German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2021-01 | NOVEMBER 18, 2021

Privacy Preserving Record Linkage in the Context of a National Statistical Institute

Rainer Schnell

Contents

Privacy Preserving Record Linkage in the Context of a National Statistical Institute

Rainer Schnell

November 18, 2021

Abstract: Linking databases containing records of the same person is an increasingly used technique within national statistical institutes (NSIs). No unique personal identification number is available for linkage in many countries (such as New Zealand, Australia, the UK, and Germany). In such cases, the linkage is most often based on quasi-identifiers such as surname, first name, address, and place of birth. Since knowledge of this kind of identifier is widely seen as private information, different methods for Privacy-Preserving Record Linkage (PPRL) have been developed. The report reviews existing methods for PPRL, considers privacy vs linkage quality of record linkage within an NSI, discusses the kind and probability of privacy attacks and gives recommendations for linkage practice in official statistics.

Keywords: Official Statistics, Record-Linkage, PPRL, privacy, five safes

1 Introduction[†]

Linking existing administrative data sets on the same units is used increasingly as a research strategy in many different fields. Depending on the academic field, this kind of operation has been given different names, but in application areas, this approach is mostly denoted as record linkage. Although linking data on organisations or economic entities is common, the most interesting applications of record linkage concern data on persons. Starting in medicine, this approach is now also being used in the social sciences and official statistics. Furthermore, the joint use of survey data with administrative data

[†] This paper was prepared as a contribution to the Government Statistical Service (GSS) Methodology Advisory Committee (GSS MAC) National Statistician's Quality Review 2020. The paper was published in 2020 on the homepage of the UK National Statistician's Quality Review: https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/privacy-preserving-record-linkage-in-the-context-of-a-national-statistics-institute

is now standard practice. For example, victimisation surveys are linked to police records, labour force surveys are linked to social security databases, and censuses are linked to surveys.

Merging different databases containing information on the same unit is technically trivial if all involved databases have a common identification number, such as a social security number or, as in the Scandinavian countries, a permanent personal identification number. Most of the modern identification numbers contain checksum mechanisms so that errors in these identifiers can be easily detected and corrected. Due to the many advantages of permanent personal identification numbers, similar systems have been introduced or discussed in some European countries outside Scandinavia.

In many jurisdictions, no permanent personal identification number is available for linkage. Examples are New Zealand, Australia, the UK, and Germany. Here, the linkage is most often based on alphanumeric identifiers such as surname, first name, address, and place of birth. In the literature, such identifiers are most often denoted as indirect or quasi-identifiers. Such identifiers are prone to error, for example, due to typographical errors, memory faults (previous addresses), different recordings of the same identifier (for example, swapping of substrings: reversal of first name and last name), deliberately false information (for example, year of birth) or changes of values over time (for example name changes due to marriages). Linking on exact matching information, therefore, yields only a non-randomly selected subset of records.

Furthermore, the quality of identifying information in databases containing only indirect identifiers is much lower than usually expected. Error rates in excess of 20% and more records containing incomplete or erroneous identifiers are encountered in practice.

1.1 Deterministic and probabilistic linkage

Special technical procedures are required to merge databases with errors in identifiers. Collectively these procedures are usually referred to in the statistical literature as record linkage techniques (Herzog/Scheuren/Winkler 2007).
Some systems in use are based on fixed rules if records don't match exactly.[1] Such systems are often used in local administration processes. In the literature, these rule-based systems are denoted as deterministic record linkage systems.[2]

For population covering databases, deterministic linkage systems are usually outperformed by more elaborate procedures. Despite 50 years of research in computer science and statistics, a record linkage method developed at Statistics Canada in the late 1960s is still considered as the gold standard of record linkage, regularly performing better than all other suggested procedures. Named after the authors, the Fellegi-Sunter model is based on a statistical decision model of what pair of records could be considered as a match or a non-match. The decision is based on the estimated likelihood that a set of identifiers match when the pair of records is actually a match. Methods based on this decision model are denoted as probabilistic record linkage. This approach is the base of most linkage systems in use in official statistics and many other fields.

1.2 Privacy Preserving Record Linkage (PPRL)

To protect the identity of research subjects, in many settings, identifiers and payload data are separated before linking. The identification of matching records is then done in a linkage unit. However, in some situations, even the linkage unit is distrusted and the identifiers have to be encoded. If the identifiers cannot be used as cleartext, linkage based on imperfect identifiers becomes challenging. Using traditional encodings, the difference of a single character in a name will cause about 50% of different bits in an encoded identifier. Solving the problem of matching encoded identifiers in a way that the identity of persons is not revealed is called Privacy-Preserving Record Linkage (PPRL).

2 PPRL in a National Statistics Institute

Currently, National Statistics Institutes (NSIs) are using record linkage in many steps of its production, for example, preparing sampling frames by deduplication.[3] One major

[1] A simple example for a potential rule: If the first name and last name, address, and place of birth matches, birthday could disagree.

[2] Sometimes, transformations (phonetic codes) of similar-sounding names to a code string are used in such systems. The most widely used phonetic code is Soundex. For example, Smith and Smyth both yield the Soundex code S530. It has to be noted that matching on phonetic codes is still a deterministic matching procedure, although some users of phonetic codes prefer to call them otherwise.

[3] Although the text refers mainly to official statistics, the main arguments may also apply to other governmental agencies. However, because I am more familiar with the legal framework of official statistics, I recommend that legal experts evaluate whether the statements given may also apply to other agencies. Furthermore, the analysis of the data situation (see section 2) requires detailed knowledge of the data processing details specific to other government agencies.

application of record linkage is census operations. Here record linkage is used, for example, for identification of household members and deduplicating persons with more than one place of residence.

Increasingly, many NSIs will use external databases for the production of official statistics. Here different databases of various administrations will be linked. If no national identification number is available, the linkage will be based on quasi-identifiers such as names, dates of birth and addresses. Although lower linkage quality has to be expected, standardisation of identifiers will, in many cases, be done within each administration. If the identifiers have to be encoded, the decentralisation of encodings will most likely result in the loss of matching records.

However, to evaluate the usefulness and protection given by PPRL, a model for the risks of revealing identifiers is needed. In the literature on PPRL and on Statistical Disclosure Control (SDC) such models are usually denoted as attack scenarios. Most texts concerning PPRL don't discuss attack scenarios in any detail. Since external attacks on encrypted databases are rarely mentioned in PPRL texts, implicitly, an insider attack by the linkage unit seems to be assumed.

Recently, attention of research in SDC has been directed from evaluating the disclosure risk by just looking at data to the situation, where the data is processed (Elliot et al. (2016)). This perspective is not prevalent in the current technical PPRL literature. Most texts seem to assume the same kind of worst-case models as criticized with regard to statistical disclosure control (SDC) by Ritchie and Welpton (2014):

> "Almost all papers in statistical disclosure control (...) focus on the concept of an 'intruder'. An intruder is an individual whose primary intention is to breach SDC protection, and who is prepared to devote resources and expertise to that end. In many cases, the resources available to the intruder are considerable: a matching database containing the same individuals as the target data set, full knowledge of the data collection mechanism, infinite time and patience (...). The use of worst-case scenarios makes sense in an academic context: when comparing alternative SDC measures, it provides a convenient common base against which to judge contending ideas."

In a later paper, Hafner et al. (2019) argues that "the dominant worst-case scenarios (...) are typically not that: they are the mathematically tractable worst cases. A realistic worst-case might [be] the unplanned release of some of the original data on the internet, against which no anonymisation can protect." Another example is the re-identification of a scientific-use file based on administrative data by administrative staff in the supplier organisation who have access to the original data (Hafner et al. 2019). The authors summarise their critique of worst-case scenarios in SDC as:

1. inefficient for society
2. not required by legislation

3. not supported by evidence
4. mathematically convenient rather than true worst case
5. as subjective as any other modelling base

The same kind of arguments can be made for attack scenarios on PPRL (see section 5).

3 Approaches to Privacy Preserving Record Linkage (PPRL)

Linking sensitive databases across organisations without the need to reveal any private or confidential information is usually denoted as PPRL (Christen et al. 2020). PPRL is most often applied in medical settings, but its use is increasing in other fields, for example in statistical research, when existing administrative records of different organisations are used for generating new micro-data sets.

PPRL solutions appeared about 25 years ago (Quantin et al. 1996)and are in practical use in large scale applications such as cancer registries all over the world. With the publication of a first demonstration that PPRL is possible beyond the use of hashed phonetic codes (Churches and Christen, 2004), the number of publications increased sharply. During the last decade, many different approaches to PPRL have been published. The currently known approaches can be classified according to various criteria (Vatsalan et al. 2013). For the purpose of this paper, a classification according to the central method used for the computation of distances between records might be most useful (see figure 1).

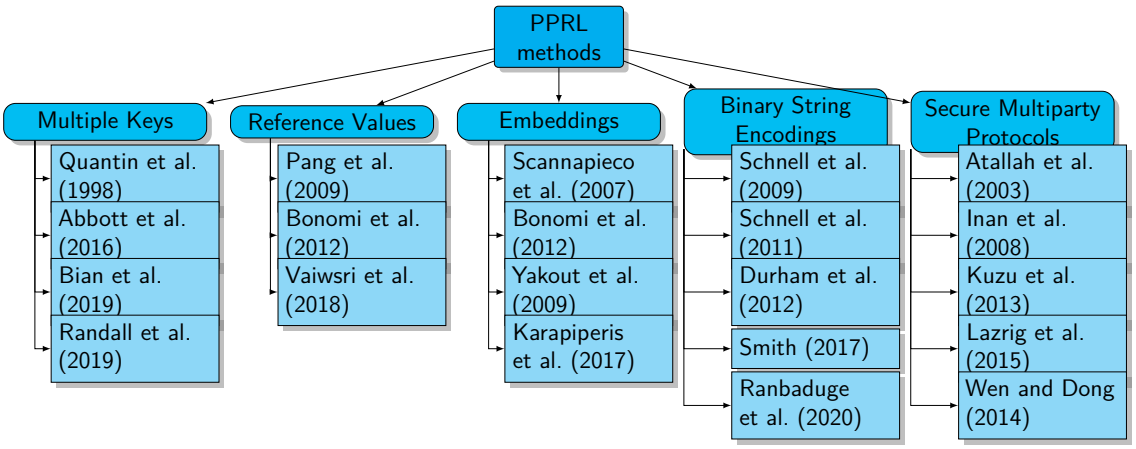

Figure 1: Main PPRL approaches, classified according to the method of distance computation

The main approaches are:

1. Multiple keys are based on combinations of transformed identifiers, such as the Soundex-code of standardized names, which are then hashed.
2. Reference values are based on the distance of records to somehow agreed tables of identifiers.

3. Embeddings transform the distance between records to distances in a different space.
4. Binary string encodings are based on hashing substrings of identifiers to a binary string, for example, Bloom Filters (BF).
5. Secure multiparty protocols use modern cryptographic tools such as homomorphic transformations to compute distances between records.

Multiple key approaches, as well as reference values and embeddings usually suffer from insufficient recall and are rarely used in practical applications. The exceptions are variants of multiple key methods which have been in use in medical settings for decades but have not been subjected to modern attacks.

Secure multiparty protocols (SMC) have the advantage that they are provable secure (given the assumptions, neither party can learn anything more than permitted by the protocol). Still, these protocols usually require either high computational or communication costs (or both). They, therefore, do not scale to population size data. For data sets of this size, computational efforts may amount to years of computation on current hardware. Hence, a demonstration of SMC based record linkage on a full census is not to be expected soon.

Binary string encodings, mostly variants of BFs, have gained popularity in research on PPRL and are beginning to be used in real-world applications. BF based techniques are currently widely considered as a reference standard for PPRL. A recent review of all approaches and an extensive treatment of BF variants is given by Christen et al. (2020).

3.1 Binary string encodings

The most popular version of binary string encodings for PPRL are Bloom filters. These have been invented by Bloom (1970) for fast set membership checks. Their use for PPRL were suggested by Schnell et al. (2009). The construction of a Bloom filter can best be explained with an example (see figure 2).

A name is split into consecutive pairs of letters (q-grams) and mapped with multiple one-way functions to a binary array, with all positions initially set to zero.

The attractive main property of Bloom filters is that they can be used to encode strings in a similarity-preserving way. One method to compute the similarity of two sets A and B of bigrams is the Dice coefficient, which is calculated as the doubled intersect of the two sets divided by the number of elements in both sets:

$$D = \frac{2|A \cap B|}{|A| + |B|}. \tag{1}$$

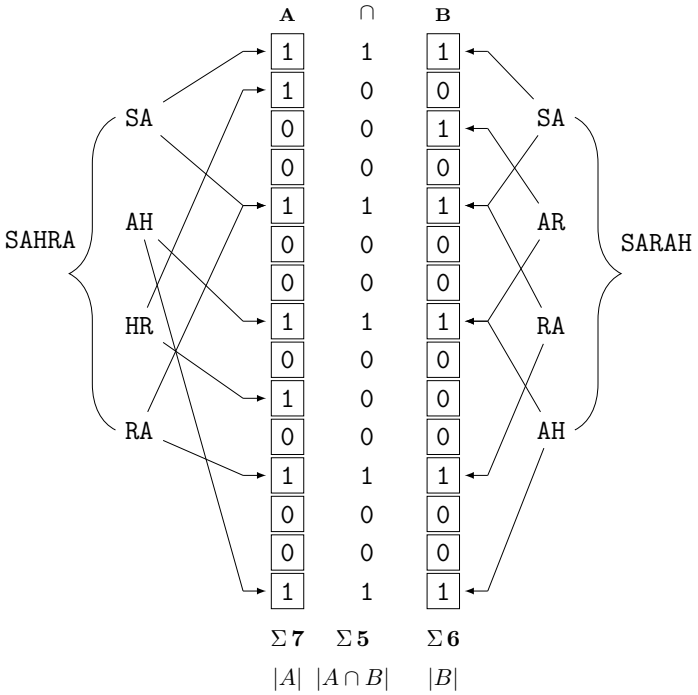

Figure 2: Example of Bloom filters for two similar names ($l = 15$, $k = 2$)

As can be seen in Figure 2, the names Sahra and Sarah share three out of four bigrams (subsets of $q = 2$). This gives an unencoded cleartext Dice similarity of $D = \frac{2*3}{4+4} = 0.75$.

For Bloom filters, the Dice similarity can be computed by comparing the sets of bit positions of two Bloom filters. Here, the Dice similarity of the Bloom filters is very close to the unencoded bigram similarity, as both Bloom filters have 6 respective 7 bits set to one, while sharing 5 bit positions. This gives a Dice coefficient for the encoded names of $D_{BF} = \frac{2*5}{7+6} \approx 0.77$.

Bloom filters have been used for encoding numerical attributes, dates (Vatsalan and Christen, 2016) and geographical information (Farrow, 2014) as well. Recently, a technique for encoding hierarchical codes such as disease classifications (for example, ICD-10) or occupational codes (for example, ISCO-08) has been suggested by Schnell and Borgs (2019).
The basic variant of BF has been replaced by modern versions using different hash functions (random-hashing, Niedermeyer et al. 2014), an encoding of many identifiers into one binary string, and the use of different encodings for different identifiers (salting, see section 3.5). Therefore, it is crucial which kind of BFs have been used in an application or an attack. It has to be emphasized that no successful attack has been reported on BFs using the mentioned combination of different encodings, many identifiers, salting, and random hashing.

Since BF encodings are subject of intensive research in theoretical PPRL, many different versions of binary encodings based on the same idea (splitting identifiers, mapping to a binary array, comparing the mappings) have been suggested. Two recent examples will be detailed.

3.2 Tabulation Min-hash Encoding

Smith (2017) suggested a complex, multi-step encoding, where sensitive attributes are mapped to binary arrays of length l. First, l sets of look-up tables c, containing keys, each pointing to a random bit string are created. One-way hashing each element yields a binary value, which is split into c sub-keys used as an index to the look-up-tables. The returned random strings are XORed to finally get a bit string for each of the l positions. Using only the least significant bit of a min-hash signature of the bit string yields finally the encoding of the identifiers.

The overall computational time required is increased, but a direct attack on the encoding seems to be complicated. Large scale evaluations of precision and recall of this approach have not been published yet. Furthermore, none of the published attacks included this proposal. However, research under review seems to indicate the same vulnerabilities as in other PPRL approaches to similarity alignment attacks.

3.3 Two-step Hash Encoding Process

Recently, Ranbaduge et al. (2020) suggested a new encoding method based on a two step approach (see table 1):

1. Quasi-identifying values in records are first converted into character q-gram sets Q (for example, peter $-> \{pe, et, te, er\}$).
2. In the first hashing step k hash functions are used where each hash function H encodes Q into a bit vector B of length l.
3. In the second hashing step a hash function (G) encodes the concatenated bit values in each column position in all bit vectors into an encoded integer value, e.
4. Finally, all encoded integer values are added into a list E which is used to compute similarities between records.

Computational speed and linkage quality are similar to previous approaches. The approach seems to be more resilient against known attacks. However, as with all other PPRL approaches, unsuitable parameter choices may still allow similarity alignments. This is the subject of ongoing work at the Australian National University (Canberra).

3.4 Attacks on PPRL

The goal of PPRL is record linkage without revealing identifiers. To verify the attainment of this goal, PPRL encodings need to be attacked. Since PPRL uses either pseudo-random

$Q_1 = \{\text{pe, et, te, er}\}$

h_1	0	1	0	1	0	1	1	0
h_2	1	1	0	0	1	0	1	0
h_3	1	0	0	1	1	0	1	0
h_4	0	0	1	0	1	1	0	1
$G(B_p)$	↓	↓	↓	↓	↓	↓	↓	↓

$E_1 = \{7,\ 28,\ 42,\ 66,\ 97,\ 110,\ 137,\ 158\}$

(a) Encoding 'peter'

$Q_2 = \{\text{pe, et, te}\}$

h_1	0	1	0	1	0	1	0	0
h_2	1	1	0	0	1	0	0	0
h_3	1	0	0	1	1	0	0	0
h_4	0	0	0	0	1	1	0	1
$G(B_p)$	↓	↓	↓	↓	↓	↓	↓	↓

$E_2 = \{7,\ 28,\ 66,\ 97,\ 110,\ 158\}$

(b) Encoding 'pete'

Table 1: 2-Step Column Hashing (example taken from Ranbuge/Christen/Schnell (2020))

number streams (PRNG, Johnston, 2018) or hash functions (Stallings, 2017) as building blocks, the encodings themselves are hard to attack.

Up to now, only one direct attack on BF-encodings has been published (Niedermeyer et al. 2014). This attack exploited a weakness of the computationally efficient but insecure hash-function used for the encoding (double hashing, Kirsch and Mitzenmacher, 2006) as an attack vector. Based on this attack, many modern PPRL implementations use PRNG instead of hash-functions for encoding.

Therefore, all other published attacks on PPRL are either frequency attacks or similarity alignments (or both). Many, but not all, of the attacks are assuming an attacker with a de-identified copy of the data set to be attacked. Other approaches assume an attacker with detailed knowledge of all encoding parameters. Both assumptions are unrealistic in practice.

However, since BF encoding has gained popularity in recent years and is now considered as the de-facto standard for PPRL (Smith, 2017), most recently published attacks concentrate on BF (for an extensive review, see Christen et al. 2020). An example for a modern attack is the use of pattern mining (Christen et al. 2018); Vidanage et al. 2019), where jointly occurring bit-positions are used for the identification of q-grams and in the end, identifiers. A graph-matching attack was used by Culnane et al. (2017) to match similarities between encoded and unencoded identifiers. The same basic idea of graph-matching can be used for attacking other encodings.

The partial success of some attacks might give the false impression that other PPRL approaches are safer than BF – except for SMC, which are provable secure, but unusable in practical applications for years to come.

The false impression that more traditional approaches are more resilient against frequency attacks or similarity alignments has been invalidated repeatedly, for example by attacks described by Culnane et al. (2017) and Vidanage et al. (2020).

At present, we have neither generally accepted privacy metrics nor realistic attack scenarios. Given this lack of standard evaluation criteria, it is impossible to give mathematical proofs of security. However, there is ongoing research on such standards. Currently, we have to evaluate PPRL proposals by comparing the efforts required and the reported success of attacks. The main problem here is the absence of a realistic attack model.

3.5 Bloom Filter Hardenings

Hardening methods are modifications of bit-arrays representing identifiers to make attacks more difficult. If, for example, an attack is based on an analysis of logical implications of column or row combinations, adding bits set to one randomly would reduce the number of records not contradicting the assumption tested. Therefore, the number of records available for a frequency attack would be smaller, and the uncertainty of alignments larger.

Examples of hardening methods are:

1. encoding of different identifiers in the same bit-string (CLK, Schnell et al. 2011 and RBF, Durham et al. 2014),
2. using different encodings depending on a stable identifier (Salting, Niedermeyer et al. 2014),
3. removing information on the number of encoded elements (Balancing, Schnell and Borgs 2016a),
4. adding random bits to implement a random response mechanism to each bit (BLIP alias RAPPOR, Alaggan et al. 2012, Erlingsson et al. 2014, Schnell and Borgs 2016a),
5. preventing row-wise logical deductions (XOR-Folding, Schnell and Borgs 2016b),
6. adding random identifiers (Markov-Chain-BF, Schnell and Borgs 2018b) and
7. increasing error propagation by local XOR-ing (Rule-90, Schnell and Borgs 2018a).

Further methods are available in print and two others are currently under review. A review of all hardening methods published so far will be published in 2020 (Christen et al. 2020).

Many of the hardening methods have been developed to prevent a specific kind of attack. Most likely, combining different methods will be more difficult to attack than each technique as such. Studies on the effect of combining different hardening methods on attacks have not been published yet. Currently, at least two groups are working on this.

However, it should be made clear, that no success of *any* attack has been reported after salting. Furthermore, since salting of bit-arrays will pose the problem of identifying mixtures of binary strings for an attacker, there is currently no known computational

approach for such an attack. So, salting should be an essential building block for PPRL techniques.

4 Trade-off between preserving privacy and linkage quality

Like other linkage techniques, most PPRL methods can be tuned to yield either optimal recall or optimal precision. The choice of an optimal threshold depends on the loss function associated with the application.

However, even for the same application, the loss functions may vary over time or between NSIs, depending, among other constraints, on political requirements.

For example, most census operations in Germany prefer a high rate of false-negatives to even a modest number of false-positives. That might result in counter-productive decisions: If we restrict the set of records to be linked to exact unique matching records, we could standardize, concatenate and hash identifiers. This line of reasoning ends in a recently suggested modification of a protocol implying deleting all records which might compromise security in a specific sense. Although this solution might be of help in some exceptional cases, in general deleting information is not the aim of information processing.

To be useful, identifiers have to be linkable in principle. We can impose restrictions on linkability, but we have to accept a non-zero risk of potential re-identifications. Otherwise, no linkage can be done. However, as discussed above, the attack models used for the evaluation of risk are not realistic in the real world and especially not in the context of an NSI. Therefore, in a secure environment of a national statistical institute, there is hardly any reason to compromise linkage quality in favour of an abstract re-identification risk based on an unclear attack model.

4.1 Advantages of Cleartext

The main problems of record linkage are errors in identifiers and missing identifiers. The last topic is neglected in the literature, but of utmost importance in real-world applications. In practice, most often a hierarchical structure of secondary linkage rules is used for missing identifiers. Applying a hierarchical scheme will result in intransitive linkage decisions, but these are usually accepted as unavoidable. For a certain amount of records, none of the secondary linkage rules will yield a clear linkage decision. For this set of unlinked records, clerical decisions are needed. These clerical decisions are either based on human processing of incomplete or erroneous records (guessing probable and plausible links) or on acquiring additional information not already in the data set. Both human interventions are close to impossible with encoded identifiers in a PPRL process.

The amount of clerical post-processing in large scale record linkage applications such as a census is widely underestimated by those not involved in such operations in practice. In general population data sets, after pre-processing, more than half of the records can

be linked without any problems. Most of the remaining records do not match exactly, but with small edit-distances between their identifiers. The remaining records usually have either large edit distances (implying many errors) or require additional information to be resolved, most likely by a manual search for additional information not contained in the data sets. Records from the last group are usually either lost by PPRL or linked with a high risk of false-positive links.

Therefore, in large scale real-world applications, linking cleartext with unstable identifiers will most likely outperform any PPRL solution. Furthermore, cleartext linkage is usually much faster. Finally, cleartext allows modifying procedures after additional data or new techniques have become available.

4.2 Advantages of PPRL within an NSI

Although within an NSI the disadvantages of PPRL outweigh their advantages in general, there are some aspects, which might require PPRL techniques or where these techniques might be useful. Even in applications, where privacy is no legal requirement, some PPRL approaches may improve linkages. A simple example is the reversal of identifiers. Another example is the use of error-tolerant blocking methods within a blocking strategy.

If, however, organisations mistrust each other, there is hardly any alternative to PPRL. If databases across organisations competing for control or resources have to be shared, cleartext linkage – which might provide the other organisation with knowledge about the intersection of databases – might be difficult to arrange.

In general, PPRL might protect against 'honest, but curious' employees of an NSI. That will be a smaller problem in the UK than in other countries. However, even a negligible chance of misuse, whatever that might be in the case of information available in a statistical database, might gain the attention of concerned citizens, interest groups, or disinformation campaigns. If such problems are to be expected, PPRL might add an additional layer of legitimation. However, official statistics and record linkage in general, should be based on citizens' trust and a social licence and not a technical solution alone (see, for example, Jones and Ford, 2018; Moore et al. 2016).

Finally, there is an essential statistical advantage of PPRL in an NSI compared to PPRL in other environments, for example, a medical application. The main disadvantages of PPRL, false positive or false negative links, are less important in an NSI than in a medical application. In medicine (for example, in a non-linear statistical model), single cases might result in the rejection of a model or heavily biased estimates. In official statistics, low rates of false positive (and false negative) links will have less impact.

5 Attacks on PPRL within an NSI

Usually, payload data and identifiers are separated as early as possible in data processing tasks. Given this separation principle and the widely adopted five safes policy (Desai et al. 2016), the PPRL literature concentrates on attacks either within a linkage unit or on encoded identifiers from the outside. These attack models are hardly realistic within an NSI.

Employees of an NSI are aware of constant monitoring of their data access, and proper IT management should be able to limit physical access to the sensitive data. Attacks by the linkage unit are, therefore, implausible. Given nearly no incentive, a high probability of detection and the expectation of severe sanctions will limit attacks by employees to rare exceptions. Not one case of such attacks has been publicly described.

An attack from the outside on encoded data is inefficient since the personal information within an NSI is economically useless. Public blaming of the NSI seems to be the only rational motive for an outside attack. Defending against such an attack is mostly an IT security problem.

Social engineering or carelessness of employees seem to be much larger risks caused by data leakage of unencoded payload data than an attack on encoded identifiers.

If external data sets are used, and the resulting micro-data set is released, linking the external data to the resulting data set – as described above – is a much more efficient attack than breaking the encoding. Technically, this is a statistical disclosure control problem, not a PPRL problem. Therefore, SDC techniques might be more appropriate to minimize this risk.

In sum, in the context of an NSI inside attacks on PPRL are psychologically implausible and outside attacks will most likely concentrate on other attack vectors. To cite Hafner et al. (2019) again: "(. . .) protecting against any attack by any person at any time in the future is an impossible standard, and no law requires it."

6 Recommendations for UK government statistics

1. In the long run, establishing a unique national identification number to be used in many different contexts seems unavoidable if linking data will be the central process for producing official statistics. As in other countries, the national identification number will form the base for project-specific identifiers (Winterleitner and Spichiger, 2018; Körner et al. 2017).

2. However, as long as no unique and population covering national identification number is available in all data sets of interest, the linkage has to use unstable identifiers such as names, date of birth, place of birth, and address. Cleartext for linking such identifiers will always outperform PPRL solutions. Therefore, since PPRL is not required by law for the current tasks of the Office for National

Statistics (ONS), it should not be used for linking ONS internal data sets.

3. If resources are limited, research efforts for linking procedures should concentrate on the use of additional information. Either models for changes to identifiers over time, for example, due to movements or family processes can be used or data from additional sources, such as utility companies.

4. Since the demand for linking external data sets not containing other identifiers than names or similar identifying attributes will increase in the next decade, developing PPRL solutions will be required for linking those data sets.

5. In the absence of realistic attack scenarios, the development and evaluation of PPRL solutions are challenging. A general agreement on the kind of attacks to be prevented would be an important first step for new technical solutions. The clarification of the kind of attacks to be expected will require joint efforts by statisticians, computer scientists, cryptographers, social scientists, and lawyers.

6. Involving universities in the development of new technical PPRL solutions is unavoidable. One of the main obstacles in the development of better PPRL solutions is the lack of large-scale test data containing all errors and problems found in practice and nearly impossible to simulate. Therefore, official statistics should provide test data (including synthetic data sets) resulting from national linkage projects using cleartext for academic research developing new PPRL approaches and attacks. No other single measure will more improve the invention and testing of new methods for record linkage.

References

Abbott, O., Jones, P., and Ralphs, M. (2016). Large-scale linkage for total populations in official statistics. In Harron, K., Goldstein, H., and Dibben, C., editors, *Methodological Developments in Data Linkage*, pages 170–200. Wiley, Chichester.

Alaggan, M., Gambs, S., and Kermarrec, A.-M. (2012). BLIP: non-interactive differentially-private similarity computation on Bloom filters. In *Symposium on Self-Stabilizing Systems*, pages 202–216.

Atallah, M., Kerschbaum, F., and Du, W. (2003). Secure and private sequence comparisons. In *Workshop on Privacy in the Electronic Society*, pages 39–44. ACM.

Bian, J., Loiacono, A., Sura, A., Mendoza Viramontes, T., Lipori, G., Guo, Y., Shenkman, E., and Hogan, W. (2019). Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. *JAMIA Open*, 2(4):562–569.

Bloom, B. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426.

Bonomi, L., Xiong, L., Chen, R., and Fung, B. (2012). Frequent grams based embedding for privacy preserving record linkage. In *ACM CIKM*, pages 1597–1601, Maui, Hawaii.

Christen, P., Ranbaduge, T., and Schnell, R. (2020). *Linking Sensitive Data. Methods and Techniques for Practical Privacy-Preserving Information Sharing*. Springer, Berlin, in press.

Christen, P., Ranbaduge, T., Vatsalan, D., and Schnell, R. (2018). Precise and fast cryptanalysis for bloom filter based privacy-preserving record linkage. *IEEE Transactions on Knowledge and Data Engineering*.

Churches, T. and Christen, P. (2004). Some methods for blindfolded record linkage. *BioMed Central Medical Informatics and Decision Making*, 4(9).

Culnane, C., Rubinstein, B. I., and Teague, V. (2017). Vulnerabilities in the use of similarity tables in combination with pseudonymisation to preserve data privacy in the UK Office for National Statistics' privacy-preserving record linkage. *arXiv preprint arXiv:1712.00871*.

Desai, T., Ritchie, F., and Welpton, R. (2016). Five safes: Designing data access for research. Technical report, Department of Accounting, Economics and Finance, Bristol Business School, University of the West of England.

Durham, E., Kantarcioglu, M., Xue, Y., Toth, C., Kuzu, M., and Malin, B. (2014). Composite Bloom filters for secure record linkage. *IEEE TKDE*, 26(12):2956–2968.

Durham, E., Xue, Y., Kantarcioglu, M., and Malin, B. (2012). Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. *Information Fusion*, 13(4):245–259.

Elliot, M., Mackey, E., O'Hara, K., and Tudor, C. (2016). *The anonymisation decision-making framework*. UKAN Manchester.

Erlingsson, Ú., Pihur, V., and Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 1054–1067.

Farrow, J. M. (2014). Comparing geospatial distance without revealing location. Presenation given at the International Health Data Linkage Conference, Vancouver.

Hafner, H.-P., Lenz, R., and Ritchie, F. (2019). User-focused threat identification for anonymised microdata. *Statistical Journal of the IAOS*, 35(4):703–713.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer, New York.

Inan, A., Kantarcioglu, M., Bertino, E., and Scannapieco, M. (2008). A hybrid approach to private record linkage. In *IEEE ICDE*, pages 496–505.

Johnston, D. (2018). *Random Number Generators—Principles and Practices: A Guide for Engineers and Programmers*. Walter de Gruyter GmbH & Co KG.

Jones, K. H. and Ford, D. V. (2018). Population data science: advancing the safe use of population data for public benefit. *Epidemiology and health*, 40.

Karapiperis, D., Gkoulalas-Divanis, A., and Verykios, V. S. (2017). Distance-aware encoding of numerical values for privacy-preserving record linkage. In *IEEE ICDE*, pages 135–138.

Kirsch, A. and Mitzenmacher, M. (2006). Less hashing, same performance: building a better Bloom filter. In *European Symposium on Algorithms*, pages 456–467.

Körner, T., Krause, A., Ramsauer, K., and Ullmann, P. (2017). *Registernutzung in Zensus und Bevölkerungsstatistik in Österreich und der Schweiz*. Destatis, Wiesbaden.

Kuzu, M., Kantarcioglu, M., Inan, A., Bertino, E., Durham, E., and Malin, B. (2013). Efficient privacy-aware record integration. In *ACM EDBT*.

Lazrig, I., Moataz, T., Ray, I., Ray, I., Ong, T., Kahn, M., Cuppens, F., and Cuppens, N. (2015). Privacy preserving record matching using automated semi-trusted broker. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 103–118. Springer.

Moore, H. C., Guiver, T., Woollacott, A., de Klerk, N., and Gidding, H. F. (2016). Establishing a process for conducting cross-jurisdictional record linkage in Australia. *Australian and New Zealand Journal of Public Health*, 40(2):159–164.

Niedermeyer, F., Steinmetzer, S., Kroll, M., and Schnell, R. (2014). Cryptanalysis of basic Bloom filters used for privacy preserving record linkage. *Journal of Privacy and Confidentiality*, 6(2):59–79.

Pang, C., Gu, L., Hansen, D., and Maeder, A. (2009). Privacy-preserving fuzzy matching using a public reference table. *Intelligent Patient Management*, pages 71–89.

Quantin, C., Bouzelat, H., Allaert, F.-A., Benhamiche, A.-M., Faivre, J., and Dusserre, L. (1998). Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. *Methods of Information in Medicine*, 37(3):271–277.

Quantin, C., Bouzelat, H., and Dusserre, L. (1996). Irreversible encryption method by generation of polynomials. *Medical Informatics and the Internet in Medicine*, 21(2):113–121.

Ranbaduge, T., Christen, P., and Schnell, R. (2020). Secure and accurate two-step hash encoding for privacy-preserving record linkage. In *PAKDD*, Singapore.

Randall, S., Brown, A. P., Ferrante, A. M., and Boyd, J. H. (2019). Privacy preserving linkage using multiple dynamic match keys. *International Journal of Population Data Science*, 4(1).

Ritchie, F. and Welpton, R. (2014). Addressing the human factor in data access: incentive compatibility, legitimacy and cost-effectiveness in public data resources. Economics Working Paper Series 141, University of the West England, Bristol.

Scannapieco, M., Figotin, I., Bertino, E., and Elmagarmid, A. (2007). Privacy preserving schema and data matching. In *ACM SIGMOD*, pages 653–664.

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

Schnell, R., Bachteler, T., and Reiher, J. (2011). A novel error-tolerant anonymous linking code. *German Record Linkage Center*, (WP-GRLC-2011-02).

Schnell, R. and Borgs, C. (2016a). Randomized response and balanced Bloom filters for privacy preserving record linkage. In *ICDMW DINA*, Barcelona.

Schnell, R. and Borgs, C. (2016b). XOR-folding for Bloom filter-based encryptions for privacy-preserving record linkage. *German Record Linkage Center*, (WP-GRLC-2016-03).

Schnell, R. and Borgs, C. (2018a). Hardening encrypted patient names against cryptographic attacks using cellular automata. In *ICDMW DINA*, Singapore.

Schnell, R. and Borgs, C. (2018b). Protecting record linkage identifiers using a language model for patient names. *Studies in Health Technology and Informatics*, 253:91–95.

Schnell, R. and Borgs, C. (2019). Encoding hierarchical classification codes for privacy-preserving record linkage using Bloom filters. In *ECML/PKDD DINA*, Würzburg.

Smith, D. (2017). Secure pseudonymisation for privacy-preserving probabilistic record linkage. *Journal of Information Security and Applications*, 34:271–279.

Stallings, W. (2017). *Cryptography and Network Security: Principles and Practice*. Pearson Education Limited, Boston, 7 edition.

Vaiwsri, S., Ranbaduge, T., and Christen, P. (2018). Reference values based hardening for bloom filters based privacy-preserving record linkage. In *Australasian Conference on Data Mining*, pages 189–202, Bathurst.

Vatsalan, D. and Christen, P. (2016). Privacy-preserving matching of similar patients. *Journal of Biomedical Informatics*, 59:285–298.

Vatsalan, D., Christen, P., and Verykios, V. (2013). A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969.

Vidanage, A., Ranbaduge, T., Christen, P., and Randall, S. (2020). A privacy attack on multiple dynamic match-key based privacy-preserving record linkage. *under review*.

Vidanage, A., Ranbaduge, T., Christen, P., and Schnell, R. (2019). Efficient pattern mining based cryptanalysis for privacy-preserving record linkage. In *IEEE ICDE*, Macau.

Wen, Z. and Dong, C. (2014). Efficient protocols for private record linkage. In *ACM SAC*, pages 1688–1694.

Winterleitner, A. D. and Spichiger, A. (2018). Personenidentifikatoren: Analyse der gesamtschweizerischen Kosten. In Stember, J., Eixelsberger, W., and Spichiger, A., editors, *Wirkungen von E-Government: Impulse für eine wirkungsgesteuerte und technikinduzierte Verwaltungsreform*, pages 383–424. Springer, Wiesbaden.

Yakout, M., Atallah, M., and Elmagarmid, A. (2009). Efficient private record linkage. In *IEEE ICDE*, pages 1283–1286.

IMPRINT

Publisher

German Record-Linkage Center
Regensburger Str. 100
D-90478 Nuremberg

Editor

Rainer Schnell

Template layout

Christine Weidmann

Download

www.record-linkage.de

The German Record Linkage Center is funded by the German Research Foundation (DFG).

DOI: 10.17185/duepublico/75488
URN: urn:nbn:de:hbz:465-20220317-181910-3