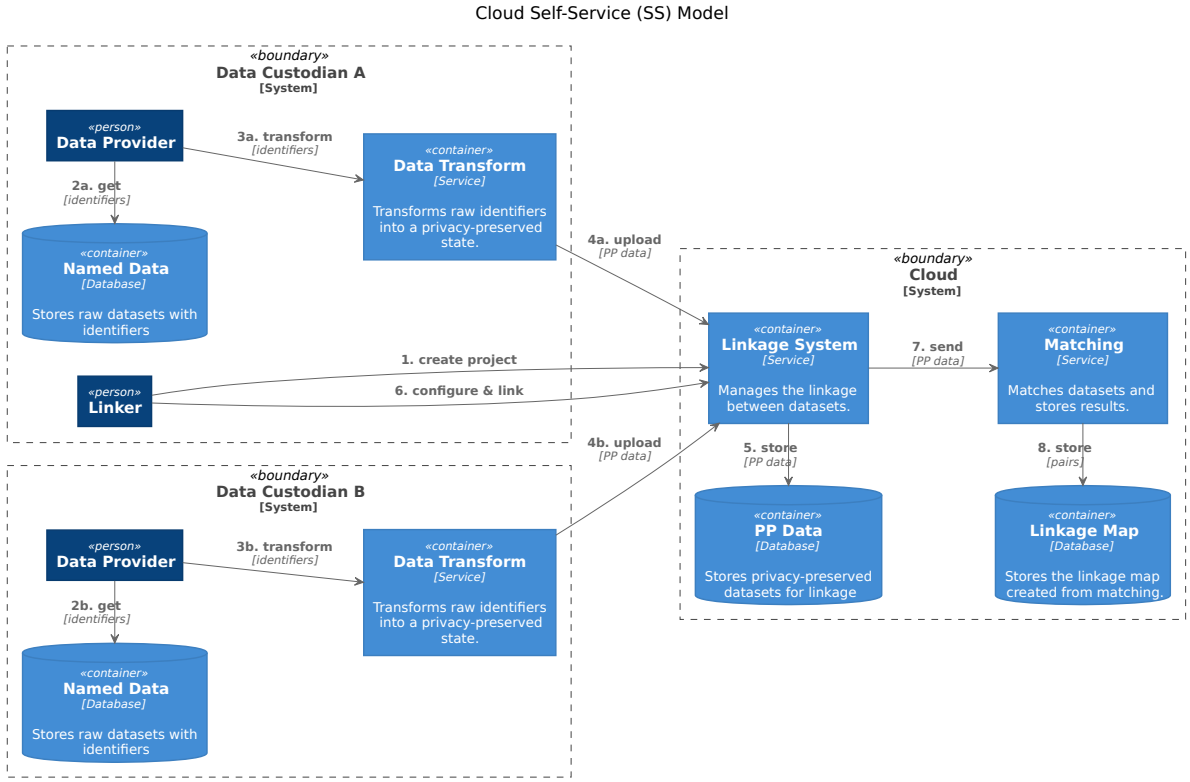


FIGURE 4.4: A self-service full cloud model for record linkage

One of the advantages of this model is providing full autonomy to data custodians. However, it is unclear how useful this model would be to those custodians who typically rely on specialised data units to undertake data linkage. A more suited use would be for linkages that involve two data linkage units, where these units are acting as secondary custodians to different datasets, looking to link their already internally linked data together. For example, cross-sectoral or cross-jurisdictional linkages could provide valuable insights for specific research questions and evidence-based decision making.

4.5 Conclusion

Different linkage scenarios can be addressed with cloud models for record linkage that specifically target that scenario. All cloud models described in this chapter utilise cloud infrastructure for the computationally intensive matching part of linkage using data encoded for privacy. While the use of a managed container cluster for linkage is described as a solution, the cloud models do not dictate the algorithm required here. Further work to refine matching algorithms so as to make full use of cloud computing infrastructure will improve the efficiency of the linkage.

The subsequent chapter describes some case studies that have used PPRL. Additionally, it describes how the cloud models presented in this chapter could fit some of those case studies. The findings from this chapter, and the previous chapters on quality, privacy and performance, are discussed in the final chapter.

4.6 Published manuscript(s)

4.6.1 Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model

Brown AP, Randall SM (2020). *Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model.* JMIR Medical Informatics, 8(9), e18920. https://doi.org/10.2196/18920

JMIR MEDICAL INFORMATICS Brown & Randall

Original Paper

Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model

Adrian Paul Brown, BSc; Sean M Randall, DPhil

Centre for Data Linkage, Curtin University, Bentley, Australia

Corresponding Author:
Adrian Paul Brown, BSc
Centre for Data Linkage
Curtin University
Kent Street
Bentley, 6021
Australia
Phone: 61 892669253
Email: adrian.brown@curtin.edu.au

Abstract

Background: The linking of administrative data across agencies provides the capability to investigate many health and social issues with the potential to deliver significant public benefit. Despite its advantages, the use of cloud computing resources for linkage purposes is scarce, with the storage of identifiable information on cloud infrastructure assessed as high risk by data custodians.

Objective: This study aims to present a model for record linkage that utilizes cloud computing capabilities while assuring custodians that identifiable data sets remain secure and local.

Methods: A new hybrid cloud model was developed, including privacy-preserving record linkage techniques and container-based batch processing. An evaluation of this model was conducted with a prototype implementation using large synthetic data sets representative of administrative health data.

Results: The cloud model kept identifiers on premises and uses privacy-preserved identifiers to run all linkage computations on cloud infrastructure. Our prototype used a managed container cluster in Amazon Web Services to distribute the computation using existing linkage software. Although the cost of computation was relatively low, the use of existing software resulted in an overhead of processing of 35.7% (149/417 min execution time).

Conclusions: The result of our experimental evaluation shows the operational feasibility of such a model and the exciting opportunities for advancing the analysis of linkage outputs.

(JMIR Med Inform 2020;8(9):e18920) doi: 10.2196/18920

KEYWORDS

cloud computing; medical record linkage; confidentiality; data science

Introduction

Background

In the last 10 years, innovative development of software apps, wearables, and the internet of things has changed the way we live. These technological advances have also changed the way we deliver health services and provide a rapidly expanding information resource with the potential for data-driven breakthroughs in the understanding, treatment, and prevention of disease. Additional information from patient devices, including mobile phone and Google search histories [1], wearable devices [2], and mobile phone apps [3], provides new opportunities for monitoring, managing, and improving health outcomes in new and innovative ways. The key to unlocking these data is in relating details at the individual patient level to provide an understanding of risk factors and appropriate interventions [4]. The linking, integration, and analysis of these data has recently been described as *population data science* [5].

Record linkage is a technique for finding records within and across one or more data sets thought to refer to the same person, family, place, or event [6]. Coined in 1946, the term describes the process of assembling the principal life events of an individual from birth to death [7]. In today's digital age, the capacity of systems to match records has increased, yet the aim

remains the same: linking records to construct individual chronological histories and undertake studies that deliver significant public benefit.

An important step in the evolution of data linkage is to develop infrastructure with the capacity to link data across agencies to create enduring integrated data sets. Such resources provide the capability to investigate many health and social issues. A number of collaborative groups have invested in a large-scale record linkage infrastructure to achieve national linkage objectives [8,9]. The establishment of research centers specializing in the analysis of *big data* also recognizes the issue of increasing data size and complexity [10].

As the demand for data linkage increases, a core challenge will be to ensure that the systems are scalable. Record linkage is computationally expensive, with a potential comparison space equivalent to the Cartesian product of the record sets being linked, making linkage of large data sets (in the tens of millions or greater) a considerable challenge. Optimizing systems, removing manual operations, and increasing the level of autonomy for such processes is essential.

A wide range of software is currently used for record linkage. System deployments range from single desktop machines to multiple servers, with most being hosted internally to organizations. The functional scope of packages varies greatly, with manual processes and scripts required to help manage, clean, link, and extract data. Many packages struggle with large data set sizes.

Many industries have moved toward cloud computing as a solution for high computational workloads, data storage, and analytics [11]. An overview of cloud computing types and service models is shown in Table 1. The business benefits of cloud computing include usage-based costing, minimal upfront infrastructure investment, superior collaboration (both internally and externally), better management of data, and increased business agility [12]. Despite these advantages, uptake by the record linkage industry has been slow. One reason for this is that the storage of identifiable information on cloud infrastructure has been assessed as high risk by data custodians. Although security in cloud computing systems has been shown to be more robust than some in-house systems [13], the media reporting of data breaches has created an impression of insecurity and vulnerability [14]. A culture of risk aversion leaves the record linkage units with expensive, dedicated equipment and computing resources that require managing, maintaining, and upgrading or replacing regularly.

Table 1. Overview of cloud computing types and service models.

Name	Description
Types of cloud computing	
Public	All computing resources are located within a cloud service provider that is generally accessible via the internet.
Private	Computing resources for an organization that are located within the premises of the organization. Access is typically through local network connections.
Hybrid	Cloud services are composed of some combination of public and private cloud services. Public cloud services are typically leveraged in this situation for increasing capacity or capability.
Service models	
IaaS[a]	The provider manages physical hardware, storage, servers, and virtualization, providing virtual machines to the consumer.
PaaS[b]	In addition to the items managed for IaaS, the provider also manages operating systems, middleware, and platform runtimes. The consumer leverages these platform runtimes in their own apps.
SaaS[c]	The provider manages everything, including apps and data, exposing software endpoints (typically as a website) for the consumer.

[a]IaaS: Infrastructure as a Service.

[b]PaaS: Platform as a Service.

[c]SaaS: Software as a Service.

To leverage the advantages of cloud computing, we need to explore operational cloud computing models for record linkage that consider the specific requirements of all stakeholders. In addition, linkage infrastructure requires the development and implementation of robust security and information governance frameworks as part of adopting a cloud *solution*.

Related Work

Some research on algorithms that address the computational burden of the comparison and classification tasks in record linkage has been undertaken. Most work on distributed and parallel algorithms for record linkage is specific to the MapReduce paradigm [15], a programming model for processing large data sets in parallel on a cluster. Few sources detail the comparison and classification tasks themselves, with the focus on load balancing algorithms to address issues associated with data skew. These works attempt to optimize the workload distribution across nodes while removing as many true negatives from the comparison space as possible [16-19]. Load balancing algorithms typically use multiple MapReduce jobs and different indexing methods to tackle the data skew problem. Indexing methods include standard blocking [17,18], density-based blocking [16], and locality sensitive hashing (LSH) [20], with varying success in optimizing the workload distribution.

Pita et al [21] have built on the MapReduce-based work and demonstrated good performance and quality using a Spark-based workflow for probabilistic linkage. Spark was chosen for

in-memory processing, ease of programming, scalability, and the new resilient distributed data set model. Like MapReduce, Spark continues to be used to address the issues with linkage and data skew on larger data sets. Spark solutions for full entity resolution are being developed, with different indexing techniques used to address workload distribution. The SparkER tool by Gagliardelli et al [22] uses LSH, meta-blocking, and a block purging process to remove high-frequency blocking keys. Mestre et al [23] presented a sorted neighborhood implementation with an adaptive window size, which uses three Spark transformation steps to distribute the data and minimize data skew.

Outside of the Hadoop ecosystem, which MapReduce and Spark are a part of, there have been some efforts to address the linkage of larger data sets through other parallel processing techniques. Sehili et al [24] presented a modified version of PPJoin, called P4Join, that can parallelize record matching on graphics processing units (GPUs), claiming an execution time improvement of up to 20 times. Despite its potential for significant improvements in runtime performance, there has not been any further work published on P4Join using larger data sets or on clusters of GPU nodes. More recently, Boratto et al [25] evaluated a hybrid algorithm using both GPUs and central processing units (CPUs) with much larger data sets. Although restricted to single (highly specified) machines, these evaluations show promise provided that the approach can be applied within a compute cluster. Again, there is not yet any further work available.

The blocking techniques used in these studies are based on the same techniques used for traditional probabilistic and deterministic linkages [15]. There are many blocking techniques used in these conventional approaches to record linkages that reduce the comparison space significantly, even when running a linkage on a single machine [26]. However, these approaches become inefficient as data set sizes become larger. They also come with a trade-off; the creation of blocks that reduce the comparisons required for linkage will inevitably reduce the coverage of true matches, resulting in more missed matches.

Much of the work in distributed linkage algorithms is focused on performance, often at the expense of linkage accuracy. Adapting these blocking techniques to distribute workload across many compute nodes has reduced the comparisons efficiently. Unfortunately, this increased efficiency has impacted the accuracy further, reducing comparisons at the expense of missing more true matches. There is still a trade-off between performance and accuracy, and further work is required to address it.

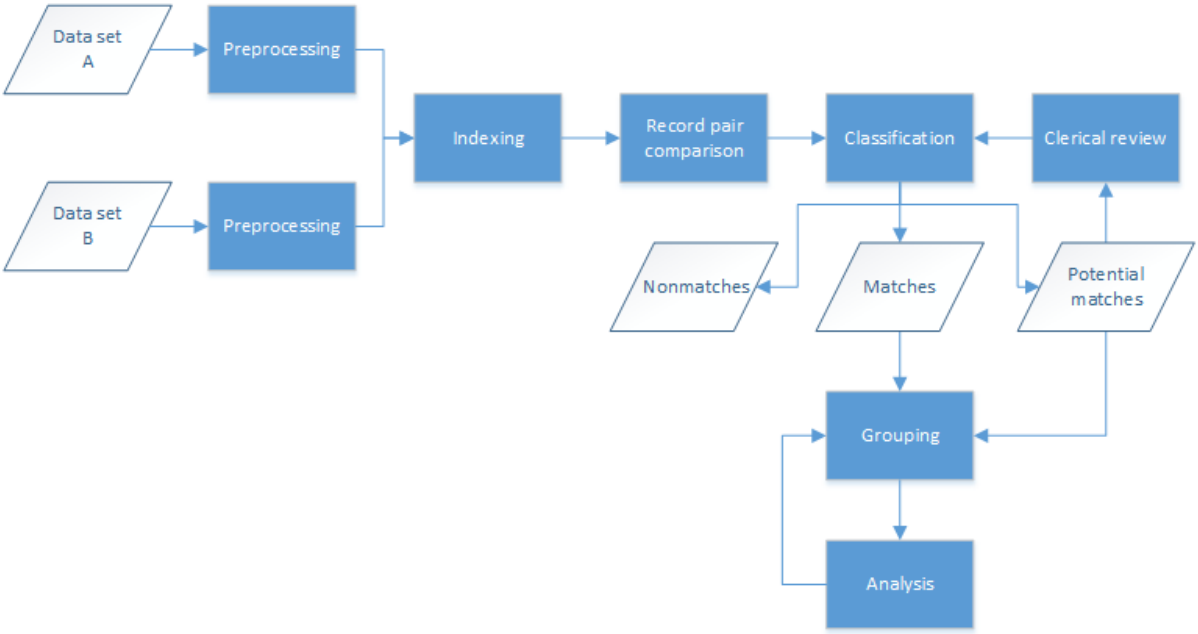
Data Flow and Release for Record Linkage

As data custodians are responsible for the use of their data, researchers must demonstrate to custodians that all aspects of privacy, confidentiality, and security have been addressed. The release of personal identifiers for linkage can be restricted, with privacy regulations such as the Health Insurance Portability and Accountability Act Privacy Rules [27] or EU regulations [28] mandating the use of encrypted identifiers. Standard record linkage methods and software are often unsuitable for linkage based on encrypted identifiers. Privacy-preserving record linkage (PPRL) techniques have been developed to enable linkage on encrypted identifiers [29]. These techniques typically use Bloom filters to store encrypted identifiers, a probabilistic data structure that can be used to approximate the equality of two sets. The emergence of these PPRL methods means that data custodians are not required to release personal identifiers. The use of PPRL methods in operational environments is still in its infancy, with limited tooling available. Available software includes the proprietary LinXmart [30], an R package called PPRL developed by the German Record Linkage Center [31], LSHDB [32], LinkIT [33], and Secure Open Enterprise Master Patient Index [34]. There is little published data on how much these systems are used outside of the organizations that created them. PPRL is a key technology that greatly opens the acceptability of cloud solutions for record linkage.

Record Linkage Process

Record linkage typically follows a standard process for the matching of two or more data sets, as shown in Figure 1. The data sets first undergo some preprocessing, a cleaning and standardization step to ensure consistency with the formatting of fields across data sets. The next step (indexing) attempts to reduce the number of record-level comparisons required (the latter often referred to as the comparison space), removing comparisons that are most likely to be false matches. The indexing step typically groups data sets into overlapping blocks or clusters based on sets of field values and can provide up to 99% reduction in the comparison space. Record pair comparisons occur next, within the blocks or clusters determined during the indexing step; this comparison step is the most computationally expensive and often requires large data sets to be broken down into smaller subsets. Classification of the record pairs into matches, nonmatches, and potential matches results in groups of entities (or individuals) based on the match results. Potential matches can be assessed manually or through special tooling to determine whether they should be classified as matches or nonmatches. A common approach to grouping matches is to merge all records that link together into a single group; however, different approaches can be used to reduce linkage error [35]. Analysis of the entity groups is the last step, where candidate groups are clerically reviewed to determine if and how the records in these groups should be regrouped.

Figure 1. Typical data matching process.



This paper presents 2 contributions to record linkage. First, it offers a model for record linkage that utilizes cloud computing capabilities while providing assurance that data sets remain secure and local. Lessons learned from many real-world record linkage projects, including several PPRL projects, have been instrumental in the design of this cloud model [30,36,37]. Second, the use of containers to distribute linkage workloads across multiple nodes is presented and evaluated within the cloud model.

Methods

Design of a Cloud Model for Record Linkage

The standard record linkage process relies on one party (known as the trusted third party [TTP]) having access to all data sets. Handling records containing identifiable data requires a sound information governance framework with controls in place that manage potential risks. Even with a well-managed information security system in place, access to some data sets may still be restricted. The TTP also requires infrastructure that can help manage data sets, matching processes and linkage key extractions over time. As the number and size of data sets grow, the computational needs and storage capacity must grow with it. However, the computation requirements for data linkage are often sporadic bursts of intense workloads, leaving expensive hardware sitting idle for extended periods.

Dedicated data linkage units in government and academic institutions exist across Australia, Canada, and the United Kingdom, acting as trusted third parties for data custodians. These data linkage units were established from the need to link data for health research at the population level. Some data linkage units are involved in the linkage of other sectors such as justice; however, the primary output of these organizations is linked data for health research. It is essential that a cloud model for record linkage takes into account the linkage practices and processes that have been developed by these organizations.

Our cloud model for record linkage addresses the limitations of data release and the computational needs of the linkage process. Data custodians and linkage units retain control of their identifiable information, while the matching of data sets between custodians occurs within a secure cloud environment.

Tenets of the Record Linkage Cloud Model

The adopted model was founded on 3 overarching design principles:

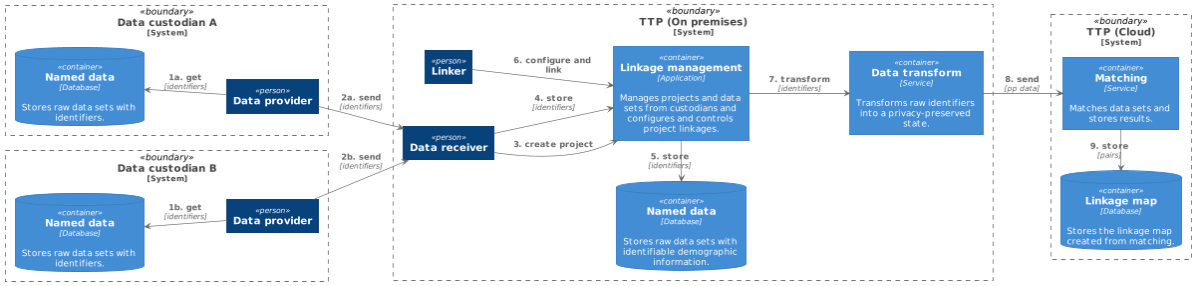
1. *The privacy of individuals in the data is protected.* One of the most important responsibilities for data custodians and linkage units is information security. Data sets contain private, and often sensitive, information on people, and it is vital that appropriate controls are in place to mitigate any potential risks. Some data sets have restrictions on where they can be held, requiring them to be kept local and protected. All computation and storage within the cloud infrastructure must be done on privacy-preserved versions of these data sets.
2. *Computation and storage are outsourced to* the cloud infrastructure. Computation requirements for data linkage are often sporadic bursts of intense workloads, followed by periods of low use or even inactivity. The ability to provision resources for computation as and when required means you only pay for what you use. This computation is generally associated with large sets of input and output data, so it makes sense to keep these data as close to the computation as possible. Storage may not necessarily be cheap, but many cloud computing providers guarantee high levels of durability and availability, with encryption and redundancy capabilities.
3. *Cloud platform services are used over infrastructure services.* Once data are stored within a cloud environment,

additional Platform as a Service offerings for analysis of the data should be leveraged. These are managed services over the top of infrastructure services (such as virtual machines) and can be started and stopped as needed.

High-Level Architectural Model

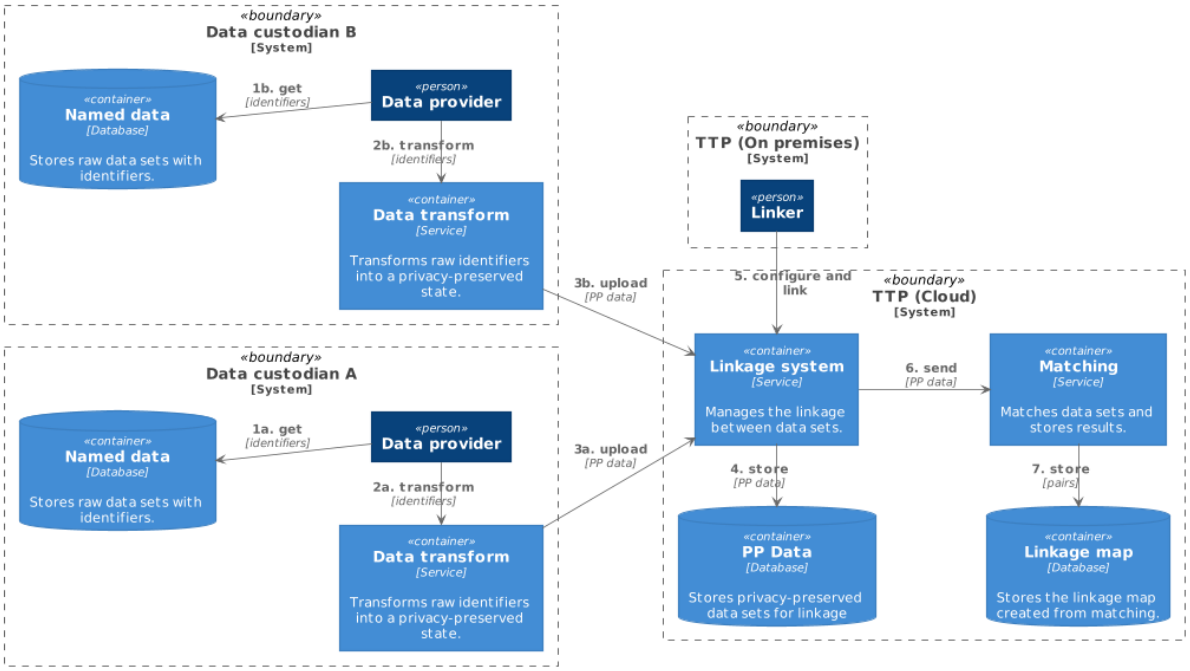
Not all storage and computation can be performed within a cloud environment without impacting privacy; the storage of raw identifiers (such as name, date of birth, and address) must often remain on-premises. The heavy-computational workloads for record linkage, the record pair comparisons and classification, are therefore undertaken on privacy-preserved versions of these data sets. These privacy-preserved data sets must be created on premises and uploaded to cloud storage. The remainder of the linkage process continues within the cloud environment. However, some parts of the classification and analysis steps may be done interactively by the user from an on-premises client app, annotating results from cloud-based analytics with locally stored details (ie, identifiers). An overview of the components and data flows involved in the hybrid TTP model is shown in Figure 2. This model satisfies our cloud model tenets and provides the linkage unit with the ability to scale their infrastructure on-demand. The matching (classification) component can utilize scalable platform services available by the cloud provider to match large privacy-preserved data sets as required. All major cloud providers have platform services that can provide computation on-demand for the processing of big data. The linkage map persists as it contains no identifiable information and can also be analyzed using available cloud platform services.

Figure 2. Hybrid cloud trusted third party model. PP: privacy-preserved; TTP: trusted third party.



Keeping identifiers at the data custodian level (on-premises) while matching on privacy-preserved data within cloud infrastructure enables linkages of data sets *between* data custodians. This model does not require any raw identifiers to be released, and thus, a hybrid model is no longer necessary. The TTP can then be hosted fully in the cloud, as shown in Figure 3. There are 2 immediate ways to achieve this: either one of the custodians manages the cloud infrastructure themselves or an independent third party controls it and provides it as a service to all custodians. A custodian could act as a TTP for all custodians involved in the linkage if this is acceptable to the parties involved. Otherwise, it may be more amenable to go with an independent TTP.

Figure 3. Full cloud trusted third party model. PP: privacy-preserved; TTP: trusted third party.



Although the full cloud TTP model may be useful in some situations, it is unlikely that this would be a desirable model with the dedicated data linkage units. Processes in cleaning, standardization, and quality analysis with personal identifiers have developed and matured over many years. Switching to a model where they no longer have access to personal identifiers would affect the accuracy of the linkage and ultimately the quality of the health research that used the linked data. The hybrid model replaces only the matching component, allowing many existing linkage processes to remain.

Scaling Computation-Heavy Workloads

Record pair comparison and classification tasks are the most computationally intensive tasks in the linkage process, although they are heavily affected by the indexing method used. The single process limitation of most linkage apps makes it difficult to cater to increasingly large data sets, regardless of indexing. Increasing memory and CPU resources for these single-process apps provides some ability to increase capacity, but this may not be sustainable in the longer term.

Although MapReduce appears to be a promising paradigm for addressing large-scale record linkage, 2 issues emerge. First, they consider only the creation of record pairs, whether matches or potential matches, without any thought as to how these record pairs are to come together to form entity groups. The grouping task is also an important part of the data matching process, and the grouping method used can significantly reduce matching errors [35]. Second, MapReduce algorithms do not appear to be readily used, if at all, within an operational linkage environment. Organizational change can be slow, and there is much investment in the existing matching algorithms and apps currently used. It may be operationally more acceptable to continue using these apps where possible.

The comparison and classification tasks of the record linkage process are an embarrassingly parallel problem if the indexing task can produce disjoint sets of record pairs (blocks) for comparison. With the rapid uptake of containerization and the availability of container management and orchestration capability, a viable option for many organizations is to reuse existing apps deployed in containers and run in parallel. Matching tasks on disjoint sets can be run independently and in parallel. The matches and potential matches produced by each matching task can, in turn, be processed independently by grouping tasks. The number of sets that are run in parallel would then only be limited by the number of container instances available.

Indexing solutions are imperfect on real-world data; however, producing disjoint sets for matching is difficult without an unacceptable drop in *pairs completeness* (a measure of the coverage of true positives). There is inevitably some overlap between blocks, as multiple passes with different blocking keys are typically used to ensure accurate results. This overlap prevents independent processing and can be handled in 1 of the 2 ways: (1) the blocks of pairs for classification can be calculated in full before duplicates are removed and the classification task can be run or (2) duplicate matches and potential matches are removed following the classification task. The main disadvantage of option 1 is that this requires a potentially massive set of pairs to be created upfront, as the comparison space is typically orders of magnitude larger than the set of matches and potential matches. Many linkage systems combine their indexing and classification tasks for efficiency, and it is often easier to ignore duplicate matches until completion. The disadvantage of option 2 is that overlapping block sets result in overlapping match sets, preventing the independent grouping of matches from each classification task.

Regardless of the indexing method used to reduce the comparison space for matching, the resulting blocks require grouping into manageable size bins that can be distributed to parallel tasks. A *bin*, therefore, refers to a subset of record pairs grouped together for efficient matching. Block value frequencies are calculated across data sets and used to calculate the size of the total comparison space. Records from these data sets are then copied into separate bins such that each bin has a comparison space of approximately equal size to every other bin.

Using this method, the comparison and classification of each bin are free to be executed on whatever compute capability is available. A managed container cluster is an ideal candidate; however, the container's resources (CPU, memory, and disk) and the bin characteristics (eg, maximum comparison space) need to be carefully chosen to ensure efficient resource use.

Development and Experimental Evaluation of the Prototype

An evaluation of the hybrid cloud linkage model was conducted through the deduplication of different sized data sets on a prototype system. The experiments were designed to evaluate parallel matching using an existing matching app on a cluster of containers; to measure encryption, transfer, and execution times; and to assess the remote analysis of the matching pairs created.

A prototype system was developed with the on-premises component running on Microsoft Windows 10 and the cloud components running on Amazon Web Services (AWS). The prototype focused on the matching part of the linkage model and utilized platform services where available. These services are described in Table 2.

Table 2. Amazon Web Services used.

AWS[a]	Description
S3	Provides an object (file) storage service with security, scalability, and durability.
Glue	A fully managed extract, transform, and load service, providing table definition, schema discovery, and cataloging. Used in conjunction with S3 to expose cataloged files to other AWS services.
Step function	A managed state machine with workflows involving other AWS. The output of a step that uses a particular service can then be used as the input for the next step.
Batch	A fully managed service for running batches of compute jobs. Compute resources are provisioned on-demand.
Athena	An interactive query service for analyzing data in S3 using standard Structured Query Language.

[a]AWS: Amazon Web Services.

Test Data

Three synthetic data sets were generated to simulate population-level data sets: 7 million records, 25 million records, and 50 million records. Although 7 million records may not necessarily represent a large data set, a 50 million record data set is challenging for most linkage units. The data sets were created with a deliberately large number of matches per entity to increase the comparison space and to challenge the matching algorithm.

Data generation was conducted using a modified version of the Febrl data generator [38], an open-source data linkage system written in Python. Frequency distributions of the names and dates of birth of the population of Western Australia were used to generate the synthetic data sets. Randomly selected addresses were sourced from Australia's National Address File, a publicly available data set [39]. Each data set contained first name, middle name, last name, date of birth, sex, address, and postcode fields. Each field had its own rate of errors and distribution of types of errors. These were based on previously published synthetic data error rates, deliberately set high to challenge matching accuracy [40]. Type of errors included replacement of values, field truncation, misspellings, deletions, insertions, use of alternate names, and values set to missing. Records had anywhere between zero to many thousands of duplicates within the data sets.

All available fields were used for matching in a probabilistic linkage. Two separate blocks were used: first name initial and last name Soundex, and date of birth and sex. Each pair output from the matching process included two record IDs, a score, a block (strategy) name, and the individual field-level comparison weights used to calculate the score.

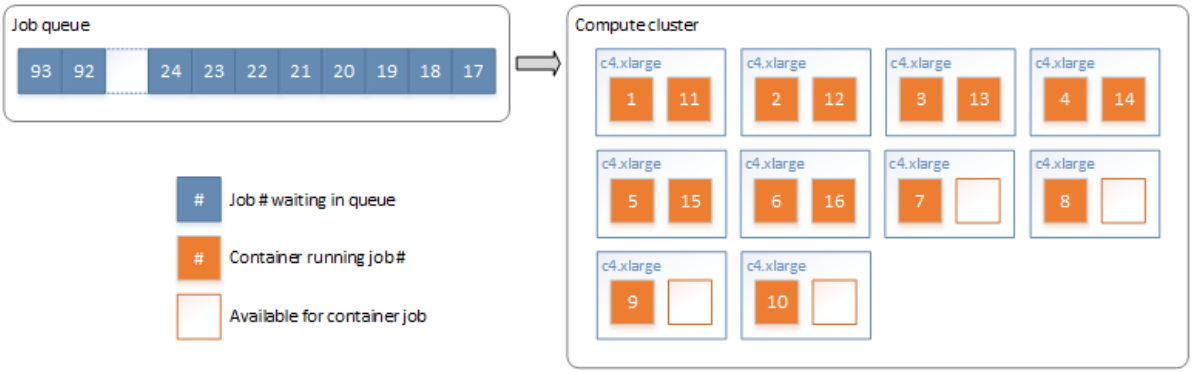
Experiments

The on-premises component first transformed data sets containing named identifiers into a privacy-preserved state using Bloom filters. String fields were split into bigrams that were hashed 30 times into Bloom filters 512 bits in length. Numeric fields (including the specific date of birth elements) were cryptographically hashed using hash-based message authentication code Secure Hash Algorithm 2 (SHA2). These privacy-preserved data sets were compressed (using gzip) before being uploaded to Amazon's object storage, S3. A configuration file was also uploaded, containing the necessary linkage parameters required for the probabilistic linkage. An AWS step function (a managed state machine) was then triggered to run through a set of tasks to complete the deduplication of the file as defined in the parameter file.

All step function tasks used on-demand resource provisioning for computation. A compute cluster managed by AWS Batch was configured with a maximum CPU count of 40 (10×c4.xlarge instance type). Each container was configured with 3.5 GB RAM and 2 CPUs, allowing up to 20 container instances to run at any one time.

The first task ran as a single job, splitting the file into many bins of approximately equal comparison space, using blocking variables specified in the configuration file. By splitting on the blocking variables, the comparison space for the entire linkage remains unchanged. Each bin was stored in an S3 location with a consistent name suffixed with a sequential identifier. The second task ran a node array batch job, with a job queued for each bin to run on the compute cluster. Docker containers running a command-line version of the LinXmart linkage engine were executed on the compute nodes to deduplicate each bin independently. AWS Batch managed the job queue, assigning jobs to available nodes in the cluster, as shown in Figure 4.

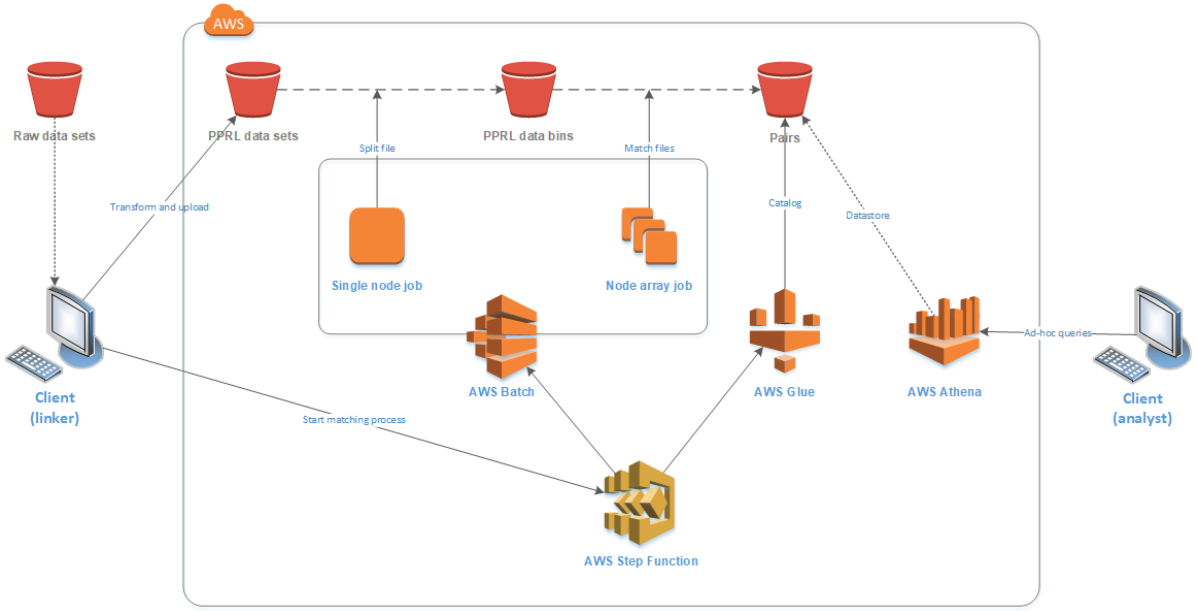
Figure 4. Matching jobs running on compute cluster (one job per bin).



LinXmart is a proprietary data linkage management system, and the LinXmart linkage engine was used because of our familiarity with the program and its ability to run as a Linux command-line tool. It accepts a local source data set and parameter file as inputs and produces a single pairs file as output. There were no licensing issues running LinXmart on AWS in this instance, as our institution has a license allowing unrestricted use. This linkage engine could be substituted, if desired, for others that similarly produce record pair files. The container was bootstrapped with a shell script that downloaded and decompressed the source files from S3 storage, ran the linkage engine program, and then compressed and uploaded the resulting pairs file to S3 storage. Each job execution was passed a sequential identifier by AWS Batch, which was used to identify a source bin datafile to download from S3 and mark the resulting pair file to upload to S3.

The third step function task classified and cataloged all new pairs files, using AWS Glue, making them available for use by other AWS analytical services. The results for each original data set were then able to be presented as a single table, although the data itself were stored as a series of individual text files. The prototype's infrastructure and data flow are shown in Figure 5.

Figure 5. Prototype on Amazon Web Services. AWS: Amazon Web Services; PPRL: privacy-preserved record linkage.



Once the deduplication linkages were complete, the on-premises component of the prototype was employed to query each data set's pair table. The queries were typical of those used following a linkage run: pair count, pair score histogram, and pairs within a pair score range. This query component used the AWS Athena application programming interface (API) to execute the queries, which used Presto (an open-source distributed query engine) to apply the ad hoc structured query language queries to the cataloged pairs tables.

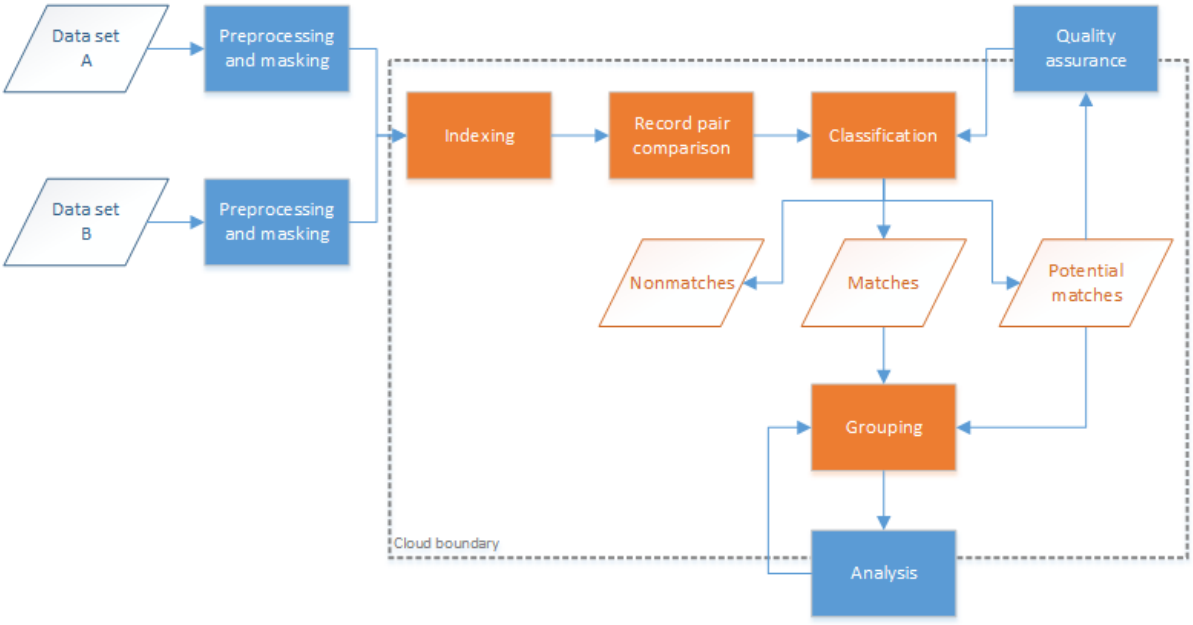
Results

Design of a Cloud Model for Record Linkage

The cloud model data matching process is shown in Figure 6. Essentially, every step in the record linkage process from indexing to group analysis is pushed to cloud infrastructure. Preprocessed data sets are transformed into a privacy-preserved state (masking) and uploaded to the cloud service for linking. The services within the cloud boundary now act as a TTP. The quality assurance and analysis steps sit on the boundary of the

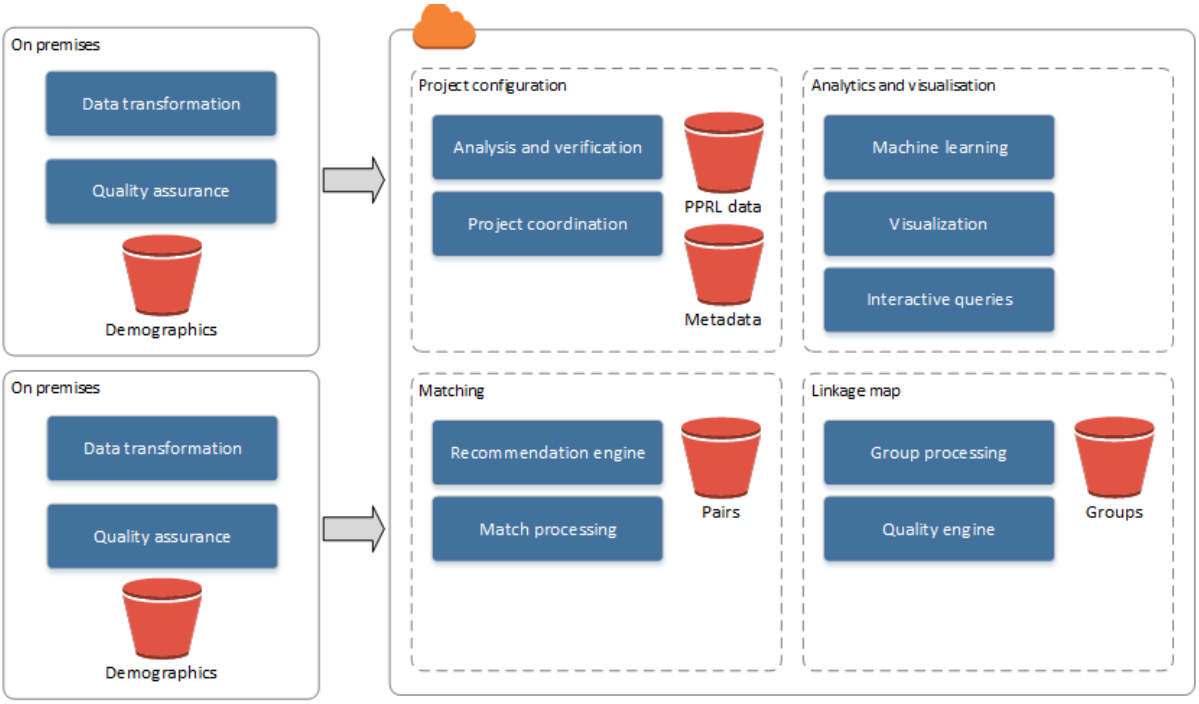
cloud as computation and query occur on the hosted cloud infrastructure, but the interactive analysis is performed by the analysis client on premises. If the analysis client has access to one or more of the raw data sets used in the linkage, these data can be annotated onto query results, giving the clerk more informed decisions and an experience to which they are accustomed.

Figure 6. New cloud model data matching process.



As shown in the high-level architectural model in Figure 7, the demographic data (containing personal identifiers) continue to remain on premises with the data custodian. Responsibilities of the data custodians are limited to data transformation and quality assurance management. The responsibilities of the cloud services are covered under 4 main categories: project configuration, matching, linkage map, and analytics and visualization. The project configuration includes the services required for coordinating projects within and across separate data custodians. Privacy-preserved data sets are stored here as well as metadata on the data sets as a result of analysis and verification performed on the uploaded data sets. The matching category includes all match processing (classification) and pairs output as well as services for providing recommendations on linkage parameters (such as *m* and *u* likelihood estimates for probabilistic linkages) for linkages between privacy-preserved data sets [41]. The linkage map category holds the entity group information, the map between individual records, and the group in which they belong. This category also contains services for processing and creating groups from pairs as well as quality estimation and analysis. Analytics and visualization contain all analytical services provided to the on-premises clients.

Figure 7. High-level architecture of record linkage cloud model. PPRL: privacy-preserved record linkage.

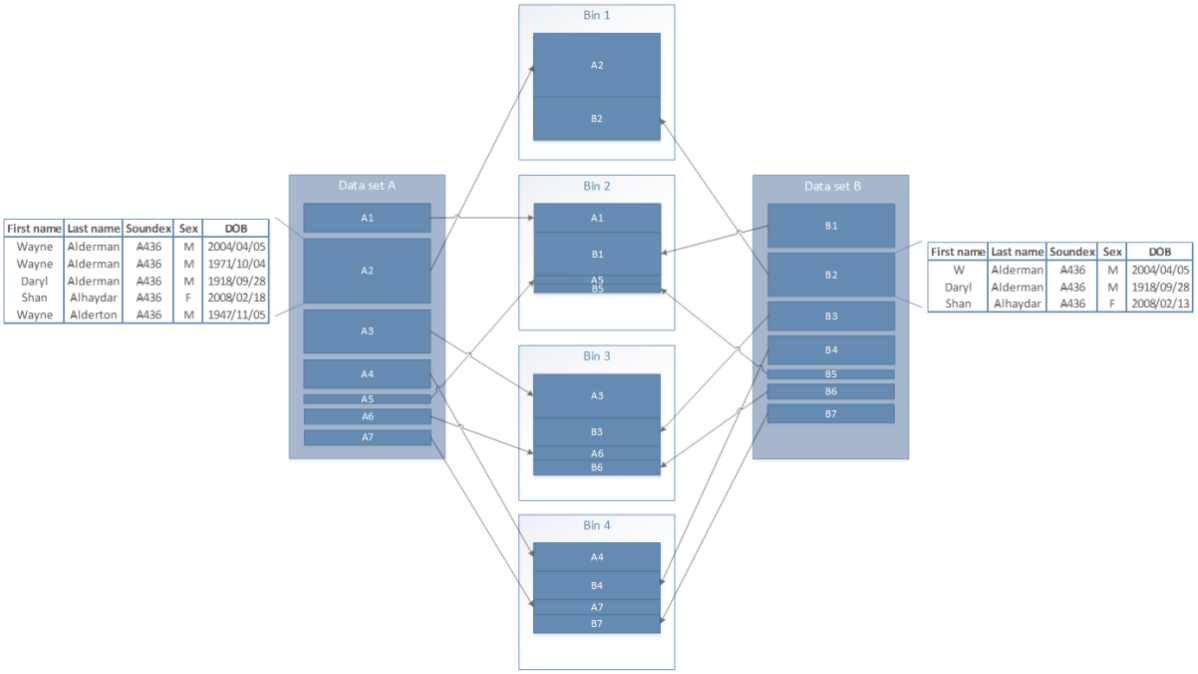


This model also allows computation to be pushed onto inexpensive, on-demand hardware in a privacy-preserving state while retaining the advantage of seeing raw identifiers during other phases of the linkage process (eg, quality assurance and analysis).

Experimental Evaluation of the Prototype

Each deduplication consisted of a single node job to split the data set into multiple bins, followed by a node array job for the matching of records within each bin. The split of data into bins is shown in Figure 8. In this example, all records with the same Soundex value will end up in the same bin.

Figure 8. Datafiles split into independent bins (by Soundex block values) for matching. DOB: date of birth.



The total comparison space was calculated using the blocking field frequencies in the data set. These frequencies represent the number of times each blocking field value occurs in the data, providing the ability to calculate the number of comparisons that will be performed for each blocking field value. First, the comparison space for each blocking field was

calculated using the frequency of the value within the file. The total comparison space was the sum of each, and the bin count was determined by dividing this by the maximum desired comparison space for a single bin. The blocking field value with the largest comparison space was assigned to the first bin. The blocking field value with the next largest comparison space was assigned to the second bin. This process continued for each blocking field value, returning to the first bin when the end was reached. A file was created for each bin, which was then independently deduplicated. Blocking field values with a very high frequency are undesirable as they are usually less useful for linkage and are costly in terms of computation. Any blocking field value with a frequency higher than the maximum desired comparison space was discarded.

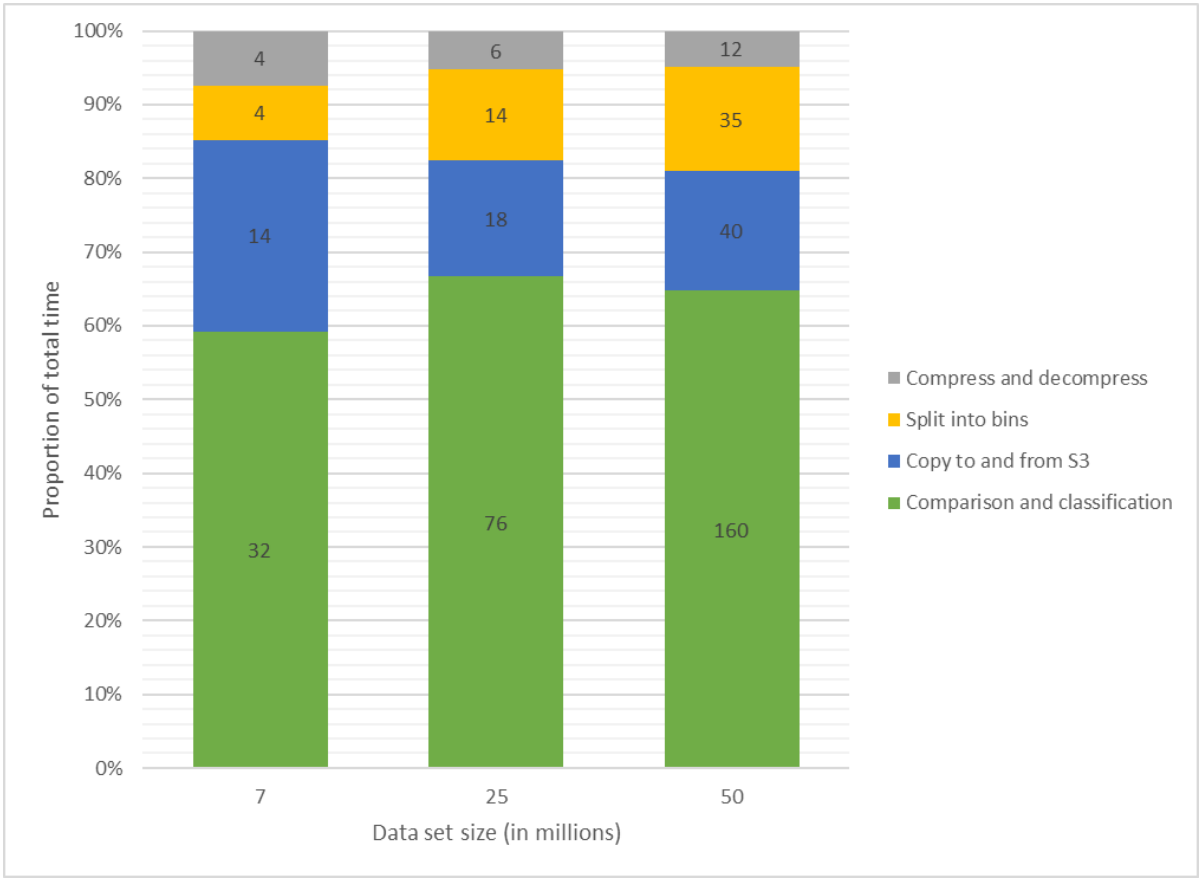
The total comparison space used for each data set, along with the bin count and pair count, is presented in Table 3. The two blocks used for the creation of separate bins for distribution across the processing cluster resulted in some duplication of comparisons and, thus, duplication of pairs.

Table 3. Comparison space and pairs created during classification.

Data set size (millions)	Comparison space, n	Bins, n	Total pairs, n	Unique pairs, n	Pairs files size (GB)
7	2,745,977,009	28	634,544,432	415,444,583	9
25	18,458,616,866	93	2,169,337,646	1,594,343,961	22
50	53,848,633,907	270	4,424,983,776	3,260,509,561	44

Approximately 60% of the time was spent on comparison and classification by each container (Figure 9). Much of the time was spent managing data in and out of the container itself. Splitting a data set into bins for parallel computation took between 7% (4/54 minutes) and 14% (35/247 minutes) of the total task time, a reasonable sacrifice considering the scalability factor this gives for the classification jobs. Provisioning of the compute resources took between 2 and 4 min for each data set.

Figure 9. Task execution time (in minutes) and the proportion of total time.



Running times of ad hoc queries on data sets are shown in Table 4; these were each executed 5 times on the client and run through the AWS Athena API. The mean execution time did not vary greatly across the differently sized data sets. With a simple count query taking around 25 seconds, there appears to be some initial setup time for provisioning the backend Presto cluster. This is expected and should not be considered an issue, particularly with all queries of the largest data set of 4.4 billion pairs taking less than 1 min to execute.

Table 4. Mean execution times for sample queries on full pairs set.

Data set size (millions)	Pairs count (millions)	Sample queries		
		Count (seconds)	Pair score histogram (seconds)	Fetch pairs in score range 15-16 (seconds)
7	635	26	52	51
25	2169	27	56	53
50	4424	24	52	54

In terms of costs associated with the use of AWS cloud services for our evaluation, there were 2 main types. First, the cost of on-demand processing, which is typically charged by the second. This totaled just over US $20 for the linkage processing used for all 3 data sets. The second is the cost of storage, which is charged per month. To retain the pairs files generated for all 3 data sets, it cost only US $2 per month. Querying data via the Athena service is currently charged at US $5 per terabyte scanned.

Discussion

Principal Findings

Our results show that an effective cloud model can be successfully developed, which extends linkage capacity into cloud infrastructure. A prototype was built based on this model. The execution times of the prototype were reasonable and far shorter than one would expect when running the same software on a single hosted machine. Indeed, it is likely that on a single hosted machine, the large data set (50 million) would need to be broken up into smaller chunks and linkages on these chunks run sequentially.

The splitting of data for comparison into separate bins worked well for distributing the work and mapped easily to the AWS Batch mechanism for execution of a cluster of containers. The creation of an AWS step function to manage the process from start to end was relatively straightforward. Step functions provide out-of-the-box support for AWS Batch. However, custom Lambda functions were required to trigger the AWS Glue crawler and retrieve the results from the first data-split task so that the appropriate size batch job could be provisioned.

As the fields used for splitting the data were the same as those used for blocking on each node, the comparison space was not different from running a linkage of the entire data set on a single machine. With the same comparison space and probabilistic parameters, the accuracy of the linkage is also identical. Having a mechanism for distributing linkage processing on multiple nodes with no reduction in accuracy is certainly a massive advantage for data linkage units looking to extend their linkage capacity.

The AWS Batch job definition's retry strategy was configured with five attempts, applying to each job in the batch. This provides some resilience to instance failures, outages, and failures triggered within the container. However, in our evaluation, this feature was never triggered. The timeout setting was set to a value well beyond what was expected as jobs that time out are not retried, and our prototype did not handle this particular scenario. Although our implementation of the step function provided no failure strategies for any task in the workflow, handling error conditions is supported and retry mechanisms within the state machine can be created as desired. An operational linkage system would require these failure scenarios to be handled.

Improvements to the prototype will address some of the other limitations found in the existing implementation. For example, S3 data transfer times could be reduced by using a series of smaller result files for pairs and uploading all of these in parallel. The over-matching and duplication of pairs could be addressed by improving the indexing algorithm used to split data. Although there is inevitably going to be some overlap of blocks, our naïve implementation could be improved. Our algorithm for distributing blocks attempts to distribute workload as evenly as possible based on the estimated comparison space. Discarding overly large blocks helps prevent excessive load on single matching nodes. However, it relies on secondary blocks to match the records within and only partly prevents imbalanced load distribution. The block-based load balancing techniques developed for the MapReduce linkage algorithms can be applied here to mitigate data skew further, where record pairs are distributed for matching instead of blocks.

As improvements to PPRL techniques are developed over time, these changes can be factored in with little change to the model. Future work on the prototype will look to extend the capability of PPRL to use additional security advances such as homomorphic encryption [42] and function-hiding encryption [43].

Conclusions

The model developed and evaluated here successfully extends linkage capability into the cloud. By using PPRL techniques and moving computation into cloud infrastructure, privacy is maintained while taking advantage of the considerable scalability offered by cloud solutions. The adoption of such a model will provide linkage units with the ability to process increasingly larger data sets without impacting data release protocols and individual patient privacy. In addition, the ability to store detailed linkage information provides exciting opportunities for increasing the quality of linkage and advancing the analysis of linkage outputs. Rich analytics, machine learning, automation, and visualization of these additional data will enable the next generation of quality assurance tooling for linkage.

Acknowledgments

AB has been supported by an Australian Government Research Training Program Scholarship.

Authors' Contributions

The core of this project was completed by AB as part of his PhD project. SR provided assistance and support for requirements analysis, testing, and data analysis. Both authors read, edited, and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Abebe R. Using search queries to understand health information needs in Africa. ArXiv 2019 (Forthcoming) [FREE Full text]
2. Radin JM, Wineinger NE, Topol EJ, Steinhubl SR. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. Lancet Digit Health 2020 Feb;2(2):e85-e93. [doi: 10.1016/s2589-7500(19)30222-5]
3. Lai S, Farnham A, Ruktanonchai NW, Tatem AJ. Measuring mobility, disease connectivity and individual risk: a review of using mobile phone data and mHealth for travel medicine. J Travel Med 2019 May 10;26(3) [FREE Full text] [doi: 10.1093/jtm/taz019] [Medline: 30869148]
4. Khoury MJ, Iademarco MF, Riley WT. Precision public health for the era of precision medicine. Am J Prev Med 2016 Mar;50(3):398-401 [FREE Full text] [doi: 10.1016/j.amepre.2015.08.031] [Medline: 26547538]
5. McGrail K, Jones K. Population data science: the science of data about people. Int J Population Data Sci 2018 Sep 6;3(4). [doi: 10.23889/ijpds.v3i4.918]
6. Christen P. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin/Heidelberg: Springer Science & Business Media; 2012.
7. Dunn HL. Record linkage. Am J Public Health Nations Health 1946 Dec;36(12):1412-1416. [Medline: 18016455]
8. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61-81 [FREE Full text] [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
9. Population Health Research Network: 2013 Independent Panel Review. The Population Health Research Network (PHRN). 2014. URL: https://www.phrn.org.au/media/80607/phrn-2013-independent-review-findings-and-recommendations-v2-_final-report-april-17-2014-2.pdf [accessed 2020-09-15]
10. Annual Report 2015. Centre for Big Data Research in Health (CBDRH). 2015. URL: https://cbdrh.med.unsw.edu.au/sites/default/files/CBDRH_Annual%20Report_2015_160609_Final.pdf [accessed 2020-09-15]
11. Predicts 2019: Cloud Adoption and Increasing Regulation Will Drive Investment in IT Vendor Management. Gartner. 2019. URL: https://www.gartner.com/en/documents/3896211/predicts-2019-cloud-adoption-and-increasing-regulation-w [accessed 2020-09-15]
12. Vasiljeva T, Shaikhulina S, Kreslins K. Cloud computing: business perspectives, benefits and challenges for small and medium enterprises (case of Latvia). Procedia Eng 2017;178:443-451. [doi: 10.1016/j.proeng.2017.01.087]
13. Gunadham T, Kuacharoen P. Security concerns in cloud computing for knowledge management systems. J Appl Stat 2019;1:52-60. [doi: 10.1109/itng.2013.127]
14. John J. Major vulnerabilities and their prevention methods in cloud computing. In: Advances in Big Data and Cloud Computing. New York, USA: Springer; 2019.
15. El-Ghafar R. Record Linkage Approaches in Big Data: a State of Art Study. In: 13th International Computer Engineering Conference. 2017 Presented at: ICENCO'17; December 27-28, 2017; Cairo, Egypt. [doi: 10.1109/icenco.2017.8289792]
16. Dou C. Probabilistic Parallelisation of Blocking Non-matched Records for Big Data. In: IEEE International Conference on Big Data. 2016 Presented at: Big Data'16; December 5-8, 2016; Washington, DC, USA. [doi: 10.1109/bigdata.2016.7841009]
17. Chu X, Ilyas IF, Koutris P. Distributed Data Deduplication. In: Proceedings of the VLDB Endowment. 2016 Presented at: VLDB'16; October 14, 2016; New Delhi, India. [doi: 10.14778/2983200.2983203]
18. Gazzarri L, Herschel M. Towards task-based parallelization for entity resolution. SICS Softw-Inensiv Cyber-Phys Syst 2019 Aug 26;35(1-2):31-38. [doi: 10.1007/s00450-019-00409-6]
19. Papadakis G, Skoutas D, Thanos E, Palpanas T. Blocking and filtering techniques for entity resolution. ACM Comput Surv 2020 Jul;53(2):1-42. [doi: 10.1145/3377455]
20. Karapiperis D, Verykios VS. A fast and efficient Hamming LSH-based scheme for accurate linkage. Knowl Inf Syst 2016 Feb 3;49(3):861-884. [doi: 10.1007/s10115-016-0919-y]
21. Pita R, Pinto C, Melo P, Silva M, Barreto M, Rasella D. A Spark-based Workflow for Probabilistic Record Linkage of Healthcare Data. CEUR-WS. 2015. URL: http://ceur-ws.org/Vol-1330/paper-04.pdf [accessed 2020-09-14]
22. Gagliardelli L, Simonini G, Beneventano D, Bergamaschi S. SparkER: scaling entity resolution in spark. Adv Data Technol 2019;2019:602-605. [doi: 10.5441/002/edbt.2019.66]

23. Mestre DG, Pires CE, Nascimento DC, de Queiroz AR, Santos VB, Araujo TB. An efficient spark-based adaptive windowing for entity matching. J Syst Software 2017 Jun;128:1-10. [doi: 10.1016/j.jss.2017.03.003]
24. Sehili Z, Kolb L, Borgs C, Schnell R, Rahm E. Privacy Preserving Record Linkage with PPJoin. Abteilung Datenbanken Leipzig - Universität Leipzig. 2015. URL: https://dbs.uni-leipzig.de/file/P4Join-BTW2015.pdf [accessed 2020-09-15]
25. Boratto M, Alonso P, Pinto C, Melo P, Barreto M, Denaxas S. Exploring hybrid parallel systems for probabilistic record linkage. J Supercomput 2018 Mar 21;75(3):1137-1149. [doi: 10.1007/s11227-018-2328-3]
26. Christen P. A survey of indexing techniques for scalable record linkage and deduplication. IEEE Trans Knowl Data Eng 2012 Sep;24(9):1537-1555. [doi: 10.1109/tkde.2011.127]
27. Trinckes Jr JJ. The Definitive Guide to Complying With the HIPAA/HITECH Privacy and Security Rules. Boca Raton, Florida, United States: CRC Press; 2012.
28. Regulation (EU) 2016/679 of the European Parliament. Publications Office of the EU - Europa EU. 2016. URL: https://op.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en [accessed 2020-09-15]
29. Vatsalan D, Sehili Z, Christen P, Rahm E. Privacy-preserving record linkage for big data : current approaches and research challenges. Handbook Big Data Technol 2017:851-895 [FREE Full text] [doi: 10.1007/978-3-319-49340-4_25]
30. Boyd JH, Randall S, Brown AP, Maller M, Botes D, Gillies M, et al. Population data centre profiles: centre for data linkage. Int J Population Data Sci 2020 Mar 11;4(2). [doi: 10.23889/ijpds.v4i2.1139]
31. Schnell R. PPRL: Privacy Preserving Record Linkage. Springer. 2019. URL: https://link.springer.com/content/pdf/10.1007%2F978-3-319-63962-8_17-1.pdf [accessed 2020-09-15]
32. Karapiperis D, Gkoulalas-Divanis A, Verykios VS. LSHDB: A Parallel and Distributed Engine for Record Linkage and Similarity Search. In: 16th International Conference on Data Mining Workshops. 2016 Presented at: ICDMW'16; December 12-15, 2016; Barcelona, Spain. [doi: 10.1109/icdmw.2016.7867099]
33. Bonomi L, Xiong L, Lu JJ. LinkIT: Privacy Preserving Record Linkage and Integration via Transformations. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. 2013 Presented at: SIGMOD'13; June 22-27, 2013; New York, USA URL: http://dl.acm.org/citation.cfm?doid=2463676.2465259 [doi: 10.1145/2463676.2465259]
34. Toth C, Durham E, Kantarcioglu M, Xue Y, Malin B. SOEMPI: a secure open enterprise master patient index software toolkit for private record linkage. AMIA Annu Symp Proc 2014;2014:1105-1114 [FREE Full text] [Medline: 25954421]
35. Randall S, Boyd J, Ferrante A, Brown A, Semmens J. Grouping Methods for Ongoing Record Linkage. In: Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference. 2015 Presented at: ACM-SIGKDD PopInfo'15; April 11, 2015; Sydney, Australia.
36. Irvine K, Smith M, de Vos R, Brown A, Ferrante A, Boyd J, et al. Real world performance of privacy preserving record linkage. Int J Population Data Sci 2018 Sep 10;3(4). [doi: 10.23889/ijpds.v3i4.990]
37. Lee YA. Medicineinsight: Scalable and Linkable General Practice Data Set. In: Health Data Analytics. 2019 Presented at: HDA'19; October 16, 2019; Sydney, Australia.
38. Christen P, Pudjijono A. Accurate synthetic generation of realistic personal information. Adv Know Discovery Data Mining 2009;5476:507-514 [FREE Full text] [doi: 10.1007/978-3-642-01307-2_47]
39. PSMA Geocoded National Address File (G-NAF). Australian Government. 2016. URL: https://data.gov.au/data/dataset/19432f89-dc3a-4ef3-b943-5326ef1dbecc [accessed 2020-09-15]
40. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating data Linkage software. J Biomed Inform 2012 Feb;45(1):165-172 [FREE Full text] [doi: 10.1016/j.jbi.2011.10.006] [Medline: 22061295]
41. Brown AP, Randall SM, Ferrante AM, Semmens JB, Boyd JH. Estimating parameters for probabilistic linkage of privacy-preserved datasets. BMC Med Res Methodol 2017 Jul 10;17(1):95 [FREE Full text] [doi: 10.1186/s12874-017-0370-0] [Medline: 28693507]
42. Randall S. Privacy Preserving Record Linkage Using Homomorphic Encryption. In: Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference. 2015 Presented at: ACM-SIGKDD PopInfo'15; April 11, 2015; Sydney, Australia.
43. Lee J, Kim D, Song Y, Shin J, Cheon J. Instant Privacy-Preserving Biometric Authentication for Hamming Distance. Semantic Scholar. 2018. URL: https://api.semanticscholar.org/CorpusID:57760611 [accessed 2020-09-15]

Abbreviations

API: application programming interface
AWS: Amazon Web Services
CPU: central processing unit
GPU: graphics processing unit
LSH: locality sensitive hashing
PPRL: privacy-preserving record linkage
TTP: trusted third party

Edited by G Eysenbach; submitted 27.03.20; peer-reviewed by S Albakri, A Al-Muqarm, T Enamorado; comments to author 23.06.20; revised version received 02.08.20; accepted 23.08.20; published 23.09.20

Please cite as:
Brown AP, Randall SM
Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model
JMIR Med Inform 2020;8(9):e18920
URL: http://medinform.jmir.org/2020/9/e18920/
doi: 10.2196/18920
PMID:

Chapter 5

Implementation and translation

Aim 5 of this thesis is to *validate privacy-preserving techniques in real-world projects*. This chapter addresses this aim by detailing some real-world privacy-preserving linkages through a series of case studies that incorporate one or more of the techniques developed through this thesis.

5.1 Use and availability of privacy-preserving record linkage

Much of the work in this thesis has been on record linkage methodology and algorithmic improvements to linkage quality, privacy and performance. This research has been incorporated into the LinXmart linkage software [39], including changes to data encoding, probability estimation, matching algorithms and grouping techniques. Embedding such features in software has enabled a number of PPRL evaluations and real-world linkage projects in Australia and internationally. In addition, a separate, standalone tool – the LinXmart Envelope Builder - for data custodians has been created to assist in the encoding of raw personal identifiers into a privacy-preserved state. This tool provides the capability for data custodians to easily transform identifiers into a consistent state for linkage. It has also been converted to a code library that can be incorporated into other products to convert datasets on the fly.

All of the software developments have been undertaken by me.

5.2 Research translation evaluation case studies

The following case studies describe how the elements of this thesis have been translated and evaluated for use by data linkage units. All of the case studies used the LinXmart software to perform the privacy-preserving record linkage.

5.2.1 Strategic project on PPRL – Population Health Research Network

The Population Health Research Network (PHRN) in Australia endorsed a joint project between several nodes in the PHRN to assess the viability of PPRL within an operational context. The CDL, CHeReL and WA Data Linkage Branch worked together to develop an agreed operational model (compatible with existing operational models) that could be used to complement existing record linkage services.

The project's main aim was to assess Bloom filters for probabilistic, privacy-preserving record linkage using. The project examined privacy, scalability, error tolerance and security using real-world administrative data. Two separate evaluations, using data from WA and NSW, were included in the evaluation process. Both evaluations received ethical approval.

WA evaluation

The evaluation with WA data involved a comparison between privacy-preserving record linkage and unencrypted record linkage using two datasets: 5,580,353 records from the Western Australian Hospital Morbidity Data Collection (2011-2015) and 68,955 death registration (2011-2015) records. Identical blocking strategies were used for both encrypted and unencrypted

linkages. The WA Data Linkage Branch (WA-DLB) used their operational linkage system to link the unencrypted datasets. Encoded versions of the same datasets were sent to the CDL and the LinXmart software was used to link these using PPRL techniques. The linkage map generated by the CDL on privacy-preserved data was sent to the WA-DLB for comparison against the unencrypted linkage. Further in-depth analysis of the results enabled the WA-DLB to determine the reasons for differences between the linkages.

Of the 68,955 entity groups containing a mortality record, 68,028 of them (98.7%) were identical between the encrypted and unencrypted linkages. Approximately half of the remaining entity groups were found through PPRL linkage but not through clear-text linkage. The remainder was found through clear-text linkage and not through PPRL linkage. In all instances, the majority of additional links were determined to be correct matches.

As a result of this evaluation it was determined that privacy-preserving linkage methods appear ready for real-world use and should be considered in situations where clear-text data are not available.

NSW evaluation

The Centre for Health Record Linkage (CHeReL) in the NSW Ministry of Health evaluated PPRL techniques using Bloom filters for a linkage of primary and secondary health care data. Primary health care data included 272,202 records from 16 general practices in NSW. This was linked to 42.8 million records from seven years of emergency presentations, hospital morbidity events and death registrations [136].

Again, two linkages were performed to assess the performance of PPRL. The first was an unencrypted linkage of primary care data with the secondary care data using CHeReL's existing linkage system. The second used privacy-preserved versions of the same datasets and LinXmart software to probabilistically link them. The quality of PPRL linkage was estimated by comparing results against the unencrypted, 'gold standard' linkage using full personal identifiers.

The PPRL linkage produced quality metrics of precision, recall and F-measure all in excess of 0.90. When configured with the best-link grouping method to leverage the pre-existing links between secondary care data (emergency department, hospital and mortality data), quality metrics of 0.98-50.99 were achieved. Some GP datasets produced lower rates of linkage quality, which was attributed to missing demographic information. Some residual variation in linkage quality across practices was also observed.

5.2.2 PPRL evaluation – Population Data BC, Canada

Population Data BC, a trusted third-party linkage unit in Vancouver, Canada, holds and links data from multiple organisations [4]. It provides full life-cycle management of requests for data access, operates secure environments for analysis, and offers education and training on using linked data to researchers. As with most jurisdictions, some data collections have been

excluded from record linkage by Population Data BC due to legislative and administrative restrictions. This project aimed to determine whether privacy-preserving record linkage methods could be appropriately adapted and used in real-world settings in Canada.

Two different datasets were selected for linkage against the population spine that Population Data BC currently manages (8M records). The first, Vital Statistics, was selected as a small dataset with high-quality data (36K records). The second, WorkSafe BC, was selected as a much larger dataset with lower quality (2.5M records). For the purposes of the evaluation, linkages were performed using unencrypted identifiers with Population Data BC's standard linkage process, as well as using privacy-preserved data with the LinXmart software. The unencrypted linkage was used as a truth-set to evaluate the performance of the privacy-preserving linkage.

Linkage of the Vital Statistics records against the population spine resulted in an F-measure value of 0.99. This was an excellent result and attributed to the high-quality of the source data. Initial linkage of the WorkSafe BC dataset with the population directory produced precision, recall and F-measure values of 0.99, 0.80 and 0.88 respectively. The very high precision combined with low recall suggests a potential issue with the threshold value; it is likely too high, and pairs are being missed. This evaluation remains ongoing.

5.3 Real-world project case studies

The next case studies detail how the work in this thesis has been utilised in real-world projects, highlighting the importance of this work for many organisations moving forward. These can be categorised into three main types of use case. The first is the use of PPRL to overcome issues with data sovereignty, where policies restrict the release of clear text data outside of a jurisdiction. This restriction typically requires a trusted fourth party to link data (using PPRL) between jurisdictions. The second type involves horizontal and vertical integration of general practice data, where each data custodian is potentially a small, independent organisation. These organisations are risk averse and protective of their client's data, and the logistics involved in bringing data together from all custodians is complex. The third type is multi-sectoral linkage, where laws restrict some of the data custodians from releasing named-identified data. This type uses a trusted third party model with additional internal separation of traditional and privacy-preserving linkage.

All linkage projects used the LinXmart software to perform the privacy-preserving record linkage.

5.3.1 Continuity of primary care on secondary care – WA

This project aimed to use whole-population Commonwealth and State linked person-level data to evaluate the influence of patterns of primary care contact on emergency department (ED) visits and potentially preventable hospitalisations (PPHs) [199].

The primary objective was to examine the continuity of primary contact against secondary care data. This included the effect (as well as the impact on costs) on ED visits and PPHs for any condition defined as 'chronic/complex' and 'acute medical' in the National Health Performance Framework Performance Indicators in WA from 2005-2014. Determining the socio-demographic and clinical predictors of primary care contact patterns for patients with these conditions was also an objective.

The study was conducted using a whole-population longitudinal study design. The Medicare Benefits Schedule (MBS) dataset was used to identify primary care (GP) attendances in each year from 1990-1998 (pre-EPC Medicare era); 1994-2004 (Medicare EPC era) and 2005-2014 (consolidation of Medicare EPC era) for individuals (aged 18+ years) in WA.

From a linkage perspective, the study required the matching of Medicare datasets held by the Commonwealth to emergency and hospitalisation datasets held by the State of WA. Several factors, including privacy legislation and custodial restrictions, prevented the release of personal identifiers for linkage. The CDL worked with the project research team, and the State and Commonwealth data custodians to provide a solution using privacy-preserving linkage techniques.

This was the first project in Australia to use privacy-preserving methods to link State and Commonwealth data for health research. The linkage brought together Medicare Enrolment File (MEF) records and WA Hospital (WA-HMD) and Emergency Department (WA-ED) data over a fifteen-year period. The linkage was conducted by the CDL (as a trusted third-party) using the LinXmart software. Each of the three datasets (MEF, WA-HMD and WA-ED) was encrypted prior to being released by the custodians. Data flows are shown in Figure 5.1.

Acting as the trusted third-party, CDL first defined the linkage parameters for the project. This *linkage definition* included the fields that were used for linkage, all privacy-preserving parameters required for each field, and any standardisation properties required before the fields were transformed. Privacy-preserving parameters covered the transformation type (i.e. Bloom filters for fields requiring approximate comparisons and cryptographic hashes for those fields that did not) and the specific properties required for each type. Standardisation was applied to ensure that fields were privacy-preserved in a consistent way across all datasets.

The MEF dataset (11,654,126 records) contained names and addresses of individuals with an address in WA. Due to this dataset's high quality and coverage, it was not deduplicated prior to linkage with the other datasets. Rather, it was used as a *population spine* for the linkage. The WA-HMD and WA-EM datasets were each deduplicated and then linked to the MEF dataset using the weighted best-link group method. A total of 2,905,534 individuals were found at the completion of the linkage, with 73% individuals having a MEF record in addition to a WA-HMD or WA-EM record.

The research work based on these linkage results is still ongoing.

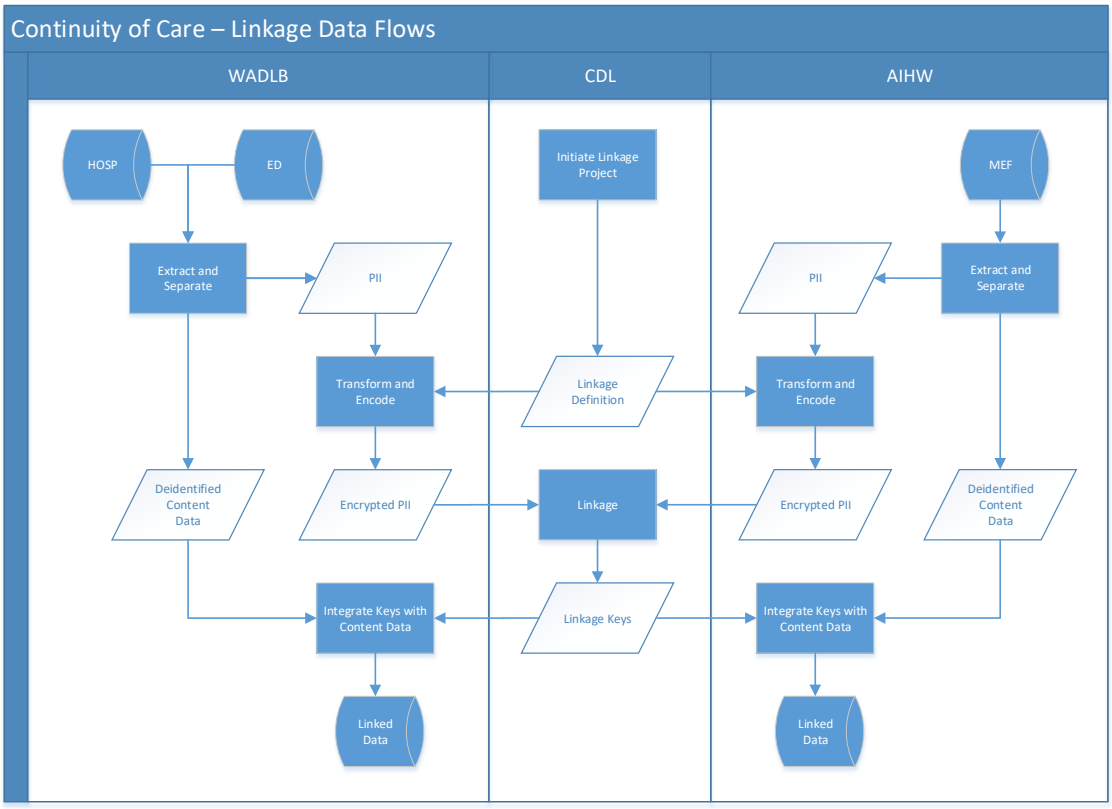


FIGURE 5.1: Continuity of care linkage data flows

5.3.2 Computerised tomography scanning – WA

The Computerised Tomography (CT) Scanning project aims to evaluate the radiation burden and estimated incident cancers attributed to CT examinations in WA from 2003 to 2016. This project builds on a previous study where one of the limitations was that only aggregate data was available. A linkage of datasets at the individual level aims to provide investigations of repeat scanning, changes in scanning rates and the associated radiation dose/cancer risk according to diagnostic cohorts. The use of PPRL techniques allowed the required data to be linked for the first time.

The Medicare Benefits Schedule (MBS) dataset was used to identify CT scans undertaken in WA from 2003-2016. Picture archival communication system (PACS) service data, WA Cancer Registry data, hospital morbidity data (WA-HMD) and emergency department data (WA-EM) were required to meet the project's aims.

From a linkage perspective, the methodology used for this project was similar to the Continuity of care project. The linkage team built upon the lessons learnt from the Continuity of care project, incorporating the additional datasets. The Medicare Enrolment File (MEF) data was used for linkage for the MBS data. This formed a population spine, with the remaining datasets

deduplicated and then linked to the spine using the weighted best-link group method. A total of 2,096,458 individuals were found across all datasets.

There are a number of research articles submitted for publication using this linked data, and additional research work is ongoing.

5.3.3 Lumos - Ministry of Health, NSW

The success of the PHRN Strategic project on PPRL in NSW (described earlier) confirmed the viability of PPRL methods for linkage of primary and secondary care data and led to a much larger project called Lumos [207], tracking the patient journey through the NSW primary and secondary health care systems. Lumos greatly expanded the number of datasets to be linked by PPRL, with as many as 500 GP practices across all 10 PHNs expected to participate in the project between 2019 and 2022. Lumos has been designed such that privacy-preserving techniques to encode patient information occurs at source (i.e. at the practice). This encoded data is securely transferred to CHeReL where it is linked to the NSW master linkage key (MLK), a set of linked secondary care datasets [134]. With the volume of data being transmitted and processed as part of Lumos, the ability for the solution to scale to meet demand is essential; in particular, for data analysis, quality verification, parameter estimation and data flow. The use of a trusted third-party hybrid cloud model, as described in Chapter 4, is a candidate solution to scaling the system as the size and number of datasets increase over time.

5.3.4 Linked primary care data warehouse – NPS MedicineWise

The MedicineInsight program run by NPS MedicineWise aims to provide a set of de-identified general practice records that can be used by researchers to derive valuable insights into primary care data to enable better patient outcomes. Data are sourced directly from the general practice and sent to NPS MedicineWise through a custom middleware solution.

A data warehouse was established in an IRAP compliant Dell Data Centre, containing de-identified data from 730 general practices using Medical Director and Best Practice clinical information systems [181]. The data warehouse provides a standardised view of the general practice data to a select group of MedicineWise users for reporting and insights derivation. This includes data sharing for research, machine learning and visualisation.

An important element in the design of the warehouse was the inclusion of individual but anonymous person ids. Adding person keys to the data warehouse through linkage required extending the middleware to extract encrypted personal identifiers from the source general practice databases. This was done by including the LinXmart Envelope Builder libraries to transform the personal identifiers into a privacy-preserved state as they are read from the database. Consistency in the transformations across all general practices was critical and achieved through the creation and distribution of a standard linkage definition file. This file was distributed to all participating general practices. The transformed data was then sent to NPS, where it was compiled and forwarded to the CDL for linkage. Following linkage, the

linkage keys were sent back to NPS where they were added to the data warehouse. These data flows are described in Figure 5.2.

Some initial testing of these systems and data flows were undertaken. An initial test included three general practices (58K, 82K and 139K records respectively) to verify and validate encoding consistency, format and data flow. These general practices were selected due to their close geographic proximity. The linkage revealed a number of duplicate records within each practice, with 5.1%, 16.1% and 3.3% of records found as duplicates across the three GP sites. The linkage also found 8% of individuals had records in two or three of these sites. While the linkage could not be compared to a truth set, samples of duplicate records and patients across two or more practices were selected for NPS to verify at their source; minimal discrepancy was found.

The next step in the project was to extend and verify the linkage of approximately 150 general practices. The size of these sites varied from 3K records to 265K records, with a total number of 5.8M records across all sites. The linkage identified 4.9M individuals with 14% having two or more records within a single general practice. Approximately 7% of individuals had records at more than one practice.

NPS MedicineWise stores and analyses its linked data in AWS on the Snowflake Data Warehouse platform. This hosting platform provides the capability to scale relatively easily as required. A future step for the linkage component would include moving to a trusted third-party full cloud model (described in Chapter 4). The linkage would remain independently managed, but hosting the linkage services within the same datacentre as the NPS data warehouse minimises data transfer speeds and costs while gaining all the benefits of the clouds' elastic scaling.

5.3.5 Social Investment Data Repository (SIDR) – WA

The Social Investment Data Resource (SIDR) is a large, de-identified researchable database, created by the Western Australian government and the CDL at Curtin University. Data across government have been brought together into a linked repository of de-identified information on individuals who have had contact with key government agencies [286]. This resource was initially created to support the Government's Target 120 program, assessing the long-term costs and benefits of interventions to young offenders and their families.

A distributed linkage model between the Department of Health Data Linkage Branch (DLB) and the CDL was established to create links between health and non-health data [98]. This model included the use of a common spine of birth records and education attendance records to integrate data between the organisations. The health datasets were linked (unencrypted) by the DLB as part of their ongoing curation of health data in WA. Some of the non-health datasets were also linked using unencrypted identifiers by the CDL. However, several datasets were required to be encoded into a privacy-preserving state before being released from the data providers (see Figure 5.3). The unencrypted datasets were first linked using traditional methods, the linked results then encoded into a privacy-preserving state and linked with the remaining privacy-preserved datasets.

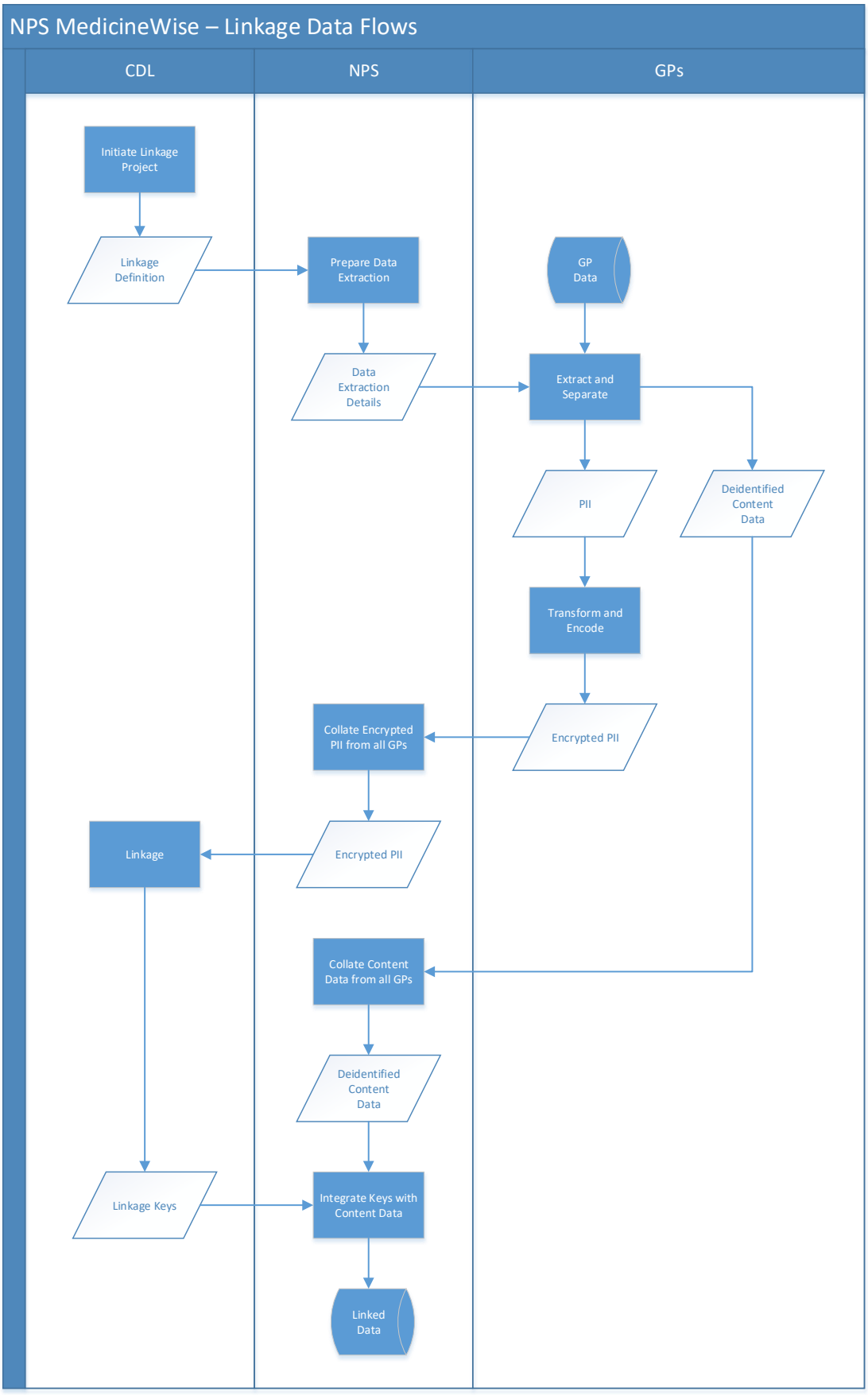


FIGURE 5.2: NPS MedicineWise linkage data flows

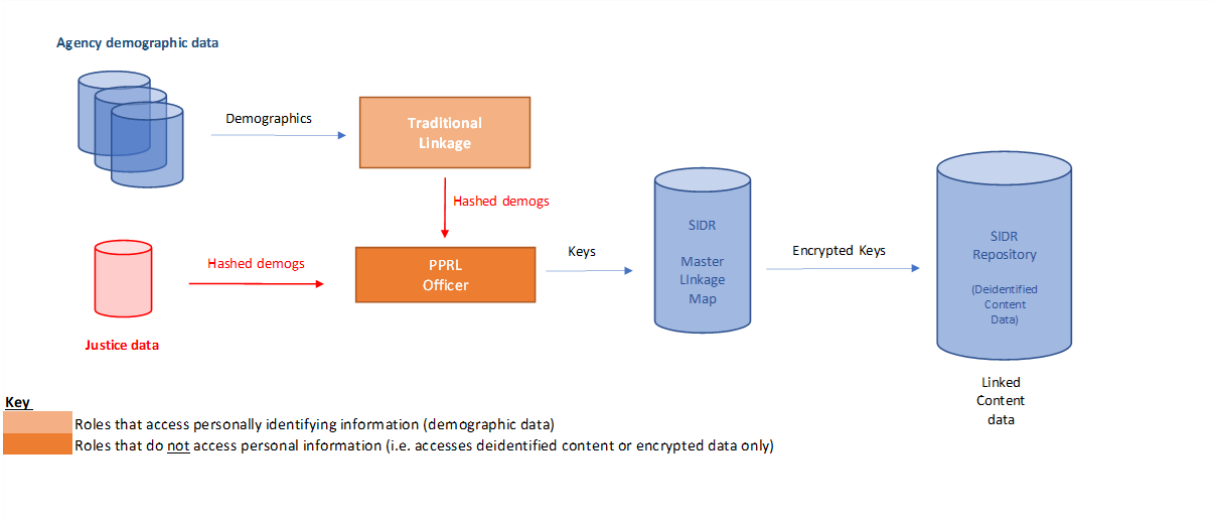


FIGURE 5.3: Combination of PPRL and traditional linkage used for SIDR

The use of this distributed linkage model, combining both traditional and PPRL methods, provided an innovative yet pragmatic way of delivering data linkage services to a large, cross-sectoral research project. These PPRL methods enabled the inclusion of otherwise excluded datasets in the project. SIDR includes health data, education records, justice, child protection, disability and housing data. This repository provides a resource for whole-of-government policy development, service evaluation, academic research and social investment analytics.

The SIDR project will refresh data sources annually, requiring a refresh of linkages, and there are plans to extend the cohort further. This provides an opportunity to enhance SIDR linkage infrastructure. While this linkage infrastructure is currently hosted on-premises, a move to a trusted third-party hybrid cloud model (described in Chapter 4) would provide the advantages of elastic scale while keeping linkage quality high and named identifiers local. These identifiable datasets could then be encoded before being passed to the cloud-hosted infrastructure. Datasets that were provided in a privacy-preserved format can simply be uploaded to the cloud as is.

Chapter 6

Conclusion

Present operational models of record linkage commonly use trusted third parties to link data using personal identifiers. Increasing growth in data creation and health research innovations from non-traditional data sources are challenging current processes and providing data linkage units with an opportunity to embrace emerging PPRL technologies. Providing access to previously unattainable datasets opens up new and expanded opportunities for research on linked-data, including sources from the private health sector, which has had, to date, limited exposure to data linkage frameworks. Ensuring the accuracy of linked data, and providing the infrastructure to support the increasing volume of data, is essential.

Cloud computing infrastructure provides a solution to address the increasing size and complexity of data requiring linkage. Utilising this infrastructure can be achieved using PPRL techniques to protect the privacy of individuals, keeping personal identifiers local and using privacy-preserving techniques to perform linkage. For this to be an operationally viable option, however, the accuracy of the linkage must be assured. Optimisation of linkage parameters early in the linkage process is essential to maximising quality. The results of the work outlined in this thesis have shown that high linkage quality can be achieved while maintaining good privacy and using scalable linkage algorithms on cloud infrastructure.

This thesis presents quality optimisations of privacy-preserving record linkage techniques used within a probabilistic framework. The optimisation of parameters for probabilistic PPRL, as discussed in Chapter 2, offers some approaches to maximising the quality of linkage of privacy-preserved datasets using refined techniques built on existing, fundamental linkage methods. The evaluation of the expectation-maximisation (EM) algorithm on datasets encoded as Bloom filters (Paper 2) showed that m and u probabilities can be accurately estimated for large datasets, producing linkage quality comparable to the use of the *actual* probabilities. A small, but important, enhancement to the probability estimation was also presented, providing an estimate of a single threshold cut-off value. Together, these linkage parameters produced highly accurate results on several PPRL datasets irrespective of level of error contained in the data.

An evaluation of the effectiveness of several partial agreement methods for Bloom filters (Paper 3) revealed that the use of partial agreements with Bloom filters in probabilistic record linkage results in better linkage quality than without the use of partial agreements. An existing approach to modelling partial agreement for probabilistic linkages was applied to Bloom filters and the approximate comparisons of Jaccard, Sorensen-Dice and Hamming distance. These weight curves were generally consistent across datasets with different error rates but varied slightly per field. Matching results using Bloom filters and partial weight curves were comparable to the Jaro-Winkler comparison with weight curves on unencrypted data. Importantly, while the weight curves generated from a truth-set of data produced the best results, weight curves derived from synthetic datasets also produced excellent quality linkage results. This means that, in the absence of a truth set, synthetic datasets can be used to create weight curves that produce highly accurate linkage.

Combining the EM algorithm (for probabilistic parameter estimation) and partial weight

curves (for approximate matching) will help maximise linkage accuracy for PPRL. The use of advanced grouping methods such as Weighted Best Link (Paper 8) for the linkage of new datasets to a population-representative dataset (often referred to as a *population spine*) can further improve the accuracy of linkage. However, this type of dataset may not always be available, and the linkage accuracy will be largely dependent on the quality of the spine data itself. Nevertheless, the use of these quality maximisation techniques during the linkage process will reduce the need for manual review that typically follows.

The third aim of this thesis was to identify methods for improving privacy and performance of PPRL. This thesis presents an innovative PPRL technique using CLKs (Cryptographic Long-term Keys) and multibit trees (Paper 4) to address this aim. CLKs combine several fields within each record into a single Bloom filter, and multibit trees are used to index the CLK values for linkage efficiently. Multibit trees and these composite Bloom filters produced the highest quality linkage results when the address field and fields with missing values were excluded. The combination of fields in the CLK producing the highest quality results was still below the quality produced using field-level Bloom filters and probabilistic linkage methods. While improving on privacy protection through the use of composite Bloom filters, the performance of the multibit tree for a full linkage was sub-optimal. However, the use of multibit trees shows potential as an indexing step before matching occurs; the selection of fields for indexing must be carefully selected to maximise pairs completeness and further work in this area is required to realise its full potential.

While maximum linkage quality is essential for operational use, stronger privacy mechanisms are highly desirable if the cost to quality and performance is not too high. An application of homomorphic encryption on Bloom filters is presented in this thesis (Paper 9) as an additional level of privacy. Providing homomorphic encryption on top of Bloom filter encoding greatly increased privacy protection, with no drop in linkage accuracy. However, performance cost was high, with encryption, decryption and calculation on encrypted values requiring vastly more computation. This linkage method shows a lot of potential for the quality and privacy aspects of PPRL but in its current form is considered too slow for operational use. Improvements to the compute performance may be possible; however, further research and testing is required to achieve this.

An alternative and novel PPRL method is presented in this thesis (Paper 6) providing additional privacy protection; however, this is at the expense of partial agreements. The *match keys* approach is based on the principles of probabilistic linkage, creating a series of hashes per record based on the combinations of agreement and disagreement that may be possible with each field. An evaluation of this method showed that linkage quality was superior to linkage using an SLK-581 field, and only slightly inferior to that using traditional probabilistic linkage using field-level Bloom filters. The datasets used in the evaluation were of reasonably good quality, so the impact of removing partial agreements may have been minimised. While this is an important contribution to the field, the use of the *match keys* approach appears to be best suited to situations where the quality of data can be guaranteed.

This thesis has shown that PPRL techniques provide sufficient quality for operational use and sufficient privacy protection for use on cloud infrastructure. The fourth aim of this thesis was to develop a model for record linkage that maintains privacy protection while utilising the scalability of cloud infrastructure to efficiently link data. Chapter 4 presented several models for record linkage that meet this aim and the scenarios in which they would be used. A hybrid cloud model was presented in greatest detail (Paper 5) that keeps identifiable data on-premises and sends encoded datasets to cloud infrastructure for linkage. The quality of the linkage can be maximised using the techniques described in Chapter 2. Standard Bloom filters can be used for data privacy, or additional privacy measures such as those described in Chapter 3 could be incorporated. There are many ways to leverage the elastic scalability of commercial cloud providers, but the hybrid model presented here used containers in a managed cluster to distribute computation. An evaluation of this model showed that it is operationally feasible to use existing linkage software and scale this out, using bootstrapped containers and an effective compute cluster management tool. Commercial cloud services provide scalable compute at a reasonable cost, and additional analytical services give greater opportunity for advanced analysis, richer analytics, machine learning and automation of data processes in the cloud.

Chapter 5 focusses on the implementation and utilisation of the PPRL methods presented through this thesis. Interest in PPRL from data linkage units and research groups with specific linkage requirements has increased dramatically over the last few years. There has been a particularly strong interest within Australia to use PPRL to link primary care data with secondary care data. Primary care data has historically been difficult to obtain for linkage [207]. However, the use of PPRL in large projects like Lumos in NSW has provided access to general practice data at the state level. The quality techniques developed and presented in this thesis have been instrumental in getting these projects off the ground; first used to evaluate the quality that PPRL can provide, then to ensure each general practice dataset is accurately linked to a secondary care data population spine. While Lumos uses infrastructure within the NSW Ministry of Health, the MedicineInsight program by NPS MedicineWise runs its middleware and data warehouse on commercial cloud infrastructure. NPS does not hold personally identifiable information on this infrastructure. However, the use of a full cloud trusted third-party model for linkage (described in Chapter 4), hosted in the same datacentre, would be highly beneficial, minimising data transfer speeds and costs while leveraging the benefits of elastic scaling.

The increased demand for linkage, often with a desire to link to datasets that are historically difficult to obtain, continues to put pressure on the existing data linkage capability available. The work presented in this thesis offers the potential to alleviate that pressure by providing operationally viable solutions for handling the increase in data size and linking to datasets where the release of personally identifying information is not possible. Commercial cloud services provide the infrastructure for handling the elastic scaling desirable for managing large data linkage. The PPRL techniques presented in this thesis provide a mechanism by which these cloud services can be utilised while meeting requirements for maximum linkage accuracy and high privacy protection.

While the PPRL techniques presented in this thesis demonstrated positive outcomes for improving linkage accuracy, there may be situations where these techniques fail to deliver. Are there types of data that will consistently provide poor accurate linkage? What kinds of datasets are these and how can we identify them? If cleartext identifiers allow linkers to make better decisions for linkage, there may be other ways to reduce risk and improve privacy. Anonymising the dataset origin may be one technique. However, the identifiers of individuals are still readable, and information may be inferred. The trade-off between the accuracy provided through PPRL and the privacy provided with cleartext linkage techniques is an interesting topic for further study.

The fundamentals for a cloud model for record linkage has been provided in this thesis, providing the framework for deployment and use. While the evaluations that were run throughout this work have shown an enormous benefit to a PPRL cloud model, further evaluations with additional real-world datasets would improve this further. Obtaining access to real-world datasets is challenging, but it is essential in determining the effectiveness of these techniques in operation.

The homomorphic encryption with Bloom filters presented in Chapter 3 showed excellent privacy benefits without compromising on the accuracy of linkage. Tackling the computation challenges of homomorphic encryption within a scalable cloud environment is the next step for this technique, with the potential to significantly improve the privacy aspect of our cloud model without compromising on linkage accuracy. Combining the use of multibit trees with homomorphic Bloom filters for private indexing also show promise and further research is warranted.

The cloud models presented in Chapter 4 showed that data linkage units now have what they need to get started with scalable data linkage in a cloud environment. With some additional work upfront to split data into manageable sized 'chunks', existing linkage software can be used to match datasets across a large number of nodes within an on-demand compute cluster. Further refining of these record linkage algorithms for such a distributed paradigm would ensure the cloud infrastructure is used to its full potential. The storage of linkage maps (the results of the linkage) on cloud infrastructure also presents some exciting new opportunities for the use of 'big data' analytics on linkage maps, a relatively untapped field that could improve quality assurance and reporting capabilities.

This thesis has aimed to provide a foundation for the use of PPRL in a cloud computing environment. This cloud model for record linkage has been built on the capabilities that are available today and presents opportunities to improve with further advances in this area. Cloud computing services continue to evolve, and as enterprises continue to adopt these services, record linkage capabilities must grow with them. While many technical challenges remain, the work in this thesis provides a significant step towards the ultimate goal of providing researchers with the linked data they require, with high accuracy using processes that are both efficient and protective of privacy.

Appendix A

Conference Abstracts

The abstracts of the conference presentations delivered for work in this thesis have been published in journals. The published abstracts are included in this section. A presentation for each abstract was delivered in person at a conference.

International conference presentation(s):

11. Boyd JH, Ferrante AM, Randall SM, **Brown AP**, Semmens JB (2016). *Implementing privacy-preserving record linkage: welcome to the real world.* Conference: International Population Data Linkage Conference, Swansea, Wales, August 2016.

12. **Brown AP**, Borgs C, Randall SM, Schnell R (2016). *High quality linkage using Multibit Trees for privacy-preserving blocking.* Conference: International Population Data Linkage Conference, Swansea, Wales, August 2016.

13. **Brown AP**, Randall SM, Ferrante AM, Boyd JH (2018). *Public Cloud: The Future of Record Linkage?* Conference: International Population Data Linkage Conference, Banff, Alberta, Canada, September 2018.

International Journal of Population Data Science (2017) 1:134

International Journal of Population Data Science

Journal Website: www.ijpds.org

Implementing privacy-preserving record linkage: welcome to the real world

Boyd, James[1*], Ferrante, Anna[1], Brown, Adrian[1], Randall, Sean[1], and Semmens, James[1]

[1]Curtin University

Objective

While record linkage has become a strategic research priority within Australia and internationally, legal and administrative issues prevent data linkage in some situations due to privacy concerns. Even current best practices in record linkage carry some privacy risk as they require the release of personally identifying information to trusted third parties. Application of record linkage systems that do not require the release of personal information can overcome legal and privacy issues surrounding data integration. Current conceptual and experimental privacy-preserving record linkage (PPRL) models show promise in addressing data integration challenges but do not yet address all of the requirements for real-world operations. This paper aims to identify and address some of the challenges of operationalising PPRL frameworks.

Approach

Traditional linkage processes involve comparing personally identifying information (name, address, date of birth) on pairs of records to determine whether the records belong to the same person. Designing appropriate linkage strategies is an important part of the process. These are typically based on the analysis of data attributes (metadata) such as data completeness, consistency, constancy and field discriminating power. Under a PPRL model, however, these factors cannot be discerned from the encrypted data, so an alternative approach is required. This paper explores methods for data profiling, blocking, weight/threshold estimation and error detection within a PPRL framework.

Results

Probabilistic record linkage typically involves the estimation of weights and thresholds to optimise the linkage and ensure highly accurate results. The paper outlines the metadata requirements and automated methods necessary to collect data without compromising privacy. We present work undertaken to develop parameter estimation methods which can help optimise a linkage strategy without the release of personally identifiable information. These are required in all parts of the privacy preserving record linkage process (pre-processing, standardising activities, linkage, grouping and extracting).

Conclusion

PPRL techniques that operate on encrypted data have the potential for large-scale record linkage, performing both accurately and efficiently under experimental conditions. Our research has advanced the current state of PPRL with a framework for secure record linkage that can be implemented to improve and expand linkage service delivery while protecting an individualâĂŹs privacy. However, more research is required to supplement this technique with additional elements to ensure the end-to-end method is practical and can be incorporated into real-world models.

*Corresponding Author:
Email Address: j.boyd@curtin.edu.au (J. Boyd)

http://dx.doi.org/10.23889/ijpds.v1i1.153

International Journal of Population Data Science (2017) 1:130

International Journal of Population Data Science

Journal Website: www.ijpds.org

High quality linkage using Multibit Trees for privacy-preserving blocking

Brown, Adrian[1], Borgs, Christian[2], Randall, Sean[1], and Schnell, Rainer[2]

[1]Curtin University

[2]University of Duisburg-Essen

[3]Massachusetts Department of Public Health

Objectives

As privacy-preserving record linkage (PPRL) emerges as a method for linking sensitive data, efficient blocking techniques that help maintain high levels of linkage quality are required. This research looks at the use of a Q-gram Fingerprinting blocking technique, with Multibit Trees, and applies this method to real-world datasets.

Approach

Data comprised ten years of hospital and mortality records from several Australian states, totalling over 25 million records. Each record contained a linkage key, as defined by the jurisdiction, which was used to assess quality (i.e. used as a 'gold standard'). Different parameter sets were defined for the linkage tests with a privacy-preserved file created for each parameter set. The files contained jurisdictional linkage key and a Cryptographic Long-term Key (the CLK is a Bloom filter comprising all fields in the parameter set).

Each file was run through an implementation of the Q-gram Fingerprinting blocking algorithm as a deduplication technique, using different similarity thresholds. The quality metrics of precision, recall and f-measure were calculated.

Results

Resultant quality varied for each parameter set. Adding suburb and postcode reduced the linkage quality. The best parameter set returned an F-measure of 0.951. In general, precision was high in all settings, but recall fell as more fields were added to the CLK. We will report details for all parameter settings and their corresponding results.

Conclusion

The Q-gram Fingerprinting blocking technique shows promise for maintaining high quality linkage in reasonable time. Determining which fields to include in the CLK for the linkage of specific datasets is important to maximise linkage quality, as well as selecting optimal similarity thresholds. Developing new technology is important for progressing the implementation of PPRL in real-world settings.

*Corresponding Author:
Email Address: adrian.brown@curtin.edu.au (A. Brown)

http://dx.doi.org/10.23889/ijpds.v1i1.149

International Journal of Population Data Science (2018) 3:3:294

International Journal of Population Data Science

Journal Website: www.ijpds.org

Public Cloud: The Future of Record Linkage?

Brown, A[1], Randall, S[1], Ferrante, A[1], and Boyd, J[1]

[1]Curtin University

Introduction

Businesses worldwide are increasingly adopting the storage, compute and analytical services provided by cloud computing. Yet, few operational linkage units are keeping pace with this world of technological change - most use legacy systems approaching their limits with the rapidly increasing size and range of datasets now required for linkage.

Objectives and Approach

To meet the demands of linkage for the near future, it is important that new solutions for linkage consider the services provided by public cloud infrastructure for compute, storage and analytics. We examined Platform as a Service (PaaS) offerings for use in the development of a cost-effective cloud model for scalable, privacy-preserving record linkage (PPRL). PPRL techniques were adapted to maximise the quality of linkage and to automate as much of the process as possible. Finally, a prototype was created to demonstrate the capabilities and potential of the model.

Results

We present our cloud model for PPRL, a platform for record linkage that provides rapid scaling of resources to meet demand, and the results of how our prototype performed on massive datasets.

Conclusion/Implications

The application of record linkage using relatively inexpensive cloud infrastructure represents a significant step towards providing an efficient and scalable record linkage service to researchers and government. Larger datasets can be linked efficiently, including national or cross-jurisdictional datasets, with little investment in private infrastructure, and improved turnaround times for researchers.

http://dx.doi.org/10.23889/ijpds.v3i4.885

Appendix B

Statements of contribution

Brown AP, Ferrante, AM, Randall, SM, Boyd, JH, Semmens, JB (2017). *Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage.* Frontiers in Public Health, 5(March), 34. https://doi.org/10.3389/fpubh.2017.00034

Contribution:

AB contributed to the conception and design of the paper, co-wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

I acknowledge the above statement of contribution is accurate:

Sean Randall: _____

Anna Ferrante: _____

James Boyd: _____

James Semmens: _____

Brown AP, Randall SM, Ferrante AM, Semmens JB, Boyd JH (2017). Estimating parameters for probabilistic linkage of privacy-preserved datasets BMC Medical Research Methodology, 17(1), 95. https://doi.org/10.1186/s12874-017-0370-0

Contribution:

AB developed the research design and evaluation methodology for the paper, reviewed the literature, performed all evaluations, analysed the data, produced and interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

I acknowledge the above statement of contribution is accurate:

Sean Randall: _____

Anna Ferrante: _____

James Boyd: _____

James Semmens: _____

Brown AP, Randall, SM, Boyd, JH, Ferrante, AM (2019). E*valuation of approximate comparison methods on Bloom filters for probabilistic linkage.* International Journal of Population Data Science, 4(1). https://doi.org/10.23889/ijpds.v4i1.1095

Contribution:

AB developed the research design and evaluation methodology for the paper, reviewed the literature, performed all evaluations, analysed the data, produced and interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

I acknowledge the above statement of contribution is accurate:

Sean Randall: _____

Anna Ferrante: _____

James Boyd: _____

Brown, AP, Randall, SM (2020). *Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model.* JMIR Medical Informatics, 8(9), e18920. https://doi.org/10.2196/18920

Contribution:

AB developed the research design and evaluation methodology for the paper, reviewed the literature, created the prototype, performed all evaluations, produced and interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

I acknowledge the above statement of contribution is accurate:

Sean Randall: _____

Brown AP, Borgs C, Randall SM, Schnell R (2017). *Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets.* BMC Medical Informatics and Decision Making, 17(1), 83. https://doi.org/10.1186/s12911-017-04785

Contribution:

AB contributed to the research design and evaluation methodology for the paper, reviewed the literature, ran the simulations, produced and interpreted the results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

I acknowledge the above statement of contribution is accurate:

Sean Randall: _____

Christian Borgs: _____

Rainer Schnell: _____

Randall SM, **Brown AP**, Ferrante AM, Boyd JH (2019). *Privacy preserving linkage using multiple match-keys.* International Journal of Population Data Science, 4(1). https://doi.org/10.23889/ijpds.v4i1.1094

Contribution:

AB contributed to the research design and evaluation methodology for the paper, reviewed the new linkage methodology and evaluation results, critically reviewed the manuscript, and approved the final manuscript.

I acknowledge the above statement of contribution is accurate:

Sean Randall: _____

Anna Ferrante: _____

James Boyd: _____

Randall SM, Ferrante AM, Boyd JH, **Brown AP**, Semmens JB (2016). *Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581?* Health Information Management Journal. https://doi.org/10.1177/1833358316647587

Contribution:

AB contributed to the research design and evaluation methodology for the paper, reviewed the results, critically reviewed the manuscript and approved the final manuscript.

I acknowledge the above statement of contribution is accurate:

Sean Randall: _____

Anna Ferrante: _____

James Boyd: _____

James Semmens: _____

Randall SM, Ferrante AM, Boyd JH, **Brown AP**, Semmens JB (2015). *Grouping methods for ongoing record linkage*. Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference.

Contribution:

AB contributed to the research design and evaluation methodology for the paper, contributed to the design of the new grouping strategy, reviewed the results, critically reviewed the manuscript and approved the final manuscript.

I acknowledge the above statement of contribution is accurate:

Sean Randall: _____

Anna Ferrante: _____

James Boyd: _____

James Semmens: _____

Randall SM, **Brown AP**, Boyd JH, Ferrante AM, Semmens JB (2015). *Privacy preserving record linkage using homomorphic encryption.* Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference.

Contribution:

AB contributed to the research design and evaluation methodology for the paper, critically reviewed the encryption scheme used for linkage, reviewed the results, critically reviewed the manuscript and approved the final manuscript.

I acknowledge the above statement of contribution is accurate:

Sean Randall: _____

Anna Ferrante: _____

James Boyd: _____

James Semmens: _____

Appendix C

Copyright statements

1. **Brown AP**, Ferrante, AM, Randall, SM, Boyd, JH, Semmens, JB (2017). *Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage.* Frontiers in Public Health, 5(March), 34. https://doi.org/10.3389/fpubh.2017.00034

2. **Brown AP**, Randall SM, Ferrante AM, Semmens JB, Boyd JH (2017). *Estimating parameters for probabilistic linkage of privacy-preserved datasets* BMC Medical Research Methodology, 17(1), 95. https://doi.org/10.1186/s12874-017-0370-0

3. **Brown AP**, Randall, SM, Boyd, JH, Ferrante, AM (2019). *Evaluation of approximate comparison methods on Bloom filters for probabilistic linkage.* International Journal of Population Data Science, 4(1). https://doi.org/10.23889/ijpds.v4i1.1095

4. **Brown AP**, Borgs C, Randall SM, Schnell R (2017). *Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets.* BMC Medical Informatics and Decision Making, 17(1), 83. https://doi.org/10.1186/s12911-017-04785

5. **Brown AP**, Randall SM (2020). *Secure Record Linkage of Large Health Data Sets: Evaluation of a Hybrid Cloud Model.* JMIR Medical Informatics, 8(9), e18920. https://doi.org/10.2196/18920

6. Randall SM, **Brown AP**, Ferrante AM, Boyd JH (2019). *Privacy preserving linkage using multiple match-keys* (2019) International Journal of Population Data Science, 4(1). https://doi.org/10.23889/ijpds.v4i1.1094

7. Randall SM, Ferrante AM, Boyd JH, **Brown AP**, Semmens JB (2016). *Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581?* Health Information Management Journal. https://doi.org/10.1177/1833358316647587

The copyright of the following conference proceedings remains with the authors.

8. Randall SM, Ferrante AM, Boyd JH, **Brown AP**, Semmens JB (2015). *Grouping methods for ongoing record linkage* (2015) Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference.

9. Randall SM, **Brown AP**, Boyd JH, Ferrante AM, Semmens JB (2015). *Privacy preserving record linkage using homomorphic encryption* (2015) Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference.

The following articles require a license for inclusion in this thesis. These licenses are included at the end of this section.

10. Boyd JH, Ferrante AM, Irvine K, Smith M, Moore E, **Brown AP**, Randall SM (2016). *Understanding the Origins of record linkage error and how they affect research outcomes* Australia and New Zealand Journal of Public Health. https://doi.org/10.1111/1753-6405.12597.

The following conference abstracts are distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

11. **Brown AP**, Borgs C, Randall SM, Schnell R (2016). *High quality linkage using Multibit Trees for privacy-preserving blocking.* Conference: International Population Data Linkage Conference, Swansea, Wales, August 2016.

12. Boyd JH, Ferrante AM, Randall SM, **Brown AP**, Semmens JB (2016). *Implementing privacy-preserving record linkage: welcome to the real world.* Conference: International Population Data Linkage Conference, Swansea, Wales, August 2016.

13. **Brown AP**, Randall SM, Ferrante AM, Boyd JH (2018). *Public Cloud: The Future of Record Linkage?* Conference: International Population Data Linkage Conference, Banff, Alberta, Canada, September 2018.

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Nov 09, 2020

This Agreement between Mr. Adrian Brown ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4945120282244
License date	Nov 09, 2020
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Australian and New Zealand Journal of Public Health
Licensed Content Title	Understanding the origins of record linkage errors and how they affect research outcomes
Licensed Content Author	Sean M. Randall, Adrian Brown, Elizabeth Moore, et al
Licensed Content Date	Nov 20, 2016
Licensed Content Volume	41
Licensed Content Issue	2
Licensed Content Pages	1
Type of Use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title	Implementing Privacy-Preserving Record Linkage in a Cloud Computing Environment
Institution name	Curtin University
Expected presentation date	Dec 2020
Requestor Location	Mr. Adrian Brown Bldg 400:237, Kent Street Bentley, WA 6021 Australia Attn: Curtin University
Publisher Tax ID	EU826007151
Total	**0.00 AUD**
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.
- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared

before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Bibliography

[1] Rediet Abebe, Shawndra Hill, Jennifer Wortman Vaughan, Peter M. Small, and H. Andrew Schwartz. "Using search queries to understand health information needs in africa". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 13. ICWSM. 2019, pp. 3–14. arXiv: 1806.05740.

[2] Robert W. Aldridge, Kunju Shaji, Andrew C. Hayward, and Ibrahim Abubakar. "Accuracy of Probabilistic Linkage Using the Enhanced Matching System for Public Health and Epidemiological Studies". In: *PLOS ONE* 10.8 (2015). Ed. by Antonio Guilherme Pacheco, e0136179. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0136179. URL: http://dx.plos.org/10.1371/journal.pone.0136179.

[3] Bethania De Araujo Almeida, Mauricio Lima Barreto, Maria Yuri Ichihara, Marcos Ennes Barreto, Liliana Cabral, Rosemeire Fiaccone, Roberto P Carreiro, Carlos Teles, Robespierre Pita, Gerson Penna, Manoel Barral-Netto, M. Sanni Ali, George Barbosa, Spiros Denaxas, Laura Rodrigues, and Liam Smeeth. "The Center for Data and Knowledge Integration for Health (CIDACS)". In: *International Journal of Population Data Science* 4.2 (2019). ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i2.1140. URL: https://ijpds.org/article/view/1140.

[4] Tavinder Kaur Ark, Sarah Kesselring, Brent Hills, and Kim McGrail. "Population Data British Columbia: A data resource for research". In: *International Journal of Population Data Science* 4.2 (2020). ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i2.1133. URL: https://ijpds.org/article/view/1133.

[5] Samuel J Aronson and Heidi L Rehm. "Building the foundation for genomics in precision medicine". In: *Nature* 526.7573 (2015), pp. 336–342.

[6] Australian Bureau of Statistics. *National Early Childhood Education and Care Collection: Data Collection Guide*. 2013.

[7] Australian Cyber Security Centre. *Cloud Services*. 2020. URL: https://www.cyber.gov.au/acsc/view-all-content/programs/irap/asd-certified-cloud-services (visited on 08/16/2020).

[8] Australian Government. *Australian Government Cloud Computing Policy*. 2014. URL: http://www.finance.gov.au/cloud/.

[9] Australian Government. *High Level Principles for Data Integration involving Commonwealth Data for Statistical and Research Purposes*. Ed. by Cross Portfolio Statistical Integration Committee (CPSIC). Canberra, 2010.

[10] Australian Government. *Privacy Amendment (Enhancing Privacy Protection) Bill 2012, Explanatory Memorandum, in Australian Government (ed)*. 2012.

[11] Australian Government Digital Transformation Agency. *Secure Cloud Strategy*. 2020. URL: https://www.dta.gov.au/our-projects/secure-cloud-strategy (visited on 08/16/2020).

[12] Australian Institute of Health and Welfare. *Community Services Ministers Advisory Council*. Tech. rep. Canberra, 2004.

[13] Australian Institute of Health and Welfare. *Disability Services Minimum Data Set: data guide*. Canberra, 2013.

[14] Australian Institute of Health and Welfare. *Enhancing the Alcohol and Other Drug Treatment Services National Minimum Data Set*. 2013.

[15] Australian Institute of Health and Welfare. *Specialist Homelessness Services Collection Data Quality Statement 2013-14*. 2014. URL: http://meteor.aihw.gov.au/content/index.phtml/itemId/593778 (visited on 04/20/2015).

[16] Tobias Bachteler, Jörg Reiher, and Rainer Schnell. "Similarity Filtering with Multibit Trees for Record Linkage". In: *Nuremberg: German Record Linkage Center* (2013).

[17] G. P. Basharin. "On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables". In: *Theory of Probability & Its Applications* 4.3 (1959), pp. 333–336. ISSN: 0040-585X. DOI: 10.1137/1104033. URL: https://epubs.siam.org/doi/abs/10.1137/1104033.

[18] A J Bass and C Garfield. "Statistical linkage keys: How effective are they?" In: *Symposium on Health Data Linkage*. Sydney 2002, 2002, pp. 40–45.

[19] G John Bauman. *Computation of Weights for Probabilistic Record Linkage using the EM Algorithm*. Tech. rep. 2006, p. 107.

[20] Mihir Bellare, Ran Canetti, and Hugo Krawczyk. "Keying Hash Functions for Message Authentication". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 1996, pp. 1–15. ISBN: 3540615121. DOI: 10.1007/3-540-68697-5_1. URL: http://link.springer.com/10.1007/3-540-68697-5{_}1.

[21] Elisa Bertino, Dan Lin, and Wei Jiang. "A survey of quantification of privacy preserving data mining algorithms". In: *Privacy-preserving data mining*. Springer, 2008, pp. 183–205.

[22] Jiang Bian, Alexander Loiacono, Andrei Sura, Tonatiuh Mendoza Viramontes, Gloria Lipori, Yi Guo, Elizabeth Shenkman, and William Hogan. "Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network". In: *JAMIA Open* 2.4 (2019), pp. 562–569. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooz050. URL: https://academic.oup.com/jamiaopen/article/2/4/562/5585396.

[23] P. P. Biemer. "Total Survey Error: Design, Implementation, and Evaluation". In: *Public Opinion Quarterly* 74.5 (2010), pp. 817–848. ISSN: 0033-362X. DOI: 10.1093/poq/nfq058. URL: https://academic.oup.com/poq/article-lookup/doi/10.1093/poq/nfq058.

[24] Ingrid A Binswanger, Marc F Stern, Richard A Deyo, Patrick J Heagerty, Allen Cheadle, Joann G Elmore, and Thomas D Koepsell. "Release from prison—a high risk of death for former inmates". In: *New England Journal of Medicine* 356.2 (2007), pp. 157–165.

[25] Megan A Bohensky, Damien Jolley, Vijaya Sundararajan, Sue Evans, David V Pilcher, Ian Scott, and Caroline A Brand. "Data linkage: a powerful research tool with potential problems". In: *BMC health services research* 10.1 (2010), pp. 1–7.

[26] Flavio Bonomi, Michael Mitzenmacher, Rina Panigrahy, Sushil Singh, and George Varghese. "An improved construction for counting Bloom filters". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2006. ISBN: 3540388753. DOI: 10.1007/11841036_61.

[27] Luca Bonomi, Rui Chen, and Benjamin C M Fung. "Frequent grams based Embedding for Privacy Preserving Record Linkage". In: (2012), pp. 1597–1601.

[28] Luca Bonomi, Liyue Fan, and Li Xiong. "A Review of Privacy Preserving Mechanisms for Record Linkage". In: *Medical Data Privacy Handbook*. 2015. Chap. 10, pp. 233–265. ISBN: 9783319236322. DOI: 10.1007/978-3-319-23633-9.

[29] Luca Bonomi, Li Xiong, Rui Chen, and B Fung. "Privacy Preserving Record Linkage via grams Projections". In: *arXiv preprint arXiv:1208.2773* (2012). arXiv: arXiv:1208.2773v1. URL: http://arxiv.org/abs/1208.2773.

[30] Luca Bonomi, Li Xiong, and James J Lu. "LinkIT". In: *Proceedings of the 2013 international conference on Management of data - SIGMOD '13*. New York, New York, USA: ACM Press, 2013, p. 1029. ISBN: 9781450320375. DOI: 10.1145/2463676.2465259. URL: http://dl.acm.org/citation.cfm?doid=2463676.2465259.

[31] Murilo Boratto, Pedro Alonso, Clicia Pinto, Pedro Melo, Marcos Barreto, and Spiros Denaxas. "Exploring hybrid parallel systems for probabilistic record linkage". In: *The Journal of Supercomputing* 75.3 (2019), pp. 1137–1149. ISSN: 1573-0484. DOI: 10.1007/s11227-018-2328-3. URL: https://doi.org/10.1007/s11227-018-2328-3.

[32] F. Borst, F. A. Allaert, and C. Quantin. "The Swiss solution for anonymously chaining patient files." In: *Medinfo. MEDINFO* (2001). ISSN: 0926-9630.

[33] Andrew Boulle, Alexa Heekes, Nicki Tiffin, Mariette Smith, Themba Mutemaringa, Nesbert Zinyakatira, Florence Phelanyane, Cara Pienaar, Kasturi Buddiga, Eduan Coetzee, Renier Van Rooyen, Robin Dyers, Naadir Fredericks, Adam Loff, Lesley Shand, Melvin Moodley, Ian De Vega, and Krish Vallabhjee. "Data Centre Profile: The Provincial Health Data Centre of the Western Cape Province, South Africa". In: *International Journal of Population Data Science* 4.2 (2019). ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i2.1143. URL: https://ijpds.org/article/view/1143.

[34] J. H. Boyd, T. Guiver, S. M. Randall, A. M. Ferrante, J. B. Semmens, P. Anderson, and T. Dickinson. "A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects". In: *Methods of Information in Medicine* 55.3 (2016), pp. 276–283. ISSN: 0026-1270. DOI: 10.3414/ME15-01-0152. URL: http://www.schattauer.de/index.php?id=1214{\&}doi=10.3414/ME15-01-0152.

[35] James Boyd, Anna Ferrante, Adrian Brown, Sean Randall, and James Semmens. "Implementing privacy-preserving record linkage: welcome to the real world". In: *International Journal of Population Data Science* 1.1 (2017). ISSN: 2399-4908. DOI: 10.23889/ijpds.v1i1.153. URL: https://ijpds.org/article/view/153.

[36] James Boyd, Yuen Ai Lee, Adrian Brown, Sean Randall, and Anna Ferrante. "Unlocking the potential of health systems using privacy preserving record linkage: A pilot project exploring the research potential of developing a linkable general practice dataset". In: *International Journal of Population Data Science* 4.3 (2019). ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i3.1231. URL: https://ijpds.org/article/view/1231.

[37] James H Boyd. "The Scottish Record Linkage System". Edinburgh, UK, 2005.

[38] James H Boyd, Anna M Ferrante, Christine M O'Keefe, Alfred J Bass, Sean M Randall, and James B Semmens. "Data linkage infrastructure for cross-jurisdictional health-related research in Australia". In: *BMC Health Services Research* 12.1 (2012), p. 480. ISSN: 1472-6963. DOI: 10.1186/1472-6963-12-480. URL: https://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-12-480.

[39] James H Boyd, Sean Randall, Adrian P Brown, Max Maller, Davie Botes, Margo Gillies, and Anna Ferrante. "Population Data Centre Profiles: Centre for Data Linkage". In: *International Journal of Population Data Science* 4.2 (2020). ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i2.1139. URL: https://ijpds.org/article/view/1139.

[40] James H Boyd, Sean M Randall, and Anna M Ferrante. "Application of Privacy-Preserving Techniques in Operational Record Linkage Centres". In: *Medical Data Privacy Handbook*. Springer, 2015, pp. 267–287.

[41] James H Boyd, Sean M Randall, Anna M Ferrante, Jacqueline K Bauer, Adrian P Brown, and James B Semmens. "Technical challenges of providing record linkage services for research". In: *BMC Medical Informatics and Decision Making* 14.23 (2014), p. 23. ISSN: 1472-6947. DOI: 10.1186/1472-6947-14-23. URL: http://www.biomedcentral.com/1472-6947/14/23.

[42] James H Boyd, Sean M Randall, Anna M Ferrante, Jacqueline K Bauer, Kevin McInneny, Adrian P Brown, Katrina Spilsbury, Margo Gillies, and James B Semmens. "Accuracy and completeness of patient pathways – the benefits of national data linkage in Australia". In: *BMC Health Services Research* 15.1 (2015), p. 312. ISSN: 1472-6963. DOI: 10.1186/s12913-015-0981-2. URL: https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-015-0981-2.

[43] Matt Boyd, June Atkinson, and Tony Blakely. "Ethnic counts on mortality, New Zealand Cancer Registry and census data: 2006–2011". In: *NZ Med J* 129.1429 (2016), pp. 22–39.

[44] Douglas Iain Ross Boyle and Naomi Rafael. "BioGrid Australia and GRHANITE™: Privacy-protecing subject matching". In: *Studies in Health Technology and Informatics*. 2011. ISBN: 9781607507901. DOI: 10.3233/978-1-60750-791-8-24.

[45] Cathy J Bradley, Lynne Penberthy, Kelly J Devers, and Debra J Holden. "Health services research and data linkages: issues, methods, and directions for the future". In: *Health services research* 45.5p2 (2010), pp. 1468–1488.

[46] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. "(Leveled) fully homomorphic encryption without bootstrapping". In: *ACM Transactions on Computation Theory (TOCT)* 6.3 (2014), pp. 1–36.

[47] Zvika Brakerski and Vinod Vaikuntanathan. "Fully homomorphic encryption from ring-LWE and security for key dependent messages". In: *Annual cryptology conference*. Springer. 2011, pp. 505–524.

[48] E L Brook, D L Rosman, C D J Holman, and B Trutwein. "Summary report: research outputs project, WA Data Linkage Unit (1995-2003)". In: *Western Australian Data Linkage Unit Perth* (2005).

[49] Emma L Brook, Diana L Rosman, and C D'Arcy J Holman. "Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System". In: *Australian and New Zealand Journal of Public Health* 32.1 (2008), pp. 19–23.

[50] Adrian Brown, Christian Borgs, Sean Randall, Rainer Schnell, and Christian Borgs Adrian Brown Sean Randall and Rainer Schnell. "High quality linkage using Multibit Trees for privacy-preserving blocking". In: *2016 International Population Data Linkage Conference* 1.1 (2016). DOI: http://dx.doi.org/10.23889/ijpds.v1i1.149.

[51] Adrian P Brown, Christian Borgs, Sean M Randall, and Rainer Schnell. "Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets". In: *BMC Medical Informatics and Decision Making* 17.1 (2017), p. 83. ISSN: 1472-6947. DOI: 10.1186/s12911-017-0478-5. URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0478-5.

[52] Adrian P. Brown, Anna M. Ferrante, Sean M. Randall, James H. Boyd, and James B. Semmens. "Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage". In: *Frontiers in Public Health* 5.MAR (2017), p. 34. ISSN: 2296-2565. DOI: 10.3389/fpubh.2017.00034. URL: https://www.frontiersin.org/articles/10.3389/fpubh.2017.00034/full.

[53] Adrian P. Brown, Sean M. Randall, Anna M. Ferrante, James B. Semmens, and James H. Boyd. "Estimating parameters for probabilistic linkage of privacy-preserved datasets". In: *BMC Medical Research Methodology* 17.1 (2017), p. 95. ISSN: 14712288. DOI: 10.1186/s12874-017-0370-0. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0370-0.

[54] Deutscher Bundestag. "Gesetz ueber Krebsregister (Krebsregistergesetz KRG)". In: *Bundesgesetzblatt (in German)* 79.1 (1994), p. 994.

[55] Joan A Casey, Brian S Schwartz, Walter F Stewart, and Nancy E Adler. "Using electronic health records for population health research: a review of methods and applications". In: *Annual review of public health* 37 (2016).

[56] Centre for Big Data Research in Health (CBDRH). *Annual Report*. Tech. rep. 2015. URL: https://cbdrh.med.unsw.edu.au/sites/default/files/CBDRH{_}AnnualReport{_}2015{_}160609{_}Final.pdf.

[57] Centre for Health Record Linkage. *Quality Assurance*. 2015. URL: http://www.cherel.org.au/quality-assurance (visited on 06/03/2015).

[58] Xiao Chen, Eike Schallehn, and Gunter Saake. "Cloud-Scale Entity Resolution: Current State and Open Challenges". In: *Open Journal of Big Data (OJBD)* 4.1 (2018). URL: http://www.ronpub.com/ojbd.

[59] Peter Christen. *A comparison of personal name matching: Techniques and practical issues*. Tech. rep. TR-CS-06-02. Canberra, 2006, pp. 290–294. DOI: 10.1109/icdmw.2006.2.

[60] Peter Christen. "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication". In: *IEEE Transactions on Knowledge and Data Engineering* 24.9 (2012), pp. 1537–1555. ISSN: 1041-4347. DOI: 10.1109/TKDE.2011.127. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5887335.

[61] Peter Christen. "Advanced computational and privacy methods for data linkage". In: *2016 International Population Data Linkage Conference*. Swansea, Wales: Swansea University, 2016.

[62] Peter Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Berlin/Heidelberg, Germany: Springer Science & Business Media, 2012.

[63] Peter Christen and Tim Churches. *Febrl - Freely extensible biomedical record linkage*. Tech. rep. October. Australian National University, 2002. URL: https://digitalcollections.anu.edu.au/handle/1885/40723.

[64] Peter Christen, Tim Churches, and Markus Hegland. "Febrl–a parallel open source data linkage system". In: *Advances in knowledge discovery and data mining* (2004), pp. 638–647.

[65] Peter Christen and Agus Pudjijono. "Accurate synthetic generation of realistic personal information". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 5476 LNAI (2009), pp. 507–514. ISSN: 03029743. DOI: 10.1007/978-3-642-01307-2_47.

[66] Peter Christen, Thilina Ranbaduge, Dinusha Vatsalan, and Rainer Schnell. "Precise and Fast Cryptanalysis for Bloom Filter Based Privacy-Preserving Record Linkage". In: *IEEE Transactions on Knowledge and Data Engineering* 31.11 (2019), pp. 2164–2177. ISSN: 1041-4347. DOI: 10.1109/TKDE.2018.2874004. URL: https://ieeexplore.ieee.org/document/8481521/.

[67] Peter Christen, Rainer Schnell, Dinusha Vatsalan, and Thilina Ranbaduge. "Efficient Cryptanalysis of Bloom Filters for Privacy-Preserving Record Linkage". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. 2017.

[68] Xu Chu, IF Ihab F Ilyas, and Paraschos Koutris. "Distributed Data Deduplication". In: *Proceedings of the Very Large Data Bases Endowment* 9.11 (2016), pp. 864–875. ISSN: 21508097. DOI: 10.14778/2983200.2983203. URL: https://pdfs.semanticscholar.org/b9a7/2b6cd8a37e36a5bdf5c672823ab0f2211343.pdf.

[69] D E Clark and D R Hahn. "Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry". In: *Proceedings from the Annual Symposium on Computer Application in Medical Care*. 1995, pp. 397–401.

[70] T Clohessy, T Acton, L Morgan, and K Conboy. "The times they are a-changin for ICT service provision: A cloud computing business model perspective". In: *24th European Conference on Information Systems ECIS* (2016), pp. 1–15. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84995803001{\&}partnerID=40{\&}md5=802997b159b74d6b206feb56124f9b95.

[71] Community Services Ministers Advisory Council. *Statistical Data Linkage in Community Services Data Collections*. Tech. rep. 2004, p. 102. URL: https://www.aihw.gov.au/reports/technical-report/statistical-data-linkage-community-services-data/contents/table-of-contents.

[72] Karen T. Copeland, Harvey Checkoway, Anthony J. Mcmichael, and Robert H. Holbrook. "Bias due to misclassification in the estimation of relative risk". In: *American Journal of Epidemiology* (1977). ISSN: 00029262. DOI: 10.1093/oxfordjournals.aje.a112408.

[73] Council of European Union. *Council regulation (EU) no 679/2016*. 2016.

[74] Chris Culnane, Benjamin I. P. Rubinstein, and Vanessa Teague. "Options for encoding names for data linking at the Australian Bureau of Statistics". In: (2018). arXiv: 1802.07975. URL: http://arxiv.org/abs/1802.07975.

[75] Curtin Data Linkage. *LinXmart*. 2018. URL: https://linxmart.com.au.

[76] Ivan Damgård, Valerio Pastro, Nigel Smart, and Sarah Zakarias. "Multiparty computation from somewhat homomorphic encryption". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2012. ISBN: 9783642320088. DOI: 10.1007/978-3-642-32009-5_38.

[77] Tharaka L Dassanayake, Alison L Jones, Patricia T Michie, Gregory L Carter, Patrick McElduff, Barrie J Stokes, and Ian M Whyte. "Risk of Road Traffic Accidents in Patients Discharged Following Treatment for Psychotropic Drug Overdose". In: *CNS drugs* 26.3 (2012), pp. 269–276.

[78] Department of the Premier and Cabinet. *WA Government Cloud Policy*. 2020. URL: https://www.wa.gov.au/government/publications/cloud-policy.

[79] Lee R. Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (1945), pp. 297–302. ISSN: 00129658. DOI: 10.2307/1932409. URL: http://doi.wiley.com/10.2307/1932409.

[80] Chenxiao Dou, Daniel Sun, Yi-Cheng Chen, Guoqiang Li, and Jianquan Liu. "Probabilistic parallelisation of blocking non-matched records for big data". In: *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 3465–3473. ISBN: 978-1-4673-9005-7. DOI: 10.1109/BigData.2016.7841009. URL: http://ieeexplore.ieee.org/document/7841009/.

[81] Wenliang Du and Mikhail J Atallah. "Secure multi-party computation problems and their applications: a review and open problems". In: *Proceedings of the 2001 workshop on New security paradigms*. 2001, pp. 13–22.

[82] Wenliang Du and Zhijun Zhan. "A practical approach to solve secure multi-party computation problems". In: *Proceedings New Security Paradigms Workshop*. 2002. DOI: 10.1145/844123.844125.

[83] Halbert L Dunn. "Record Linkage". In: *American Journal of Public Health and the Nations Health* 36.12 (1946), pp. 1412–1416. ISSN: 0002-9572. DOI: 10.2105/AJPH.36.12.1412. URL: http://ajph.aphapublications.org/doi/10.2105/AJPH.36.12.1412.

[84] Elizabeth Durham, Yuan Xue, Murat Kantarcioglu, and Bradley Malin. "Private medical record linkage with approximate matching." In: *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2010 (2010), pp. 182–186. ISSN: 1942-597X.

[85] Elizabeth Durham, Yuan Xue, Murat Kantarcioglu, and Bradley Malin. "Quantifying the Correctness, Computational Complexity, and Security of Privacy-Preserving String Comparators for Record Linkage." In: *An international journal on information fusion* 13.4 (2012), pp. 245–259. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2011.04.004. URL: https://www.sciencedirect.com/science/article/pii/S1566253511000340.

[86] Elizabeth A Durham, Murat Kantarcioglu, Senior Member, Yuan Xue, Csaba Toth, Mehmet Kuzu, and Bradley Malin. "Composite Bloom Filters for Secure Record Linkage". In: *IEEE Transactions on Knowledge and Data Engineering* 26.12 (2014), pp. 2956–2968.

[87] Elizabeth Ashley Durham. "A Framework for Accurate, Efficient Private Record Linkage". PhD thesis. 2012. URL: http://etd.library.vanderbilt.edu/available/etd-03262012-144837/unrestricted/dissertation.pdf.

[88] L. Dusserre, C. Quantin, and H. Bouzelat. "A one way public key cryptosystem for the linkage of nominal files in epidemiological studies." In: *Medinfo. MEDINFO* 8 Pt 1 (1995), pp. 644–647.

[89] Scott L. DuVall, Richard A. Kerber, and Alun Thomas. "Extending the Fellegi–Sunter probabilistic record linkage method for approximate field comparators". In: *Journal of Biomedical Informatics* 43.1 (2010), pp. 24–30. ISSN: 15320464. DOI: 10.1016/j.jbi.2009.08.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S1532046409001051.

[90] "Ebola and big data - Call for help". In: *The Economist* (2014). URL: http://www.economist.com/news/leaders/21627623-mobile-phone-records-are-invaluable-tool-combat-ebola-they-should-be-made-available.

[91] Khaled El Emam, Fida K Dankar, Régis Vaillancourt, Tyson Roffey, and Mary Lysyk. "Evaluating the risk of re-identification of patients from hospital prescription records". In: *The Canadian journal of hospital pharmacy* 62.4 (2009), p. 307.

[92] Randa M. Abd El-Ghafar, Mervat H. Gheith, Ali H. El-Bastawissy, and Eman S. Nasr. "Record linkage approaches in big data: A state of art study". In: *2017 13th International Computer Engineering Conference (ICENCO)*. IEEE, 2017, pp. 224–230. ISBN: 978-1-5386-4266-5. DOI: 10.1109/ICENCO.2017.8289792. URL: http://ieeexplore.ieee.org/document/8289792/.

[93] Josie M M Evans and Thomas M MacDonald. "Record-linkage for pharmacovigilance in Scotland". In: *British journal of clinical pharmacology* 47.1 (1999), pp. 105–110.

[94] Li Fan, Pei Cao, Jussara Almeida, and Andrei Z. Broder. "Summary cache: A scalable wide-area Web cache sharing protocol". In: *IEEE/ACM Transactions on Networking* (2000). ISSN: 10636692. DOI: 10.1109/90.851975.

[95] *Farr Institute*. 2014. URL: http://www.farrinstitute.org/.

[96] Ivan Fellegi and Alan Sunter. "A Theory for Record Linkage". In: *Journal of the American Statistical Association* 64 (1969), pp. 1183–1210.

[97] A Ferrante and J Boyd. "A transparent and transportable methodology for evaluating Data Linkage software". In: *Journal of Biomedical Informatics* 45.1 (2012), pp. 165–172.

[98] Anna Ferrante, James Boyd, Tom Eitelhuber, Sean Randall, Adrian Brown, Max Maller, Davie Botes, and Kurt Sibma. "Using data linkage innovation and collaboration to create a cross-sectoral data repository for Western Australia". In: *International Journal of Population Data Science* 4.3 (2019). ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i3.1233. URL: https://ijpds.org/article/view/1233.

[99] David V Ford, Kerina H Jones, Jean-Philippe Verplancke, Ronan A Lyons, Gareth John, Ginevra Brown, Caroline J Brooks, Simon Thompson, Owen Bodger, Tony Couch, and Ken Leake. "The SAIL Databank: building a national architecture for e-health research and evaluation ". In: *BMC Health Services Research 2009* 9.157 (2009). DOI: doi:10.1186/1472-6963-9-157.

[100] Luca Gagliardelli, Giovanni Simonini, Domenico Beneventano, and Sonia Bergamaschi. "SparkER: Scaling entity resolution in spark". In: *Advances in Database Technology - EDBT* 2019-March (2019), pp. 602–605. ISSN: 23672005. DOI: 10.5441/002/edbt.2019.66.

[101] Gartner. *Predicts 2019: Cloud Adoption and Increasing Regulation Will Drive Investment in IT Vendor Management*. Tech. rep. 2019. URL: https://www.gartner.com/en/documents/3896211/predicts-2019-cloud-adoption-and-increasing-regulation-w.

[102] Leonardo Gazzarri and Melanie Herschel. "Towards task-based parallelization for entity resolution". In: *SICS Software-Intensive Cyber-Physical Systems* 35.1-2 (2019), pp. 31–38. ISSN: 2524-8510. DOI: 10.1007/s00450-019-00409-6. URL: http://link.springer.com/10.1007/s00450-019-00409-6.

[103] Craig Gentry. "A fully homomorphic encryption scheme". Thesis. 2009.

[104] Leicester Gill. *Methods for automatic record matching and linkage and their use in national statistics*. 25. Office for National Statistics, 2001.

[105] Leicester E. Gill. "OX-LINK: The Oxford Medical Record Linkage System". In: *Record Linkage Techniques* (1997), pp. 15–33.

[106] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. "Detecting influenza epidemics using search engine query data". In: *Nature* 457.7232 (2009), pp. 1012–1014.

[107] Aris Gkoulalas-Divanis and Grigorios Loukides. *Medical data privacy handbook*. Springer, 2015.

[108] Karl Goiser, Peter Christen, and Karl Goiser. "Quality and Complexity Measures for Data Linkage and Deduplication". In: *Quality Measures in Data Mining Studies in Computational Intelligence*. Ed. by F Guillet and H Hamilton. Vol. 43. Springer, 2007, pp. 127–151. ISBN: 3540449116. DOI: 10.1007/978-3-540-44918-8_6.

[109] O. Goldreich, S. Micali, and A. Wigderson. "How to play ANY mental game". In: *Proceedings of the nineteenth annual ACM conference on Theory of computing - STOC '87*. New York, New York, USA: ACM Press, 1987, pp. 218–229. ISBN: 0897912217. DOI: 10.1145/

28395.28420. URL: http://portal.acm.org/citation.cfm?doid=28395.28420.

[110] Harvey Goldstein and Katie Harron. "Record linkage: a missing data problem". In: *Methodological Developments in Data Linkage* (2015).

[111] Shanti Gomatam, Randy Carter, Mario Ariet, and Glenn Mitchell. "An Empirical Comparison of Record Linkage Procedures". In: *Statistics in Medicine* 21 (2002), pp. 1485–1496. DOI: 10.1002/SIM.1147.

[112] Sergey Gorbunov, Vinod Vaikuntanathan, and Hoeteck Wee. "Functional encryption with bounded collusions via multi-party computation". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2012. ISBN: 9783642320088. DOI: 10.1007/978-3-642-32009-5_11.

[113] Christopher Graham. "Anonymisation: managing data protection risk code of practice". In: *Information Commissioner's Office* (2012).

[114] Shaun J Grannis, J Marc Overhage, Siu Hui, and Clement J McDonald. "Analysis of a probabilistic record linkage technique without human review." In: *American Medical Infomatics Association* Figure 2 (2003), pp. 259–263. ISSN: 1942-597X. DOI: D030003683.

[115] Shaun J Grannis, J Marc Overhage, and Clement J McDonald. "Analysis of identifier performance using a deterministic linkage algorithm (ABSTRACT)". In: *PubMed Central* Proceeding (2002), pp. 305–309.

[116] E Green Ritchie, F., Mytton, J., Webber, D. J., Deave, T., Montgomery, A., Woolfrey, L., ul-Baset, K. and Chowdhury, S. *Enabling data linkage to maximise the value of public health research data*. Tech. rep. Wellcome Trust, London, 2015.

[117] Lifang Gu, Rohan Baxter, Deanne Vickers, Chris Rainsford, and Citeseer. *Record Linkage: Current Practice and Future Directions*. Tech. rep. 2003, p. 83.

[118] Tenniel Guiver, ABS, and Analytical Services Branch. *Sampling-Based Clerical Review Methods in Probabilistic Linking*. Tech. rep. ABS Cat. no.1351.0.55.034. ABS Website (http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1351.0.55.034May%202011?OpenDocumen 2011, p. 22. URL: http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1351.0.55.034May2011?OpenDocument.

[119] Thanyatida Gunadham and Pramote Kuacharoen. "Security Concerns in Cloud Computing for Knowledge Management Systems". In: *Journal of Applied Statistics and Information Technology* 1.2 (2019), pp. 52–60. ISSN: 2465-4949.

[120] David Hand and Peter Christen. "A note on using the F-measure for evaluating record linkage algorithms". In: *Statistics and Computing* 28.3 (2018), pp. 539–547. ISSN: 0960-3174. DOI: 10.1007/s11222-017-9746-6. URL: http://link.springer.com/10.1007/s11222-017-9746-6.

[121] Julie Harris. "Next Generation Linkage Management System". In: *Sixth Australiasian Workshop on Health Informations and Knowledge Management*. Ed. by Gray K and A Koronios. Vol. 142. Hikm. Adelaide, Australia: Australian Computer Society, 2013, pp. 7–12. ISBN: 978-1-921770-27-2.

[122] Katie Harron, Angie Wade, Ruth Gilbert, Berit Muller-Pebody, and Harvey Goldstein. "Evaluating bias due to data linkage error in electronic healthcare records". In: *BMC medical research methodology* 14.1 (2014), p. 36.

[123] Katie Harron, Angie Wade, Berit Muller-Pebody, Harvey Goldstein, and Ruth Gilbert. "Opening the black box of record linkage". In: *Journal of epidemiology and community health* 66.12 (2012), p. 1198.

[124] Oktie Hassanzadeh and Fei Chiang. "Framework for evaluating clustering algorithms in duplicate detection". In: *Proceedings of the VLDB ...* 2.1 (2009), pp. 1282–1293. ISSN: 21508097. DOI: 10.14778/1687627.1687771. URL: https://dl.acm.org/doi/10.14778/1687627.1687771.

[125] Mauricio A Hernández and Salvatore J Stolfo. "Real-world data is dirty: Data cleansing and the merge/purge problem". In: *Data mining and knowledge discovery* 2.1 (1998), pp. 9–37.

[126] Thomas N Herzog, Fritz J Scheuren, and William E Winkler. *Data quality and record linkage techniques*. Vol. 1. Springer, 2007.

[127] Diana Hetzel. "Data linkage research - can we reap benefits for society without compromising public confidence?" In: *Australian Health Consumer* 2 (2005), pp. 27–28.

[128] M Hobbs and M G McCall. "Health statistics and record linkage in Australia". In: *Journal of Chronic Disease* 23 (1970), pp. 375–381.

[129] C D'Arcy J Holman, John A Bass, Diana L. Rosman, Merran B. Smith, James B. Semmens, Emma J. Glasson, Emma L. Brook, Brooke Trutwein, Ian L. Rouse, Charles R. Watson, Nicholas H. de Klerk, and Fiona J. Stanley. "A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system". In: *Australian Health Review* 32.4 (2008), p. 766. ISSN: 0156-5788. DOI: 10.1071/AH080766. URL: http://www.publish.csiro.au/?paper=AH080766.

[130] Cashel D'Arcy James Holman. *Anonymity and Research: Health Data and Biospecimen Law in Australia*. Uniprint, UWA, 2012. ISBN: 0646574175.

[131] D Holman, A J Bass, I Rouse, and M Hobbs. "Population-based linkage of health records in Western Australia: Development of a health services research linked database". In: *Australian and New Zealand Journal of Public Health* 23 (1999).

[132] Holman CDJ, Meslin E, and Stanley F. "Privacy protectionism and health information: is there any redress for harms to health?" In: *Journal of Law and Medicine* 21.2 (2013), pp. 473–485.

[133] International Mathematical and Statistical Libraries Inc. *User's Manual*. English. Houston TX, 1984.

[134] Katie Irvine, Rick Hall, and Lee Taylor. "Centre for Health Record Linkage". In: *International Journal of Population Data Science* 4.2 (2019). ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i2.1142. URL: https://ijpds.org/article/view/1142.

[135] Katie Irvine and Stephanie Hollis. "Multiple operating models for data linkage: A privacy positive". In: *2016 International Population Data Linkage Conference*. Swansea, Wales: Swansea University, 2016.

[136] Katie Irvine, Michael Smith, Reinier De Vos, Adrian Brown, Anna Ferrante, James Boyd, and Sarah Thackway. "Real world performance of privacy preserving record linkage". In: *International Journal of Population Data Science* 3.4 (2018). ISSN: 2399-4908. DOI: 10.23889/ijpds.v3i4.990. URL: https://ijpds.org/article/view/990.

[137] Katie A Irvine and Lee K Taylor. "The Centre for Health Record Linkage: fostering population health research in NSW". In: *New South Wales public health bulletin* 22.2 (2011), pp. 17–18.

[138] Robert Isele and Christian Bizer. "Learning expressive linkage rules using genetic programming". In: *arXiv preprint arXiv:1208.0291* (2012).

[139] IT News. *Vodafone sacks staff over alleged security breach*. 2011. URL: http://www.itnews.com.au/News/244672,vodafone-sacks-staff-over-alleged-security-breach.aspx.

[140] Matthew A Jaro. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa , Florida". In: *Journal of the American Statistical Association* 84.406 (1989), pp. 414–420.

[141] Matthew A Jaro. "Probabilistic linkage of large public health data files". In: *Statistics in Medicine* 14 (1995), pp. 491–498. ISSN: 02776715. DOI: 10.1002/sim.4780140510. URL: http://dx.doi.org/10.1002/sim.4780140510.

[142] Jomina John and Jasmine Norman. "Major Vulnerabilities and Their Prevention Methods in Cloud Computing". In: *Advances in Big Data and Cloud Computing*. Ed. by J Dinesh Peter, Amir H Alavi, and Bahman Javadi. Singapore: Springer Singapore, 2019, pp. 11–26. ISBN: 978-981-13-1882-5.

[143] Kerina H Jones, David Vincent Ford, Simon Thompson, and Ronan Lyons. "A Profile of the SAIL Databank on the UK Secure Research Platform". In: *International Journal of Population Data Science* 4.2 (2019). ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i2.1134. URL: https://ijpds.org/article/view/1134.

[144] Douglas P. Jutte, Leslie L. Roos, and Marni D. Brownell. "Administrative Record Linkage as a Tool for Public Health Research". In: *Annual Review of Public Health* 32.1 (2011), pp. 91–108. ISSN: 0163-7525. DOI: 10.1146/annurev-publhealth-031210-100700. URL: http://www.annualreviews.org/doi/10.1146/annurev-publhealth-031210-100700.

[145] Murat Kantarcioglu, Ali Inan, Wei Jiang, and Bradley Malin. "Formal anonymity models for efficient privacy-preserving joins". In: *Data & Knowledge Engineering* 68.11 (2009), pp. 1206–1223.

[146] Alexandros Karakasidis and Vassilios S. Verykios. "Secure Blocking + Secure Matching = Secure Record Linkage". In: *Journal of Computing Science and Engineering* (2011). ISSN: 1976-4677. DOI: 10.5626/jcse.2011.5.3.223.

[147] Dimitrios Karapiperis, Aris Gkoulalas-Divanis, and Vassilios S. Verykios. "LSHDB: a parallel and distributed engine for record linkage and similarity search". In: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2016, pp. 1–4. ISBN: 978-1-5090-5910-2. DOI: 10.1109/ICDMW.2016.7867099. URL: http://ieeexplore.ieee.org/document/7867099/.

[148] Dimitrios Karapiperis and Vassilios Verykios. "A distributed near-optimal LSH-based framework for privacy-preserving record linkage". In: *Computer Science and Information Systems* 11.2 (2014), pp. 745–763. ISSN: 1820-0214. DOI: 10.2298/CSIS140215040K. URL: http://www.doiserbia.nb.rs/Article.aspx?ID=1820-02141400040K.

[149] Dimitrios Karapiperis and Vassilios S. Verykios. "A fast and efficient Hamming LSH-based scheme for accurate linkage". In: *Knowledge and Information Systems* 49.3 (2016), pp. 861–884. ISSN: 0219-1377. DOI: 10.1007/s10115-016-0919-y. URL: http://link.springer.com/10.1007/s10115-016-0919-y.

[150] Dimitrios Karapiperis and Vassilios S. Verykios. "An LSH-based Blocking Approach with a Homomorphic Matching Technique for Privacy-Preserving Record Linkage". In: *IEEE Transactions on Knowledge and Data Engineering* PP.99 (2014), pp. 1–1. ISSN: 1041-4347. DOI: 10.1109/TKDE.2014.2349916. URL: https://ieeexplore.ieee.org/document/6880802.

[151] Dimitrios Karapiperis and Vassilios S. Verykios. "Load-Balancing the Distance Computations in Record Linkage". In: *ACM SIGKDD Explorations Newsletter* 17.1 (2015), pp. 1–7. ISSN: 19310145. DOI: 10.1145/2830544.2830546. URL: http://dl.acm.org/citation.cfm?id=2830544.2830546.

[152] R Karmel. "Transitions between aged care services". In: *Data Linkage Series no. 2* cat. no. CSI 2. (2005).

[153] Rosemary Karmel. *Data linkage protocols using a statistical linkage key*. Ed. by Australian Institute of Health and Welfare. Canberra Australia, 2005.

[154] Rosemary Karmel, Phil Anderson, Diane Gibson, Ann Peut, Stephen Duckett, and Yvonne Wells. "Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study". In: *BMC Health Services Research* 10.1 (2010), p. 41. ISSN: 1472-6963. DOI: 10.1186/1472-6963-10-41. URL: https://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-10-41.

[155] Rosemary Karmel and Diana Rosman. "Linkage of health and aged care service events: comparing linkage and event selection methods". In: *BMC Health services research* 8.1 (2008), p. 149.

[156] Alan F. Karr. "The Role of Statistical Disclosure Limitation in Total Survey Error". In: *Total Survey Error in Practice*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2017, pp. 71–94. DOI: 10.1002/9781119041702.ch4. URL: https://onlinelibrary.wiley.com/doi/10.1002/9781119041702.ch4.

[157] Alan Katz, Jennifer Enns, Mark Smith, Charles Burchill, Ken Turner, and Dave Towns. "Population Data Centre Profile: The Manitoba Centre for Health Policy". In: *International Journal of Population Data Science* 4.2 (2020). ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i2.1131. URL: https://ijpds.org/article/view/1131.

[158] C.W. W Kelman, A.J. J Bass, and C.D.J. D J Holman. "Research use of linked health data - a best practice protocol". In: *Australian and New Zealand Journal of Public Health* 26.3

(2002), pp. 251–255. ISSN: 13260200. DOI: 10.1111/j.1467-842X.2002.tb00682.x. URL: http://doi.wiley.com/10.1111/j.1467-842X.2002.tb00682.x.

[159] S W Kendrick and J A Clarke. "The Scottish Medical Record Linkage System". In: *Health Bulletin (Edinburgh)* 51 (1979), pp. 72–79.

[160] Steve W. Kendrick, M. M. Douglas, D. Gardner, and D. Hucker. "Best-link matching of scottish health data sets". In: *Methods of Information in Medicine* (1998). ISSN: 00261270. DOI: 10.1055/s-0038-1634494.

[161] Minhaj Ahmad Khan. "A survey of security issues for cloud computing". In: *Journal of Network and Computer Applications* 71 (2016), pp. 11–29. ISSN: 10848045. DOI: 10.1016/j.jnca.2016.05.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S1084804516301060.

[162] Muin J Khoury, Michael F Iademarco, and William T Riley. "Precision public health for the era of precision medicine". In: *Am J Prev Med* (2015), p. 1.

[163] Adam Kirsch and Michael Mitzenmacher. "Distance-sensitive bloom filters". In: *Proceedings of the 8th Workshop on Algorithm Engineering and Experiments and the 3rd Workshop on Analytic Algorithms and Combinatorics*. 2006. ISBN: 0898716101. DOI: 10.1137/1.9781611972863.4.

[164] Lea Kissner and Dawn Song. "Privacy-preserving set operations". In: *Annual International Cryptology Conference*. Springer. 2005, pp. 241–257.

[165] Predrag Klasnja and Wanda Pratt. "Healthcare in the pocket: mapping the space of mobile-phone health interventions". In: *Journal of biomedical informatics* 45.1 (2012), pp. 184–198.

[166] Lars Kolb and Erhard Rahm. "Parallel Entity Resolution with Dedoop". In: *Datenbank-Spektrum* 13.1 (2012), pp. 23–32. ISSN: 1618-2162. DOI: 10.1007/s13222-012-0110-x. URL: http://link.springer.com/10.1007/s13222-012-0110-x.

[167] Lars Kolb, Andreas Thor, and Erhard Rahm. "Block-based load balancing for entity resolution with MapReduce". In: *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*. New York, New York, USA: ACM Press, 2011, p. 2397. ISBN: 9781450307178. DOI: 10.1145/2063576.2063976. URL: http://dl.acm.org/citation.cfm?id=2063576.2063976.

[168] Lars Kolb, Andreas Thor, and Erhard Rahm. "Load Balancing for MapReduce-based Entity Resolution". In: *2012 IEEE 28th International Conference on Data Engineering*. IEEE, 2012, pp. 618–629. ISBN: 978-0-7695-4747-3. DOI: 10.1109/ICDE.2012.22. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6228119.

[169] H. Krawczyk, M. Bellare, and R. Canetti. *RFC2104 - HMAC: Keyed-hashing for message authentication*. Tech. rep. 1997.

[170] Thomas G Kristensen, Jesper Nielsen, Christian NS Pedersen, JJ Irwin, BK Shoichet, VJ Gillet, P Willett, J Bradshaw, AR Leach, VJ Gillet, P Willett, P Willett, JM Barnard, GM Downs, SJ Swamidass, P Baldi, A Smellie, P Baldi, DS Hirschberg, RJ Nasr, N Saitou, M Nei, C Steinbeck, Y Han, S Kuhn, O Horlacher, E Luttmann, and E Willighagen. "A tree-based method for the rapid screening of chemical fingerprints". In: *Algorithms for*

Molecular Biology 5.1 (2010), p. 9. ISSN: 1748-7188. DOI: 10.1186/1748-7188-5-9. URL: http://almob.biomedcentral.com/articles/10.1186/1748-7188-5-9.

[171] Martin Kroll and Simone Steinmetzer. "Automated Cryptanalysis of Bloom Filter Encryptions of Health Records". In: *arXiv preprint arXiv:1410.6739* (2014). arXiv: 1410.6739. URL: https://arxiv.org/abs/1410.6739.

[172] Martin Kroll and Simone Steinmetzer. "Who Is 1011011111. . .1110110010? Automated Cryptanalysis of Bloom Filter Encryptions of Databases with Several Personal Identifiers". In: *Biomedical Engineering Systems and Technologies*. Vol. 25. 2015, pp. 189–201. ISBN: 978-3-540-92218-6. DOI: 10.1007/978-3-540-92219-3. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-78049368213{\&}partnerID=tZOtx3y1.

[173] Mehmet Kuzu, Murat Kantarcioglu, Elizabeth Durham, and Bradley Malin. *A Constraint Satisfaction Cryptanalysis of Bloom Filters in Private Record Linkage*. Ed. by Simone Fischer-Hübner and Nicholas Hopper. Vol. 6794. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 226–245. ISBN: 978-3-642-22262-7. DOI: 10.1007/978-3-642-22263-4. URL: https://link.springer.com/book/10.1007{\%}2F978-3-642-22263-4.

[174] Mehmet Kuzu, Murat Kantarcioglu, Elizabeth Ashley Durham, Csaba Toth, and Bradley Malin. "A practical approach to achieve private medical record linkage in light of public resources". In: *Journal of the American Medical Informatics Association* 20.2 (2013), pp. 285–292. URL: http://jamia.oxfordjournals.org/content/20/2/285.short.

[175] Mehmet Kuzu, Murat Kantarcioglu, Ali Inan, Elisa Bertino, Elizabeth Durham, and Bradley Malin. "Efficient Privacy-Aware Record Integration." In: *Advances in database technology : proceedings. International Conference on Extending Database Technology* (2013), pp. 167–178. DOI: 10.1145/2452376.2452398. URL: http://dl.acm.org/citation.cfm?id=2452376.2452398.

[176] Shengjie Lai, Andrea Farnham, Nick W. Ruktanonchai, and Andrew J. Tatem. "Measuring mobility, disease connectivity and individual risk: a review of using mobile phone data and mHealth for travel medicine". In: *Journal of travel medicine* 26.3 (2019), taz019. ISSN: 17088305. DOI: 10.1093/jtm/taz019.

[177] Joseph T Lariscy. "Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox". In: *Journal of aging and health* 23.8 (2011), pp. 1263–1284.

[178] Kristin Lauter, Michael Naehrig, and Vinod Vaikuntanathan. "Can homomorphic encryption be practical?" In: *Proceedings of the 3rd ACM workshop on Cloud computing security workshop - CCSW '11*. New York, New York, USA: ACM Press, 2011, p. 113. ISBN: 9781450310048. DOI: 10.1145/2046660.2046682. URL: http://dl.acm.org/citation.cfm?doid=2046660.2046682.

[179] G Lawrence, Isa Dinh, and Lee Taylor. "The Centre for Health Record Linkage: A New Resource for Health Services Research and Evaluation". In: *Health Information Management Journal* 37.2 (2008), pp. 60–62.

[180] Joohee Lee, Dongwoo Duhyeong Kim, Dongwoo Duhyeong Kim, Yongsoo Song, Junbum Shin, and Jung Hee Cheon. "Instant Privacy-Preserving Biometric Authentication for Hamming Distance". In: *IACR Cryptology ePrint Archive* 2018 (2018), p. 1214.

[181] Yuen Ai Lee. "Medicineinsight: Scalable and Linkable General Practice Data Set". In: *Health Data Analytics. 2019 Presented at: HDA'19*. Sydney, Australia, 2019.

[182] L Li, J Li, and H Gao. "Rule-Based Method for Entity Resolution". In: *IEEE Transactions on Knowledge and Data Engineering* 27.1 (2015), pp. 250–263. ISSN: 1558-2191. DOI: 10.1109/TKDE.2014.2320713.

[183] Meng Liu, Priyadarsi Nanda, Xuyun Zhang, Chi Yang, Shui Yu, and Jianxin Li. "Asymmetric Commutative Encryption Scheme Based Efficient Solution to the Millionaires' Problem". In: *Proceedings - 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications and 12th IEEE International Conference on Big Data Science and Engineering, Trustcom/BigDataSE 2018*. Institute of Electrical and Electronics Engineers Inc., 2018, pp. 990–995. ISBN: 9781538643877. DOI: 10.1109/TrustCom/BigDataSE.2018.00139.

[184] Ray Lovett, Jodie Fisher, Fadwa Al-Yaman, Phyll Dance, and Hassan Vally. "A review of Australian health privacy regulation regarding the use and disclosure of identified data to conduct data linkage." en. In: *Australian and New Zealand journal of public health* 32.3 (2008), p. 4. ISSN: 1326-0200. DOI: 10.1111/j.1753-6405.2008.00230.x. URL: http://europepmc.org/abstract/MED/18578830http://www.ncbi.nlm.nih.gov/pubmed/18578830.

[185] Ronan A Lyons, David V Ford, Laurence Moore, and Sarah E Rodgers. "Use of data linkage to measure the population health effect of non-health-care interventions". In: *The Lancet* 383.9927 (2014), pp. 1517–1519.

[186] Vadim Lyubashevsky, Chris Peikert, and Oded Regev. "On ideal lattices and learning with errors over rings". In: *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer. 2010, pp. 1–23.

[187] Roger S Magnusson. "Data linkage, health research and privacy: regulating data flows in Australia's health information system". In: *Sydney L. Rev.* 24 (2002), p. 5.

[188] Sean Marston, Zhi Li, Subhajyoti Bandyopadhyay, Juheng Zhang, and Anand Ghalsasi. "Cloud computing — The business perspective". In: *Decision Support Systems* 51.1 (2011), pp. 176–189. ISSN: 01679236. DOI: 10.1016/j.dss.2010.12.006. URL: http://linkinghub.elsevier.com/retrieve/pii/S0167923610002393.

[189] Patricia J Martens. "Using the repository housed at the Manitoba centre for health policy: learning from the past, planning for the future". In: *Montreal, Quebec: Conference proceedings of the Statistics Canada Conference: Longitudinal Social and Health Surveys in an International Perspective*. 2006.

[190] Michael G Maxfield, Barbara Luntz Weiler, and Cathy Spatz Widom. "Comparing self-reports and official records of arrests". In: *Journal of Quantitative Criminology* 16.1 (2000), pp. 87–110.

[191] Andrew McCallum, Kamal Nigam, and Lyle H. Ungar. "Efficient clustering of high-dimensional data sets with application to reference matching". In: *Proceedings of the sixth*

ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00. New York, New York, USA: ACM Press, 2000, pp. 169–178. ISBN: 1581132336. DOI: 10.1145/347090.347123. URL: http://portal.acm.org/citation.cfm?doid=347090.347123.

[192] Scott A McDonald, Sharon J Hutchinson, Sheila M Bird, Peter R Mills, John Dillon, Mick Bloor, Chris Robertson, Martin Donaghy, Peter Hayes, and Lesley Graham. "A population-based record linkage study of mortality in hepatitis C-diagnosed persons with or without HIV coinfection in Scotland". In: *Statistical methods in medical research* 18.3 (2009), pp. 271–283.

[193] Kimberlyn M. McGrail, Kerina Jones, Ashley Akbari, Tellen D. Bennett, Andy Boyd, Fabrizio Carinci, Xinjie Cui, Spiros Denaxas, Nadine Dougall, David Ford, Russell Kirby, Hye Chung Kum, Rachael Moorin, Ros Moran, Christine M. O'Keefe, David Preen, Hude Quan, Claudia Sanmartin, Michael Schull, Mark Smith, Christine Williams, Tyler Williamson, Grant M.A. Wyper, and Milton Kotelchuck. "A position statement on population data science: The science of data about people". In: *International Journal of Population Data Science* 3.1 (2018), pp. 1–11. ISSN: 23994908. DOI: 10.23889/ijpds.v3i1.415.

[194] Doaa Medhat, Ahmed H. Yousef, and Cherif Salama. "Cost-aware load balancing for multilingual record linkage using MapReduce". In: *Ain Shams Engineering Journal* 11.2 (2019), pp. 419–433. ISSN: 20904479. DOI: 10.1016/j.asej.2019.08.009. URL: https://doi.org/10.1016/j.asej.2019.08.009.

[195] Peter Mell and Timothy Grance. *SP 800-145. The NIST Definition of Cloud Computing*. Tech. rep. 2011. DOI: http://doi.org/10.6028/NIST.SP.800-145. URL: http://faculty.winthrop.edu/domanm/csci411/Handouts/NIST.pdf.

[196] Demetrio Gomes Mestre, Carlos Eduardo Santos Pires, Dimas Cassimiro Nascimento, Andreza Raquel Monteiro de Queiroz, Veruska Borges Santos, and Tiago Brasileiro Araujo. "An efficient spark-based adaptive windowing for entity matching". In: *Journal of Systems and Software* 128 (2017), pp. 1–10. ISSN: 01641212. DOI: 10.1016/j.jss.2017.03.003. URL: http://dx.doi.org/10.1016/j.jss.2017.03.003.

[197] Rebecca J Mitchell, Cate M Cameron, and Rod McClure. "Quantifying the hospitalised morbidity and mortality attributable to traumatic injury using a population-based matched cohort in Australia". In: *BMJ open* 6.12 (2016), e013266.

[198] Rebecca J Mitchell, Cate M Cameron, Rod J McClure, and Ann M Williamson. "Data linkage capabilities in Australia: practical issues identified by a Population Health Research Network 'Proof of Concept project'". In: *Australian and New Zealand journal of public health* 39.4 (2015), pp. 319–325.

[199] Rachael Moorin. *Use of linked administrative data to evaluate continuity of primary care. PHRN Webinar*. 2019. URL: https://www.youtube.com/watch?v=SWhLy3V0Gnw.

[200] John Neter, E Scott Maynes, and R Ramanathan. "The effect of mismatching on the measurement of response errors". In: *Journal of the American Statistical Association* 60.312 (1965), pp. 1005–1027.

[201] *New centre to unlock the secrets of cheaper, quality healthcare*. 2014. URL: https://newsroom.unsw.edu.au/news/health/new-centre-unlock-secrets-cheaper-quality-healthcare (visited on 11/23/2020).

[202] New South Wales Government. *NSW Government Cloud Policy*. 2020. URL: https://www.digital.nsw.gov.au/policy/cloud-strategy-and-policy/cloud-policy.

[203] H B Newcombe and J M Kennedy. "Record linkage: making maximum use of the discriminating power of identifying information". In: *Commun. ACM* 5.11 (1962), pp. 563–566. DOI: doi:http://doi.acm.org/10.1145/368996.369026.

[204] HB Newcombe, James M Kennedy, SJ Axford, and AP James. "Automatic Linkage of Vital Records". In: *Science* (1959), pp. 954–959.

[205] Howard B Newcombe. *Handbook of record linkage : methods for health and statistical studies, administration, and business*. New York: Oxford University Press, 1988, p. 210. ISBN: 019261732X.

[206] Frank Niedermeyer, Simone Steinmetzer, Martin Kroll, and Rainer Schnell. "Cryptanalysis of basic Bloom Filters used for Privacy Preserving Record Linkage". In: *Journal of Privacy and Confidentiality* 6.2 (2014), p. 3.

[207] NSW Health. *Lumos*. 2019. URL: https://www.health.nsw.gov.au/LUMOS/Pages/default.aspx.

[208] Office for National Statistics. *Beyond 2011: Matching Anonymous Data*. Tech. rep. 2013. URL: https://webarchive.nationalarchives.gov.uk/20150912144244/https://ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/methods-and-policies-reports/beyond-2011-matching-anonymous-data.pdf.

[209] Office of the Australian Information Commissioner. *Data breach notification guide: A guide to handling personal information security breaches*. 2014.

[210] Office of the Australian Information Commissioner. *Privacy Fact Sheet 2: National Privacy Principles*. 2011.

[211] Christine M O'Keefe and Chris Connolly. "Regulation and perception concerning the use of health data for research in Australia". In: *Electronic Journal of Health Informatics* 6.2 (2011), p. 16.

[212] Toan C. Ong, Michael V. Mannino, Lisa M. Schilling, and Michael G. Kahn. "Improving record linkage performance in the presence of missing linkage data". In: *Journal of Biomedical Informatics* 52 (2014), pp. 43–54. ISSN: 15320464. DOI: 10.1016/j.jbi.2014.01.016. URL: https://www.sciencedirect.com/science/article/pii/S1532046414000197.

[213] J. Adam Oostema, Adrienne Nickles, and Mathew J. Reeves. "A Comparison of Probabilistic and Deterministic Match Strategies for Linking Prehospital and in-Hospital Stroke Registry Data". In: *Journal of Stroke and Cerebrovascular Diseases* 29.10 (2020), p. 105151. ISSN: 10523057. DOI: 10.1016/j.jstrokecerebrovasdis.2020.105151. URL: https://linkinghub.elsevier.com/retrieve/pii/S1052305720305693.

[214] Dermot O'Reilly, Orla Bateson, Gemma McGreevy, Chris Snoddy, and Tracy Power. "Administrative Data Research Northern Ireland (ADR NI)". In: *International Journal of Population Data Science* 4.2 (2020). ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i2.1148. URL: https://ijpds.org/article/view/1148.

[215] A. & Singh P. Pandey. *Full @ Www.Frontiersin.Org*. 2016. URL: http://www.frontiersin.org/language{_}sciences/10.3389/fpsyg.2011.00054/full.

[216] Alexandros Pantelopoulos and Nikolaos G Bourbakis. "A survey on wearable sensor-based systems for health monitoring and prognosis". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40.1 (2010), pp. 1–12.

[217] George Papadakis, Dimitrios Skoutas, Emmanouil Thanos, and Themis Palpanas. "Blocking and Filtering Techniques for Entity Resolution". In: *ACM Computing Surveys* 53.2 (2020), pp. 1–42. ISSN: 0360-0300. DOI: 10.1145/3377455.

[218] Graeme Philipson. *Cloud strategy transforms Victoria's largest agency*. 2017. URL: https://www.governmentnews.com.au/cloud-strategy-transforms-victorias-largest-agency/.

[219] Clicia Pinto, Robespierre Pita, George Barbosa, Bruno Araujo, Juracy Bertoldo, Samila Sena, Sandra Reis, Rosemeire Fiaccone, Leila Amorim, Maria Yuri Ichihara, Mauricio Barreto, Marcos Barreto, and Spiros Denaxas. "Probabilistic Integration of Large Brazilian Socioeconomic and Clinical Databases". In: *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2017, pp. 515–520. ISBN: 978-1-5386-1710-6. DOI: 10.1109/CBMS.2017.64. URL: http://ieeexplore.ieee.org/document/8104248/.

[220] Robespierre Pita, Clicia Pinto, Pedro Melo, Malu Silva, Marcos Barreto, and Davide Rasella. "A Spark-based workflow for probabilistic record linkage of healthcare data". In: *CEUR Workshop Proceedings*. Vol. 1330. 2015.

[221] *Population Health Research Network 2013 Independent Panel Review*. Tech. rep. April. 2014, pp. 1–37.

[222] Edward H Porter and William E Winkler. "Approximate String Comparison and its Effect on an Advanced Record Linkage System". In: *Advanced Record Linkage System. U.S. Bureau of the Census, Research Report*. 1997, pp. 190–199. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.46.7347.

[223] Conrad Pow, Karey Iron, James Boyd, Adrian Brown, Simon Thompson, Nelson Chong, and Charlotte Ma. "Privacy-Preserving Record Linkage: An international collaboration between Canada, Australia and Wales". In: *2016 International Population Data Linkage Conference* 1.1 (2016). DOI: http://dx.doi.org/10.23889/ijpds.v1i1.101. URL: http://www.ipdlnconference2016.org/Programme/Abstract/84.

[224] Productivity Commission. *Data Availability and Use*. Tech. rep. Canberra, 2017.

[226] Python Software Foundation. *Python language reference, version 2.7*. 2010.

[227] Catherine Quantin, Hocine Bouzelat, F A A Allaert, Anne-Marie Benhamiche, Jean Faivre, and Liliane Dusserre. "How to ensure data security of an epidemiological

follow-up: quality assessment of an anonymous record linkage procedure". In: *International journal of medical informatics* 49.1 (1998), pp. 117–122.

[228] Catherine Quantin, Hocine Bouzelat, and Liliane Dusserre. "A computerized record hash coding and linkage procedure to warrant epidemiological follow-up data security". In: *Studies in Health Technology and Informatics*. Vol. 43. IOS Press, 1997, pp. 339–342. ISBN: 9051993439. DOI: 10.3233/978-1-60750-887-8-339.

[229] Queensland Government. *ICT Strategy - Cloud computing*. 2020. URL: https://www.qgcio.qld.gov.au/ict-strategy/cloud-computing.

[230] Jennifer M. Radin, Nathan E. Wineinger, Eric J. Topol, and Steven R. Steinhubl. "Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study". In: *The Lancet Digital Health* 2.2 (2020), e85–e93. ISSN: 25897500. DOI: 10.1016/S2589-7500(19)30222-5. URL: http://dx.doi.org/10.1016/S2589-7500(19)30222-5.

[231] Thilina Ranbaduge, Peter Christen, and Dinusha Vatsalan. "Tree Based Scalable Indexing for Multi-Party Privacy-Preserving Record Linkage". In: *Australasian Data Mining Conference*. 2014.

[232] Thilina Ranbaduge, Dinusha Vatsalan, and Peter Christen. "MERLIN – A Tool for Multi-party Privacy-Preserving Record Linkage". In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015, pp. 1640–1643. ISBN: 978-1-4673-8493-3. DOI: 10.1109/ICDMW.2015.101. URL: http://ieeexplore.ieee.org/document/7395877/.

[233] S. M. Randall, A. M. Ferrante, J. H. Boyd, A. P. Brown, and J. B. Semmens. "Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581?" In: *Health Information Management Journal* (2016). ISSN: 1833-3583. DOI: 10.1177/1833358316647587. URL: http://him.sagepub.com/lookup/doi/10.1177/1833358316647587.

[234] Sean M Randall, James H Boyd, Anna M Ferrante, Jacqueline K Bauer, and James B Semmens. "Use of graph theory measures to identify errors in record linkage". In: *Computer Methods and Programs in Biomedicine* 115.2 (2014), pp. 55–63. ISSN: 01692607. DOI: 10.1016/j.cmpb.2014.03.008. URL: https://doi.org/10.1016/j.cmpb.2014.03.008.

[235] Sean M Randall, James H Boyd, Anna M Ferrante, Adrian P Brown, James B Semmens, and Best Link. "Grouping methods for ongoing record linkage". In: *Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference*. August. 2015.

[236] Sean M Randall, Adrian P Brown, Anna M Ferrante, James H Boyd, and James B Semmens. "Privacy preserving record linkage using homomorphic encryption". In: *Proceedings of the ACM-SIGKDD Population Informatics 2015 Conference*. August. 2015.

[237] Sean M Randall, Anna M Ferrante, James H Boyd, and James B Semmens. "The effect of data cleaning on record linkage quality". In: *BMC Medical Informatics and Decision Making* 13.1 (2013), p. 64. ISSN: 1472-6947. DOI: 10.1186/1472-6947-13-64. URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-13-64.

[238] Sean M. Randall, Anna M. Ferrante, James H. Boyd, James B. Semmens, Jacqueline K. Bauer, and James B. Semmens. "Privacy-preserving record linkage on large real world datasets". In: *Journal of Biomedical Informatics* 50 (2014), pp. 205–212. ISSN: 15320464. DOI: 10.1016/j.jbi.2013.12.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S1532046413001949.

[239] Leslie L Roos, Marni Brownell, Lisa Lix, Noralou P Roos, Randy Walld, Leonard MacWilliam, Roos LL, Brownell M, Lix L, Roos NP, Walld R, and Leonard MacWilliam. "From health research to social research: Privacy, methods, approaches". In: *Social Science and Medicine* 66.1 (2008), pp. 117–129.

[240] Leslie L Roos, Andre Wajda, and Faculty of Medicine. *Record Linkage Strategies: Part 1: Estimating Information and Evaluating Approaches*. Tech. rep. Winnipeg, 1990, p. 28.

[241] Diana Rosman, Carol Garfield, Stuart Fuller, Alexia Stoney, Todd Owen, and Geoff Gawthorne. "Measuring data and link quality in a dynamic multi-set linkage system". In: *Book measuring data and link quality in a dynamic multi-set linkage system* (2002), p. 4.

[242] Trish Ryan, Diane Gibson, and Bella Holmes. *A national minimum data set for home and community care*. Australian Institute of Health and Welfare, 1999.

[243] Rita Sallam, Donald Feinberg, Mark Beyer, W Roy Schulte, Alexander Linden, Joseph Unsworth, Svetlana Sicular, Nick Heudecker, Ehtisham Zaidi, Adam Ronthal, Erick Brethenoux, Pieter den Hamer, and Alys Woodward. *Top 10 Data and Analytics Technology Trends That Will Change Your Business*. Tech. rep. ID G00379563. Retrieved from Gartner database, 2019. URL: https://www.gartner.com/document/3906812.

[244] C Samuels. "Using the EM Algorithm to Estimate the Parameters of the Fellegi-Sunter Model for Data Linking". 2012.

[245] Kurt Schmidlin et al. "Privacy Preserving Probabilistic Record Linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality". In: *BMC Medical Research Methodology* 15.1 (2015), p. 46. ISSN: 1471-2288. DOI: 10.1186/s12874-015-0038-6. URL: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-015-0038-6.

[246] Irene Schmidtmann, Murat Sariyar, Andreas Borg, Aslihan Gerold-Ay, Oliver Heidinger, Hans-Werner Hense, Volker Krieg, and Gaël Paul Hammer. "Quality of record linkage in a highly automated cancer registry that relies on encrypted identity data." In: *GMS Medizinische Informatik, Biometrie und Epidemiologie* 12.1 (2016).

[247] Matthias Schneider, Christopher Gordon Radbone, Stacy Ann Vasquez, Miroslav Palfy, and Andrew Kristjan Stanley. "Population Data Centre Profile: SA NT DataLink (South Australia and Northern Territory)". In: *International Journal of Population Data Science* 4.2 (2019). ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i2.1136. URL: https://ijpds.org/article/view/1136.

[248] Rainer Schnell. "An efficient privacy-preserving record linkage technique for administrative data and censuses". In: *Statistical Journal of the IAOS* 30.3 (2014), pp. 263–270. ISSN: 18747655. DOI: 10.3233/SJI-140833.

[249] Rainer Schnell. "Increasing the Security of Bloom Filter Based Privacy Preserving Record Linkage". In: *The Farr Institute International Conference 2015: Data Intensive Health Research and Care*. St Andrews, Scotland, 2015.

[250] Rainer Schnell. "Privacy-preserving record linkage". In: *Methodological Developments in Data Linkage*. John Wiley & Sons, 2015. Chap. 9, pp. 201–225.

[251] Rainer Schnell, Tobias Bachteler, and Jörg Reiher. "A Novel Error-Tolerant Anonymous Linking Code". 2011.

[252] Rainer Schnell, Tobias Bachteler, and Jörg Reiher. "Privacy-preserving record linkage using Bloom filters". In: *BMC Medical Informatics and Decision Making* 9.1 (2009), p. 41. ISSN: 1472-6947. DOI: 10.1186/1472-6947-9-41. URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-9-41.

[253] Rainer Schnell and Christian Borgs. "Building a national perinatal data base without the use of unique personal identifiers". In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE. 2015, pp. 232–239.

[254] Rainer Schnell and Christian Borgs. "Randomized Response and Balanced Bloom Filters for Privacy Preserving Record Linkage". In: *IEEE International Conference on Data Mining (ICDM'16)*. 2016.

[255] Rainer Schnell and Christian Borgs. "Secure Privacy Preserving Record Linkage of Large Databases by Modified Bloom Filter Encodings". In: *International Population Data Linkage Conference (IPDLN 2016)*. Swansea, Wales: Swansea University, 2016.

[256] Rainer Schnell and Christian Borgs. "XOR-Folding for Bloom Filter-based Encryptions for Privacy-preserving Record Linkage". In: *Working Paper WP-GRLC-2016-03, German Record Linkage Center, Nuremberg* (2016).

[257] Rainer Schnell, Anke Richter, and Christian Borgs. "A Comparison of Statistical Linkage Keys with Bloom Filter-based Encryptions for Privacy-preserving Record Linkage using Real-world Mammography Data." In: *HEALTHINF*. 2017, pp. 276–283.

[258] Rainer Schnell and Dorothea Rukasz. *PPRL: Privacy Preserving Record Linkage*. 2019. URL: https://cran.r-project.org/package=PPRL.

[259] Michael J Schull, Mahmoud Azimaee, Marcel Marra, Rosario G Cartagena, Marian J Vermeulen, Minnie M Ho, and Astrid Guttmann. "ICES: Data, Discovery, Better Health". In: *International Journal of Population Data Science* 4.2 (2020). ISSN: 2399-4908. DOI: 10.23889/ijpds.v4i2.1135. URL: https://ijpds.org/article/view/1135.

[260] Ziad Sehili, Lars Kolb, Christian Borgs, Rainer Schnell, and Erhard Rahm. "Privacy Preserving Record Linkage with PPJoin". In: (2015).

[261] Gulzar H Shah, Kaveepan Lertwachara, and Anteneh Ayanso. "Record linkage in healthcare: Applications, opportunities, and challenges for public health". In: *International Journal of Healthcare Delivery Reform Initiatives (IJHDRI)* 2.3 (2010), pp. 29–47.

[262] James G. Shanahan and Liang Dai. "Large scale distributed data science using apache spark". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol. 2015-Augus. 2015. DOI: 10.1145/2783258.2789993.

[263] Caroline Shaw, June Atkinson, and Tony Blakely. "(Mis) classification of ethnicity on the New Zealand Cancer Registry: 1981-2004". In: *The New Zealand Medical Journal (Online)* 122.1294 (2009).

[264] Ashish Singh and Kakali Chatterjee. "Cloud security issues and challenges: A survey". In: *Journal of Network and Computer Applications* 79 (2017), pp. 88–115. ISSN: 1084-8045. DOI: 10.1016/J.JNCA.2016.11.027. URL: https://www.sciencedirect.com/science/article/pii/S1084804516302983.

[265] Jillian Smith. "The history and future of record linkage in the ONS longitudinal study". In: *Statistical Journal of the United Nations Economic Commission for Europe* 16.2, 3 (1999), pp. 197–205.

[266] Fiona Stanley, Rebecca Glauert, Anne McKenzie, and Melissa O'Donnell. "Can joined-up data lead to joined-up thinking? The Western Australian developmental pathways project." eng. In: *Healthcare policy = Politiques de sante* 6.Spec Issue (2011), pp. 63–73. ISSN: 1715-6572 (Print).

[267] Lee K. Taylor, Katie Irvine, Renee Iannotti, Taylor Harchak, and Kim Lim. "Optimal strategy for linkage of datasets containing a statistical linkage key and datasets with full personal identifiers". In: *BMC Medical Informatics and Decision Making* (2014). ISSN: 14726947. DOI: 10.1186/1472-6947-14-85.

[268] The Wall Street Journal. *NSA Officers Spy on Love Interests*. 2013. URL: http://blogs.wsj.com/washwire/2013/08/23/nsa-officers-sometimes-spy-on-love-interests/.

[269] Y Thibaudeau. "Fitting log-linear models when some dichotomous variables are unobservable". In: *Proceedings of the Section on statistical computing*. 1989, pp. 283–288.

[270] Csaba Toth, Elizabeth Durham, Murat Kantarcioglu, Yuan Xue, and Bradley Malin. "SOEMPI : A Secure Open Enterprise Master Patient Index Software Toolkit for Private Record Linkage". In: *AMIA Annual Symposium Proceedings*. 2014, pp. 1105–1114.

[271] John J Trinckes Jr. *The definitive guide to complying with the HIPAA/HITECH privacy and security rules*. CRC Press, 2012.

[272] B Trutwein, D Holman, and D Rosman. "Health Data Linkage Conserves Privacy in a Research-Rich Environment". In: *Annals of Epidemiology* 16 (2006).

[273] Ross E G Upshur, Benoit Morin, and Vivek Goel. "The privacy paradox: laying Orwell's ghost to rest". In: *Canadian Medical Association Journal* 165.3 (2001), pp. 307–309.

[274] Jaideep Vaidya and Chris Clifton. "Privacy preserving association rule mining in vertically partitioned data". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002, pp. 639–644.

[275] Jaideep Vaidya and Chris Clifton. "Secure set intersection cardinality with application to association rule mining". In: *Journal of Computer Security* 13.4 (2005), pp. 593–622.

[276] Vanderbilt University. *SOEMPI: Secure Open Enterprise Master Patient Index*. 2014. URL: https://hiplab.mc.vanderbilt.edu/projects/soempi/ (visited on 08/14/2020).

[277] Tatjana Vasiljeva, Sabina Shaikhulina, and Karlis Kreslins. "Cloud Computing: Business Perspectives, Benefits and Challenges for Small and Medium Enterprises (Case

of Latvia)". In: *Procedia Engineering* 178 (2017), pp. 443–451. ISSN: 18777058. DOI: 10.1016/j.proeng.2017.01.087. URL: http://dx.doi.org/10.1016/j.proeng.2017.01.087.

[278] Dinusha Vatsalan and Peter Christen. "Scalable Privacy-Preserving Record Linkage for Multiple Databases". In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14* (2014), pp. 1795–1798. DOI: 10.1145/2661829.2661875. URL: http://dl.acm.org/citation.cfm?doid=2661829.2661875.

[279] Dinusha Vatsalan, Peter Christen, Christine O'Keefe, and Vassilios Verykios. "An Evaluation Framework for Privacy-Preserving Record Linkage". In: *Journal of Privacy and Confidentiality* 6.1 (2014). URL: http://repository.cmu.edu/jpc/vol6/iss1/3.

[280] Dinusha Vatsalan, Peter Christen, Vassilios Verykios, and Others. "An efficient two-party protocol for approximate matching in private record linkage". In: *Proceedings of the Ninth Australasian Data Mining Conference - Volume 121* (2011), pp. 125–136.

[281] Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios. "A taxonomy of privacy-preserving record linkage techniques". In: *Information Systems* 38.6 (2013), pp. 946–969. ISSN: 03064379. DOI: 10.1016/j.is.2012.11.005. URL: https://doi.org/10.1016/j.is.2012.11.005.

[282] Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios. "Efficient Two-Party Private Blocking based on Sorted Nearest Neighborhood Clustering". In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13* (2013), pp. 1949–1958. DOI: 10.1145/2505515.2505757. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84889565216{\&}partnerID=tZOtx3y1.

[283] Dinusha Vatsalan, Ziad Sehili, Peter Christen, and Erhard Rahm. "Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges". In: *Handbook of Big Data Technologies*. Cham: Springer International Publishing, 2017, pp. 851–895. DOI: 10.1007/978-3-319-49340-4_25. URL: http://link.springer.com/10.1007/978-3-319-49340-4{_}25.

[284] Anders Vind Ebbesen. "The Creation of the Central Person Registry in Denmark". In: *History of Nordic Computing 4*. Vol. 447. 2015, pp. 263–272. ISBN: 978-3-319-17144-9. DOI: 10.1007/978-3-319-17145-6. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84942876592{\&}partnerID=tZOtx3y1.

[285] Jeffrey Voas. "Cloud Computing". In: *IT Professional* April (2013), pp. 12–14.

[286] WA Treasury. *Case study: Social Investment Data Resource*. 2020. URL: https://www.wa.gov.au/organisation/department-of-treasury/case-study-social-investment-data-resource (visited on 11/25/2020).

[287] Andre Wajda and Leslie L. Roos. "Simplifying record linkage: Software and strategy". In: *Computers in Biology and Medicine* 17.4 (1987), pp. 239–248. ISSN: 00104825. DOI: 10.1016/0010-4825(87)90010-2. URL: https://linkinghub.elsevier.com/retrieve/pii/0010482587900102.

[288] Lance A Waller and Carol A Gotway. *Applied spatial statistics for public health data*. Vol. 368. John Wiley & Sons, 2004. ISBN: 0471662674.

[289] Zhiyu Wan, Yevgeniy Vorobeychik, Weiyi Xia, Ellen Wright Clayton, Murat Kantarcioglu, Ranjit Ganta, Raymond Heatherly, and Bradley A Malin. "A game theoretic framework for analyzing re-identification risk". In: *PloS one* 10.3 (2015), e0120592.

[290] S. C. Weber, H. Lowe, A. Das, and T. Ferris. "A simple heuristic for blindfolded record linkage". In: *Journal of the American Medical Informatics Association* 19.e1 (2012), e157–e161. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2011-000329. URL: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2011-000329.

[291] William E Winkler. *Frequency Based Matching in Fellegi-Sunter Model of Record Linkage*. Ed. by U S Bureau of the Census. Washington DC, 2000.

[292] William E Winkler. "Matching and Record Linkage". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 6.5 (2005). Ed. by U S Bureau of the Census, pp. 313–325. ISSN: 19390068. DOI: 10.1002/wics.1317.

[293] William E Winkler. "Preprocessing of lists and string comparison". In: *W. Alvey and B* 985 (1985), pp. 181–187.

[294] William E. Winkler. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.* en. Tech. rep. Paper presented at the Annual ASA Meeting in Anaheim. CA. Washington, DC: Statistical Research Division, U.S. Bureau of the Census, 1990. URL: http://eric.ed.gov/?id=ED325505.

[295] William E Winkler. *Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage*. Tech. rep. Washington DC: Bureau of the Census - Statistical Research Division, 2000, p. 12.

[296] William E Winkler and Yves Thibaudeau. "An application of the Fellegi-Sunter model of record linkage to the 1990 US decennial census". In: *Research Report* (1990), pp. 1–22. DOI: 10.1.1.39.2433. URL: https://www.census.gov/srd/papers/pdf/rr91-9.pdf.

[297] Mohamed Yakout, Mikhail J. Atallah, and Ahmed Elmagarmid. "Efficient private record linkage". In: *Proceedings - International Conference on Data Engineering*. 2009, pp. 1283–1286. ISBN: 9780769535456. DOI: 10.1109/ICDE.2009.221.

[298] Wei Yan, Yuan Xue, and Bradley Malin. "Scalable load balancing for mapreduce-based record linkage". In: *2013 IEEE 32nd International Performance Computing and Communications Conference (IPCCC)*. IEEE, 2013, pp. 1–10. ISBN: 978-1-4799-3214-6. DOI: 10.1109/PCCC.2013.6742785. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84897789149{\&}partnerID=tZOtx3y1.

[299] WE Yancey. "Improving EM algorithm estimates for record linkage parameters". In: *Proceedings of the Section on Survey Research ...* (2002), pp. 3835–3840. URL: https://www.amstat.org/sections/srms/proceedings/y2002/Files/JSM2002-000581.pdf.

[300] Andrew Chi-Chih Yao. "How to generate and exchange secrets". In: *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*. IEEE, 1986, pp. 162–167. ISBN: 0-8186-0740-8. DOI: 10.1109/SFCS.1986.25. URL: http://ieeexplore.ieee.org/document/4568207/.

[301] Masaya Yasuda, Takeshi Shimoyama, Jun Kogure, Kazuhiro Yokoyama, and Takeshi Koshiba. *Practical packing method in somewhat homomorphic encryption*. Ed. by Joaquin Garcia-Alfaro, Georgios Lioudakis, Nora Cuppens-Boulahia, Simon Foley, and William M. Fitzgerald. Vol. 8247. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 148–161. ISBN: 978-3-642-54567-2. DOI: 10.1007/978-3-642-54568-9. URL: http://link.springer.com/10.1007/978-3-642-54568-9.

[302] David S Zingmond, Susan L Ettner, and Honghu Liu. "Linking hospital discharge and death records - accuracy and sources of bias". In: *Journal of Clinical Epidemiology* 57.1 (2004), pp. 21–29. DOI: doi:10.1016/S0895-4356(03)00250-6.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.