Journal of Biomedical Informatics 50 (2014) 205–212

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Privacy-preserving record linkage on large real world datasets

Sean M. Randall *, Anna M. Ferrante, James H. Boyd, Jacqueline K. Bauer, James B. Semmens

Centre for Population Health Research, Faculty of Health Sciences, Curtin University, Bentley 6102, WA, Australia

ARTICLE INFO

Article history:
Received 26 June 2013
Accepted 4 December 2013
Available online 9 December 2013

Keywords:
Record linkage
Privacy preserving record linkage
Data integration
Bloom filters
Privacy preserving protocols
Population based research

ABSTRACT

Record linkage typically involves the use of dedicated linkage units who are supplied with personally identifying information to determine individuals from within and across datasets. The personally identifying information supplied to linkage units is separated from clinical information prior to release by data custodians. While this substantially reduces the risk of disclosure of sensitive information, some residual risks still exist and remain a concern for some custodians. In this paper we trial a method of record linkage which reduces privacy risk still further on large real world administrative data. The method uses encrypted personal identifying information (bloom filters) in a probability-based linkage framework. The privacy preserving linkage method was tested on ten years of New South Wales (NSW) and Western Australian (WA) hospital admissions data, comprising in total over 26 million records. No difference in linkage quality was found when the results were compared to traditional probabilistic methods using full unencrypted personal identifiers. This presents as a possible means of reducing privacy risks related to record linkage in population level research studies. It is hoped that through adaptations of this method or similar privacy preserving methods, risks related to information disclosure can be reduced so that the benefits of linked research taking place can be fully realised.

1. Introduction

1.1. Administrative data as resource

Administrative health records, containing information on an individual's health and the health services they have received, cover a large proportion of the population and are generally considered to be highly sensitive data. They are used not only for managing an individual health event, but have important uses in informing research, planning and decision making [1]. Current Australian laws provide a number of safeguards to personal privacy including the requirement that the public benefit in using health information for research outweighs with the privacy risks of doing so for the individual [2].

1.2. Record linkage of health information

The process of record linkage is often used to enable researchers to answer questions which require a picture of an individual's health over time. Record linkage is used to identify administrative records belonging to the same person from multiple datasets. In the absence of a unique person identifier, this task is typically carried out using personally identifying information such as name, date of birth and address. As these identifiers can change and/or include errors within or between datasets, probabilistic statistical methods are typically used to ensure high quality links [3]. This linkage process allows researchers to answer questions about the health of individuals over time rather than solely about discrete health events. Research using linked data has resulted in changes to health services delivery and policy [4]. Large scale investment in record linkage infrastructure has occurred in England [5], Scotland [6], Wales [7], Canada [8] and Australia [9] over the last thirty years. Each of these centres has developed linkage expertise which has enabled important research at a population level.

1.3. Record linkage processes and privacy protection

The linkage of different administrative collections across portfolios usually requires the transfer of data to a trusted party or 'linkage unit', which may or may not be external to the data custodians/owners. Various processes and protocols have been developed to protect the privacy of individual and to maintain the security of data.

1.3.1. Separation principle

One method used in many Australian linkage units to reduce privacy risks is to separate data [10]. Under this model, the data is split into personally identifying data (containing information such as name, address and date of birth) and content data (clinical

* Corresponding author.
E-mail addresses: Sean.randall@curtin.edu.au (S.M. Randall), A.Ferrante@curtin.edu.au (A.M. Ferrante), J.Boyd@curtin.edu.au (J.H. Boyd), Jacqui.Bauer@curtin.edu.au (J.K. Bauer), James.Semmens@curtin.edu.au (J.B. Semmens).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.12.003

or service information used for research). The personal identifiers are released to a linkage unit, whose sole role is to determine which records belong to a single person (the release of name-identifying information for research is typically permitted in Australia through exemptions in privacy laws). This is carried out through probabilistic linkage using personal identifiers, typically with a large manual review component. The linkage unit then sends this information back to the custodian, who uses it to supply clinical information to the researcher (see Figure 1).

This method is used by linkage units in WA and NSW to conduct state-based linkages, and has been adopted by the CDL as its best practice national linkage model [9].

1.3.2. Information governance

In addition to the separation principle, linkage units have adopted strong policies and procedures applying to the obtaining, handling, using and disclosing of personally identifying information. This includes an effort to ensure that staff understand their role and responsibilities, that information assets are protected, that policies exist surrounding breaches and disclosure and that information systems place a high priority on security in their design. These policies and procedures have been adopted and developed with input from data custodians.

1.4. Privacy preserving linkage techniques

By separating clinical data from personal identifiers during the linkage process, the risk of revealing sensitive information about individuals is dramatically reduced. Staff conducting the linkage have access only to identifying information, while researchers see only the clinical information relevant for their research questions. Appropriate information governance within linkage units further reduces the risk of information leaks, whether accidentally or maliciously by operators, or as a result of poor business processes.

Nevertheless, some residual risk to privacy remains and, for some data custodians, this is sufficient to prevent the release of personally identifying information to record linkage units. Ideally, such data custodians seek a zero-risk method of providing accurate linked research data without the need to disclose any identifying information to linkage units.

Various techniques known as privacy preserving linkage have been developed to provide lower risk solutions for record linkage. These methods engage in record linkage on encrypted information, and do not require third parties to see personal identifiers. These techniques each differ in their methods, maturity, practicality and suitability for large scale linkages (particularly of low quality data).

Privacy preserving techniques can be classified into two general categories – those that utilise a third party for performing the linkage (three party protocols) and those that do not (two party protocols). Two-party protocols often require a greater amount of necessary communication and computation [11] to compare records, but can be considered more secure as they do not rely on the existence of a trusted third party [12].

In terms of security, privacy preserving techniques generally adopt the same threat model, but differ in the particular privacy techniques used. Nearly all privacy preserving protocols adopt an ‘honest-but-curious’ threat model [12], whereby parties are expected to try to carry out the protocol correctly, but will also try and find out as much information as they can from any data they receive.

Perhaps the most important criteria in differentiating privacy preserving protocols are around performance features such as linkage quality, scalability and robustness. Privacy preserving protocols range in terms of the comparison techniques applied, from those carrying out an exact match on entire records, to protocols employing string similarity measures on individual fields. Those protocols utilising more fine-grained techniques in determining similarity will typically give higher linkage quality.

Several privacy preserving protocols are being regularly used for routine record linkage. The Australian Institute of Health and Welfare uses the 2nd, 3rd and 5th letters of surname, the 2nd and 3rd letters of forename, the full date of birth and the persons sex to create a ‘statistical linkage key’ (SLK) which is used to match records [13]. The SLK has been used successfully for a large number of linkages. The Swiss Anonymous Linkage Code [14] creates an identifier from the phonetic codes of first and last name, along with full date of birth and sex. A similar method has been used to conduct linkage in France [15]. Grhanite [16] also uses privacy preserving protocols; like some other systems, it applies a number of pre-processing steps, including phonetic encoding and nickname resolution, before creating their identifier. The process uses these pre-processing steps and fuzzy matching algorithms to produce linkage results that are probabilistic in nature.

In this paper we adopted the bloom filter method for privacy preserving record linkage, developed by Schnell et al. [11]. There were several reasons why we chose this method over other privacy preserving protocols. Firstly the bloom filter approach differs from most other privacy preserving linkage methods in that it is able to measure the similarity between two fields (for instance, between two names) – a method often used in probabilistic record linkage to ensure high quality. Evaluations of privacy preserving string comparison using bloom filters have demonstrated very high quality [11,17], including quality improvements over the SLK and the Swiss anonymous linkage code [18]. Current evaluations have focussed on small data samples, but the method appears adaptable for large-scale record linkage. The method appears robust and well-developed, with a number of papers investigating its security [19] and proposing additions to its method [18,20].

The use of bloom filters was evaluated to determine its suitability for conducting large scale privacy preserving record linkage. Two datasets, comprising in total over 26 million records, were linked using this method, with results compared to the linkage of unencrypted data. A probabilistic linkage framework was adopted to allow large-scale linkage to occur.

2. Method

2.1. Application of bloom filters

To use bloom filters for encrypted record linkage, the personal identifiers need to be encrypted by data custodians. As this process is technically complicated, data custodians would need to be supplied with software that would enable them to encrypt the records. The data custodians involved in the project would agree on a password or pass phrase used to encrypt the data, which would not be shared with the linkage unit. The encrypted data can then be passed to the linkage unit, who can use it to determine which records belong to the same person (see Figure 2).

2.1.1. Creating and comparing bloom filters

An outline of the encryption process presented by Schnell et al. [11] is shown in Fig. 3 along with the method for comparing two encrypted variables which is shown in Fig. 4. Each value (for instance the given name ‘SEAN’ on one record) is encrypted separately.

A bloom filter begins as an array of a set length, with all array elements set to zero. Firstly, bigrams (overlapping sets of two letters) of the matching variables are created. Padding has been used to give the first and last letters their own bigrams – for instance,

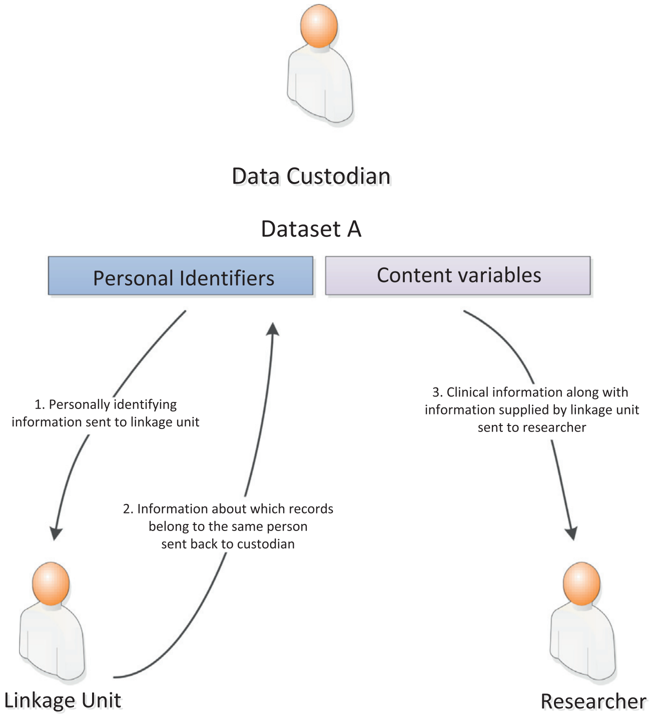

Data Custodian
Dataset A
Personal Identifiers
Content variables
1. Personally identifying information sent to linkage unit
2. Information about which records belong to the same person sent back to custodian
3. Clinical information along with information supplied by linkage unit sent to researcher
Linkage Unit
Researcher

Fig. 1. The separation principle: Personally identifying information (name, address, date of birth) is sent only to the linkage unit, while clinical information is sent only to the researcher.

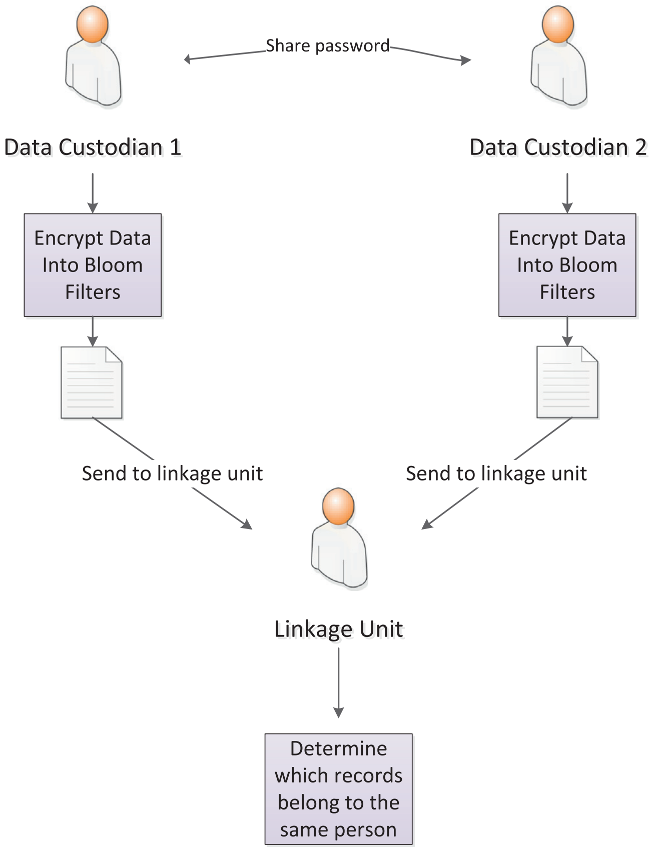

Share password
Data Custodian 1
Data Custodian 2
Encrypt Data Into Bloom Filters
Encrypt Data Into Bloom Filters
Send to linkage unit
Send to linkage unit
Linkage Unit
Determine which records belong to the same person

Fig. 2. Privacy preserving linkage using bloom filters: personal identifiers are first encrypted by the data custodian before being sent to the linkage unit.

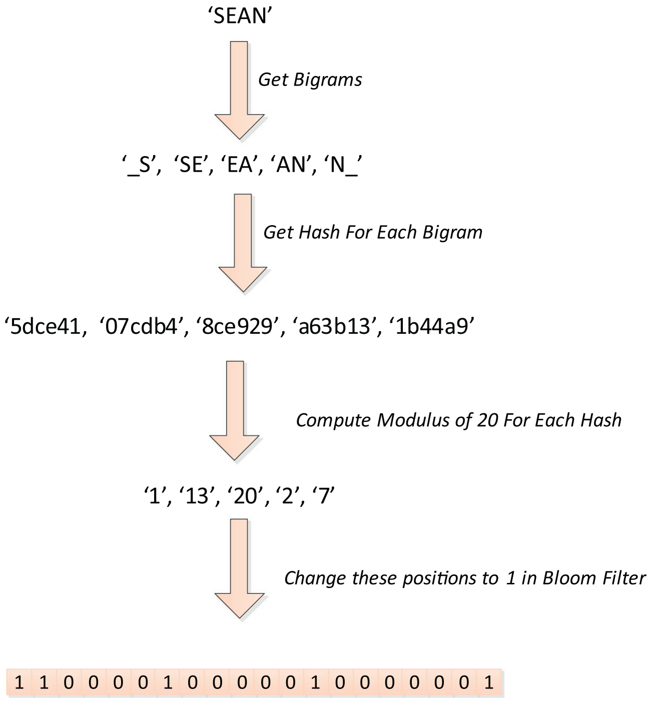

'SEAN'
Get Bigrams
'_S', 'SE', 'EA', 'AN', 'N_'
Get Hash For Each Bigram
'5dce41, '07cdb4', '8ce929', 'a63b13', '1b44a9'
Compute Modulus of 20 For Each Hash
'1', '13', '20', '2', '7'
Change these positions to 1 in Bloom Filter
1 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1

Fig. 3. Example of encoding the name 'Sean' into a Bloom Filter of length 20.

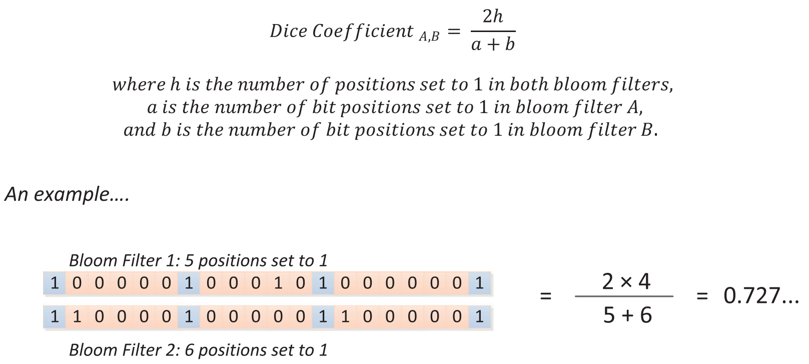
'SEAN' first becomes '_SEAN_', before being converted into the bigrams '_S', 'SE', 'EA', 'AN' 'N_'.

Each of the bigrams is passed through a hash function. The 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 bigram will always produce the same hash value). The hash functions is also one-way, meaning it is not possible to determine the encoded bigram from the given hash value.

The modulus of these hashes is then computed with respect to the length of the bloom filter. This results in each bigram having a number which corresponds to a position in the bloom filter. These positions in the bloom filter are then changed to 1. 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 reduce the effects of false positives (which occur when two hash values map to the same position in the bloom filter).

Bloom filters are a useful tool for determining set membership efficiently. Bloom filters allow us to quickly determine whether a bigram is not in the encoded bloom filter – we simply hash the bigram and check whether the positions are set to 1. If they are set to 0 we can be certain the bigram is not contained in the bloom filter. If the positions are set to 1, then the bigram is possibly contained in the bloom filter – the other possibility is a false positive (a different bigram/combination of bigrams also resulting in the same positions being set). The probability of a false positive depends on the length of the bloom filter and the number of other elements already contained within the bloom filter.

Two bloom filters can be compared to each other using the dice coefficient. This is calculated as twice the number of positions in which both bloom filters have a value of one, divided by the number of positions set to 1 in total (see Fig. 4). The dice coefficient results in a score between 0 and 1, where a higher score reflects greater similarity.



$$Dice\ Coefficient_{A,B} = \frac{2h}{a+b}$$

where h is the number of positions set to 1 in both bloom filters,
a is the number of bit positions set to 1 in bloom filter A,
and b is the number of bit positions set to 1 in bloom filter B.

Fig. 4. Example of calculating string similarity by comparing two bloom filters.

Further detailed technical information on how to implement bloom filters can be found in the original paper by Schnell et al. [11].

2.2. Evaluation of bloom filter methodology

2.2.1. Datasets

Two individual datasets were used in the evaluation; ten years of West Australian Hospital Admissions data (approximately 7 million records) along with ten years of the New South Wales Admitted Patient Data Collection (approximately 20 million records) were used in the evaluation. This data was made available as part of the PHRN Proof of Concept 1 project [9]. Each dataset had been previously internally linked (deduplicated) to a very high quality (by the Western Australia linkage unit (WA-DLB [21]) and the New South Wales linkage unit (CHeReL [22]), respectively) using probabilistic linkage methods along with rigorous manual reviews of created links, and a quality assurance program to analyse and review likely errors. These linked datasets have been use in a large number of research projects and published research articles which further validate the quality of the matching. A summary description of the datasets is provided in Table 1.

2.2.2. Evaluation strategy

For each dataset an internal linkage was conducted using this privacy preserving methodology. This linkage strategy deduplicated each file, identifying all the records belonging to an individual within each datasets. In addition, an internal/deduplicating linkage of each dataset was performed using a probabilistic matching strategy [9], with the full unencrypted personal identifiers. In all cases, the results of these linkages were compared with the results achieved previously by the West Australian and New South Wales linkage units, which were supplied to the CDL and used as a gold standard. A variation to Schnell's bloom filter methodology using trigrams instead of bigrams was also tested.

A direct comparison between Schnell's method and the one extrapolated in this paper would be desirable but is essentially not possible. Schnell does not present a method to link records together (although it is clear he sees this as the main use for his string comparison methodology) but rather a method to compare alphabetic strings. Additional information regarding how to combine multiple comparison scores together, how to handle missing values, or how to compare non-string variables is required to compare records. This paper essentially adopts Schnell's method for comparing strings, placing it within a probabilistic linkage framework.

2.2.3. Linkage strategy

For the unencrypted linkage, a probabilistic linkage approach [3,23] was used. The linkage strategy was based on a published linkage strategy used to evaluate matching quality across a number of linkage products [24]. This strategy used two blocks (Soundex of surname with first initial, and date of birth). All possible comparison variables were compared in each block. String similarity measures were used for all alphabetic variables (names, address and suburb) with exact matches being carried out on all other variables. Day, month and year of birth were all compared separately. Correct agreement and disagreement weights were calculated for each variable and used in linkage. The linkage quality at various threshold settings was measured, with the highest result reported.

The encrypted linkage also followed this strategy, with the same blocking fields, comparisons and agreement and disagreement weights used. During the creation of the encrypted file, separate fields for Soundex and first initial were created to allow blocking on these fields during encrypted linkage. The bloom filter string similarity comparison was used for names, address and suburb fields, while an exact comparator was used for all other fields.

2.2.4. Creation of encrypted dataset

Bloom filters were created and compared based on the implementation described by Schnell et al. [11] with some modifications.

Within each dataset, bloom filters were created for individual fields. The bloom filters created were smaller in size than those originally used by Schnell (from a length of 1000 to a length of 100). This dramatically reduced file sizes. Fields were split into bigrams in line with the method outline Schnell et al. [11]. Padding was used for each field. The number of hash functions for each bigram used was 3; this kept the same ratio of hash functions to bloom filter length as described by Schnell et al. [11]. The dice coefficient was used to compare bloom filters. Work by Durham et al. [17] had shown the use of bigrams and the dice coefficient provided higher accuracy than other string similarity measures. Preliminary testing was carried out experimentally comparing Schnell's method for creating and comparing bloom filters (which used 1000 bit bloom filters with 30 hash functions per qgram) to our own (which used 100 bit bloom filters with 3 hash functions per qgram). This preliminary testing used the WA dataset described above. The dataset was encrypted using both methods, and linked using the same linkage strategy, with the results compared. No difference in linkage quality was found.

Blocking variables were used in the linkage strategy, dramatically reducing the number of comparisons and allowing large scale record linkage to occur (see Table 2). These were implemented as simple hashes of the original values. Only variables agreeing on a

Table 1
Quality of datasets used in evaluation.

WA morbidity 6,772,949 records	Percentage of missing values	Average string length	NSW morbidity 19,874,083 records	Percentage of missing values (%)	Average string length
Given name	<0.1%	6	Given name	31.9	4
Surname	<0.1%	6	Surname	31.8	4
Sex	0	1	Sex	<0.1	1
Date of birth	0	8	Date of birth	<0.1	8
Address	0	18	Address	7.5	15
Suburb	<0.1%	9	Suburb	<1.0	9
Postcode	0	4	Postcode	<1.0	4
State	0	1	State	<0.1	1
Percentage of records with a missing value					
WA morbidity	<0.1%		NSW morbidity	33.0	

set of blocking variables were compared further. Missing values did not have bloom filters computed, but were left blank. This allowed the linkage program to recognise them as missing values and treat them appropriately.

The bloom filter method for approximate string similarity allows for several variations. Both bigrams and trigrams have been used widely for (unencrypted) approximate string matching [25] with trigrams recognised as performing well in privacy preserving contexts [26]. Trigrams are more sensitive to differences between strings than bigram methods [25]. The bloom filter method can be easily adapted to use trigrams instead of bigrams, with no apparent efficiency trade-offs. As well as utilising the bigram bloom filter method used by Schnell and others [11,17], we also tested the use of bloom filters using trigrams, using the same datasets.

An alternate edit distance measure for bloom filter comparisons has been developed, based on the Levenshtein distance function [27]. This edit distance measure has several differences to the dice coefficient measure, both in terms of its calculation, as well as its security properties. The edit distance measure requires a specific method of creating the bloom filters, joining individual letters of the string in question with their position in the string and hashing these (i.e. ‘1S’, ‘2E’, ‘3A’, ‘4N’ for ‘SEAN’). The edit distance measure also requires additional information to be known to the party carrying out the similarity calculation, such as the number of characters in the string. More importantly, the edit distance function requires the party carrying out the similarity calculation to be comparing a single bloom filter to a specific known word – it cannot be used to compare two bloom filters containing unknown strings. It also requires any password used to encrypt the data to be known to the evaluating party. This has significant privacy implications.

The edit distance measure involves determining whether specific letter/number combinations are contained within the bloom filter. This involves calculating hashes, a relatively slow operation. Using the smaller WA data, a single comparison of two records will require approximately 950 hashes to be calculated. Given there are over 3.5 billion comparisons to be performed, this increases the number of required hashes to 3.3 trillion. Using the standard SHA-1 hash function used by Schnell et al. [11] we can calculate approximately 700,000 hashes per second on our current hardware. This equates to a run time of over 50 days to complete a linkage of our smaller file. The edit distance measure, as currently formulated, does not appear feasible for large scale record linkage and cannot be used for third party privacy preserving linkage. For these reasons, it was not experimentally evaluated in this study.

2.2.5. Measuring linkage quality

Linkage quality was evaluated using pairwise precision, recall and *f*-measure. These measures have been previously used in the record linkage literature [24]. The *f*-measure of a linkage is the harmonic mean between precision and recall. This provides a single figure with which linkage quality can be compared. Final thresholds were set to levels which maximised the *f*-measure.

3. Results

Linkage quality (precision, recall and *f*-measure) for encrypted linkage using bigrams, and unencrypted linkages are presented in Fig. 5. The threshold here refers to the lowest acceptable linkage score used to determine results; probabilistic linkage gives each record pair combination a score based on their similarity, and it is up to the operator to determine the appropriate threshold. There appears to be very little difference in linkage quality across the threshold weight range except in the very high threshold values in the WA data, where encrypted linkage outperformed unencrypted linkage. In all graphs, there appears to be a reasonably wide range (from threshold value 14 to 18) where matching achieved near optimum linkage quality. This should make it easier for operators to extract high quality from their linkage. It is notable that, for all the linkages, the optimal threshold settings did not appear to vary highly between encrypted and unencrypted data, potentially hinting at methods of determining threshold settings in privacy preserving record linkage.

Quality results using the optimal matching thresholds for each linkage are presented in Table 3. The linkage quality was very high for all linkages. There was very little difference is linkage quality between the bigram and trigram encrypted linkages. Almost no difference in quality was found between the encrypted linkage using bloom filters and the unencrypted linkage using full personal identifiers.

Irrespective of linkage type, results for NSW data were slightly lower than those for WA. This was due to lower data quality – the NSW file was missing all name information for one third of records (see Table 1).

4. Discussion

This is the first time that privacy preserving linkage using the bloom filter method for approximate string comparison has been applied to large, population-level data collections. The results show that in terms of linkage quality, probabilistic linkage using encrypted fields is equally as effective as probabilistic linkage using unencrypted personal identifiers. In our experimental setting, the results demonstrate that it is possible to link large volumes of data and achieve high quality linkage without the need to disclose fully identifying personal information. However, there are some limited issues which may need to be overcome.

One drawback of using privacy preserving linkage is the difficulty of checking the overall quality of the linkage. The typical

Table 2
Efficiency of linkage with blocking.

	WA dataset		NSW data	
	Blocking[a]	No blocking	Blocking[a]	No blocking
Number of comparisons	3,506,013,239	22,936,415,691,826	16,226,633,871	197,489,577,608,403
Speed of record pair comparisons[b] (comparisons per second)	120,000	120,000	120,000	120,000
Approximate time required for linkage	8 h	6 years	37.5 h	52 years

[a] Using the specific blocking variables of soundex of surname with first initial, as well as date of birth.

[b] A single record pair comparison involved the comparison of five bloom filters using the dice coefficient as well as with five exact comparisons, along with the pooling of these results into a final score.

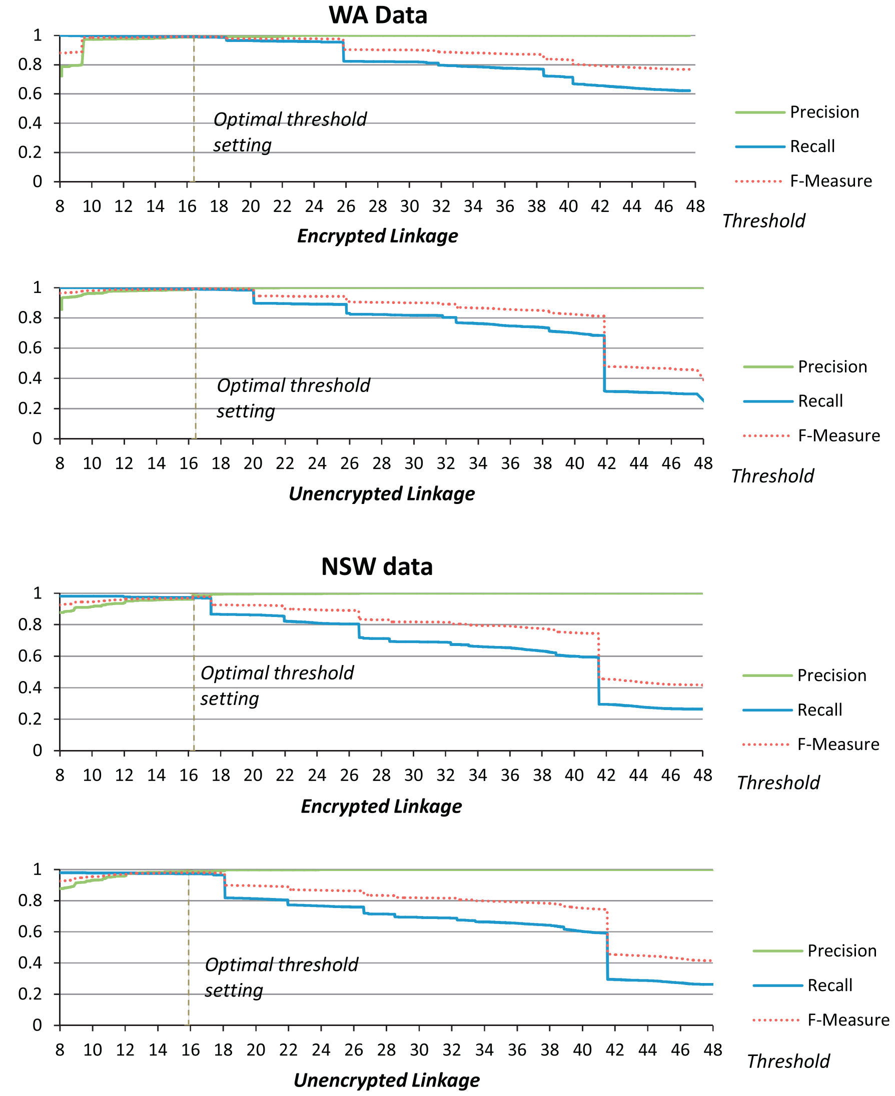


Fig. 5. Precision, recall and *f*-measure (linkage quality) for all possible threshold settings for each linkage undertaken. For almost any threshold setting, there is little difference in linkage quality between encrypted and unencrypted linkage.

quality assessment mechanism for linkage is through manual inspection of individual pairs of records which have been brought together. As all records in the privacy preserving linkage are encrypted, this is no longer possible. If a linkage was conducted that was of poor quality, this poor quality linked clinical data would be passed to the researcher. Depending on the errors, these may be immediately noticeable by the researcher, or may be completely undetectable, leading to erroneous research findings.

Manual inspection of individual pairs is also used to determine optimal threshold settings. There is currently no obvious way to determine optimal thresholds with probabilistic encrypted linkage. Additional work in this area will need to be undertaken in order to make this method useful in practice. Based on Fig. 5, the optimal threshold for an encrypted linkage appears to be very similar to that of unencrypted linkage, which both have a reasonably wide tolerance (a threshold setting 1 or 2 points either side will give

Table 3
Linkage quality of internal links of two datasets using privacy preserving bloom filters compared with unencrypted probabilistic linkage.

	Precision	Recall	*f*-Measure
WA unencrypted linkage	0.999	0.981	0.990
WA encrypted linkage: bigrams	0.998	0.981	0.990
WA encrypted linkage: trigrams	0.998	0.980	0.989
NSW unencrypted linkage	0.986	0.972	0.979
NSW encrypted linkage: bigrams	0.985	0.970	0.978
NSW encrypted linkage: trigrams	0.985	0.970	0.977

almost equal results). A rules-based, deterministic linkage paradigm using bloom filters may also avoid this problem, as this method does not involve setting a matching threshold. Deterministic linkage uses a set of rules to specifically outline which combinations of matching variables will result in a record-pair match [13,28]. The 'matching threshold' for deterministic linkage is the decision of which rules, out of all possible rules, will designate a correct match.

In addition to manual inspection for quality assessment purposes or threshold setting/checking, some linkage units have adopted a clerical review and intervention process as an overall quality improvement strategy, where a small proportion of links are routinely scanned and manually reviewed [5,22,29]. While this method results in a high quality linkage, there are drawbacks; it is very expensive and time consuming, and may not be feasible for large linkage projects. Data custodians may also feel uncomfortable about the increased privacy risk when business processes require personal identifiers to be regularly manually examined.

Clerical review processes are not used by all linkage units, however; and there is little published evidence of the extent of quality improvement provided by clerical review, or whether this improvement has any effects research outcomes. Some part of the difference in quality between the results found here and the results found by the two linkage units (i.e. the reason why the *f*-measures found were 0.99 and 0.97 instead of 1) is due to clerical review (the other difference may be due to better probabilistic linkage strategies, access and linkage to additional data or more importantly the access to additional variables used in these linkages). This clerical review process is not possible using privacy preserving linkage techniques.

Notwithstanding these issues, privacy preserving linkage shows promise as an alternative to linkage with personally identifying information. Our adoption of a probabilistic linkage framework which utilised bloom filters to encode and compare data allowed large scale privacy preserving linkage to occur while providing high linkage quality. Privacy and security are both increased, while the optimum linkage quality achievable is comparable. No members of the linkage unit are able to see any of the personal identifiers used in linkage, but are able to link it to a very high quality.

Further testing of this approach on other datasets may be useful to ensure its robustness. Additional variations to the protocol may provide greater quality. It may be useful, also, to compare the bloom filter privacy preserving approach with other privacy preserving methods, such as BioGrid's GRHANITE [16]. Based on these results it appears entirely feasible to carry out a large scale linkage study using this method for linkage. When methods for determining appropriate thresholds for privacy preserving linkage have been developed and verified, it is envisaged this methodology, or similar methodologies could replace traditional record linkage.

5. Conclusion

This study has shown the feasibility of privacy preserving record linkage of large scale datasets. By using the bloom filter method to encrypt and compare individual fields, along with a probabilistic linkage framework, large scale privacy preserving linkage can occur at no cost to linkage quality. More work is currently needed to determine appropriate methods for threshold setting.

Although more work is required to refine the process for routine record linkage, it is hoped that adaptations of this method or similar privacy preserving approach will serve to reduce the privacy risks related to record linkage and therefore increase the use of record linkage to support health research. Through these efforts, we will be able to use administrative health data resources to their full extent to improve health services without compromising on individual privacy.

Acknowledgments

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy's Population Health Research Network. The authors would like to thank the reviewers for their invaluable comments.

References

[1] Butler-Henderson K. Health reform, health data and the Health Information Manager. Health Inf Manage J 2010;39:7–8.

[2] Privacy Act 1988 (Australian Commonwealth). <http://www.comlaw.gov.au/Details/C2013C00482>.

[3] Newcombe HB. Handbook of record linkage: methods for health and statistical studies, administration and business. New York: Oxford University Press; 1988.

[4] 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 Publ Health 2008;32:19–23.

[5] Gill, L. E. OX-LINK: the oxford medical record linkage system. in W. Alvey, B. Jamerson (Eds.), Record Linkage Techniques – 1997, National Academy Press, Washington DC, 1999, pp. 15-33.

[6] Kendrick S, Clarke J. The Scottish record linkage system. Health Bull 1993;51:72.

[7] 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, 4 September 2009.

[8] Roos LL, Nicol JP. A research registry: uses, development, and accuracy. J Clin Epidemiol 1999;52:39–47.

[9] 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;12:480.

[10] Kelman C, Bass A, Holman D. Research use of linked health data: a best practice protocol. Aust N Z J Publ Health 2002;26:5.

[11] Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak 2009;9.

[12] Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. Inform Syst 2013;38:946–69.

[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 Serv Res 2010;10:41.

[14] Borst F, Allaert F-A, Quantin C. The Swiss solution for anonymously chaining patient files. Stud Health Technol Inform 2001:1239–41.

[15] 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;49:117–22.

[16] Boyle DIR, Rafael N. BioGrid Australia and GRHANITE: privacy-protecting subject matching. Stud Health Technol Inform 2011;168:24–34.

[17] Durham E, Xue Y, Kantarcioglu M, Malin B. Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. Inform Fusion 2012;13:245–59.

[18] Schnell R, Bachteler T, Reiher J. A novel error-tolerant anonymous linking code. In: Editor (ed)^(eds), A novel error-tolerant anonymous linking code. Working paper series no. WP-GRLC-2011-02. Nürnberg, Germany, City: German Record Linkage Center; 2011.

[19] Kuzu M, Kantarcioglu M, Durham E, Malin B. A constraint satisfaction cryptanalysis of Bloom filters in private record linkage. In: Privacy Enhancing Technologies. Berlin, Springer; 2011.

[20] Karakasidis A, Verykios VS. Secure blocking + secure matching = secure record linkage. JCSE 2011;5:223–35.

[21] 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, Public Health Information Development Unit, Adelaide, 2002.

[22] Lawrence G, Dinh I, Taylor L. The Centre for Health Record Linkage: a new resource for health services research and evaluation. Health Inform Manage J 2008;37:60–2.

[23] Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc 1969;64:1183–210.

[24] Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. J Biomed Inform 2012;45:165–72.

[25] Owolabi O, McGregor D. Fast approximate string matching. Softw Pract Exper 1988;18:387–93.

[26] Verykios VS, Karakasidis A, Mitrogiannis VK. Privacy preserving record linkage approaches. Int J Data Min Model Manage 2009;1:206–21.

[27] Karakasidis A, Verykios VS. Secure blocking + secure matching = secure record linkage. J Comput Sci Eng 2011;5:101–6.

[28] Gomatam S, Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. Stat Med 2002;21:1485–96.

[29] Holman CDJ, Bass J, Rouse IL, Hobbs MST. Population-based linkage of health records in Western Australia: development of a health services research linked database. Aust N Z J Publ Health 1999;23:453–9.