Privacy-preserving Methods for Linking Big Data and Survey Data sets

Rainer Schnell Christian Borgs

German Record Linkage Center
University of Duisburg-Essen, Germany

BigSurv 2018
Barcelona, Spain, 25–27 October, 2018

German RLC

UNIVERSITÄT DUISBURG ESSEN

Introduction

- For the prediction of social phenomena, a scientific explanation requires information on covariates considered to influence individual behaviour.
- Until recently, social scientists mainly used experimental or survey data to obtain this kind of information.
- The increasing production of 'big data' sets such as transactional data, sensor data, social media data and administrative data seem to open new pathways to answer research questions.
- However, most 'big data' sets contain only very few covariates.
- Therefore, most applications of big data for a predictive social science will need linkage of datasets.
- In practice, this requires the one-to-one identification of individuals in different datasets.

Big data sources

- Linking large-scale administrative databases with data of the same type with survey data is technically simple compared to other linkage problems.
- Although some other 'big data' sources are already specific to individuals, e.g. transactional data, other databases, such as sensor data or social media data, are not directly related to specific individuals.
- To use this kind of non-identified data, for most explanatory applications, it has to be linked to data specific to individuals. Therefore, identifiers have to be present in both data sets.
- In the absence of identifying information in the data, linking 'big data' to survey data on individual respondents is impossible.

Linking data

In general, there are three ways to link data on persons:

1. using a unique identifier (Personal Identification Numbers).
2. using unencrypted pseudo-identifiers (e.g. names, birthday).
3. using encrypted pseudo-identifiers (e.g. hashed names and birthday).

Linking data

- If a unique ID is available, linking is a trivial merge operation.
- If no ID is available, either unencrypted identifiers or encrypted identifiers have to be used.
- More often than not, legal constraints require encryptions.
- For example, pseudonymization is strongly recommended for record linkage under EU regulations (EU Council Regulation No. 679/2016).
- This legal requirement gave rise to the very active research field of Privacy-preserving Record Linkage (PPRL), which is devoted to methods for error-tolerant linking using encrypted identifiers only.

Privacy-Preserving Record Linkage (PPRL)

- Error-tolerant linking of encrypted pseudo-identifiers is the goal of PPRL.
- It has to be kept in mind that PPRL methods are usually slightly inferior to clear-text linkage in terms of linkage quality.
- However, PPRL enables new opportunities to link data that would otherwise not be available for research purposes.
- Some examples include linking surveys to administrative data, health records, criminal records or credit score data.

PPRL Methods

- Although many different methods for PPRL have been suggested (for an overview, see Vatsalan/Christen/Verykios (2013)), only very few techniques are suitable for the standard setting in social sciences (Schnell 2015).
- Usually, many different organizations provide data on the same unit according to a centrally specified protocol on encrypting identifiers.
- Repeated on-line communication and computations between the data providers are usually not permitted by the data protection agencies.
- Therefore, the actual linking is being done by a trusted third party using encrypted identifiers.
- In the technical literature, this kind of setting is considered as a two-party protocol.

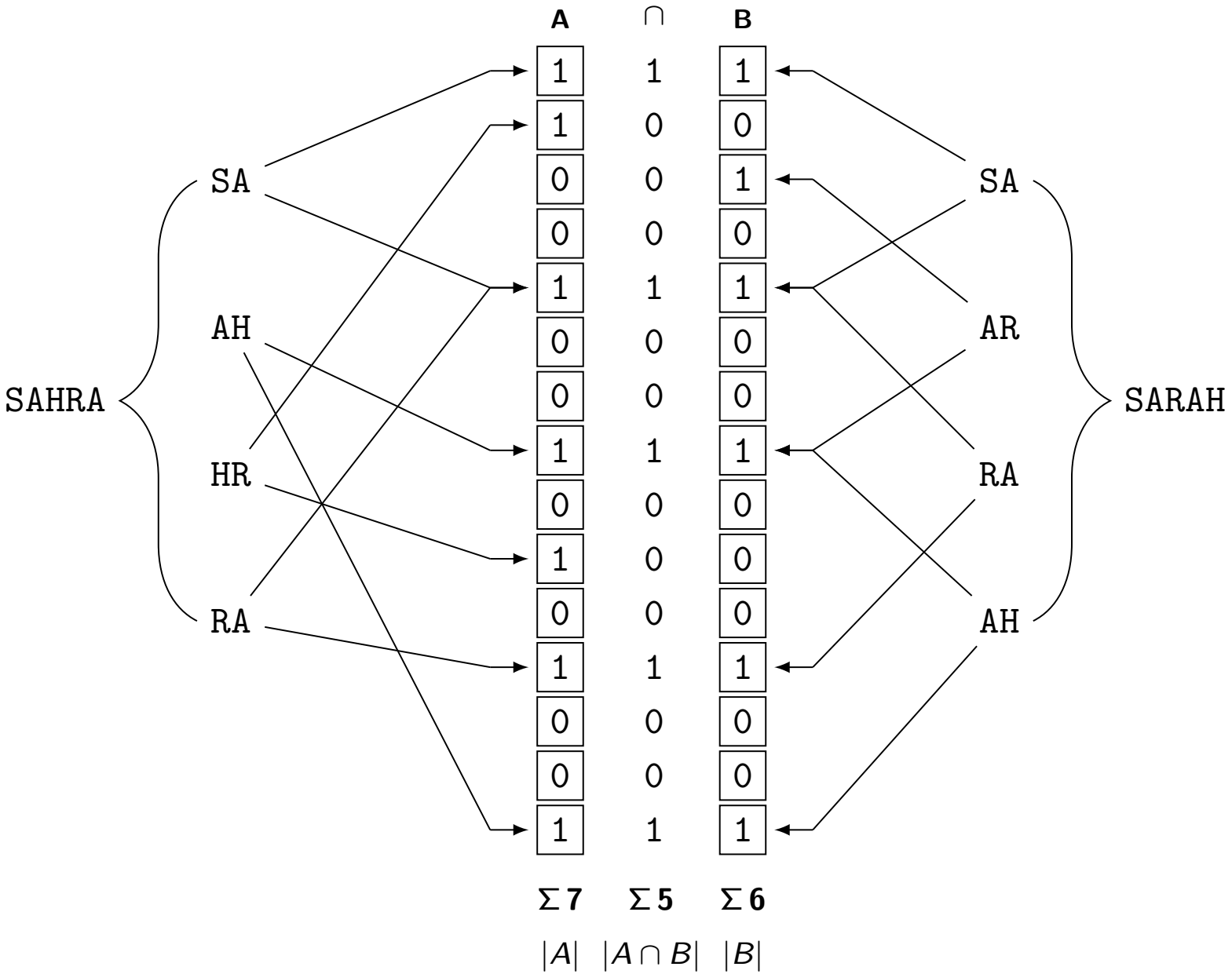
Exemplary PPRL methods: ESLs

- Only two variants are widely discussed (Randall et al. 2016): The use of Encrypted Statistical Linkage Keys (ESLs) and the use of Bloom filters.
- Linkage keys are based on encrypted phonetic codes or subsets of identifiers, for example the encryption of concatenated identifiers by cryptographic functions such as SHA-3.
- Example ESL:
  - John O'Shea, 1.9.1967, male
  - OHSHA01091967M
  - ab76990b084b82d3e06701c52d02485e8e2ba9fe

Exemplary PPRL methods: Bloom filter-based PPRL

- The application of Bloom filters for PPRL (as suggested by Schnell/Bachteler/Reiher (2009)) is now widely discussed in the literature, because the encryption is similarity-preserving.
- Here, the identifiers are split into bigrams, which are then hashed into a bit vector of a fixed length.
- To protect against cryptographic attacks, the recommended technique for mapping bigrams into a Bloom filter is the use of randomly sampled bit positions using the bigram together with a password as a seed for a random number generator, such as Salsa20 (Bernstein 2008).
- For additional security, a composite Bloom filter (CLK) can be used (Schnell 2014).

Example: Bloom Filter construction



Similarity-preserving properties

- Bigrams of two first names were mapped to Bloom filters with a length of $l = 15$ bits using $k = 2$ hash functions.
- `SAHRA` and `SARAH` both share 3 out of 4 bigrams (`SA`, `AH` and `RA`) and differ on one single bigram (`HR` and `AR`, respectively).
- The unencrypted Dice coefficient for the bigrams of both names is:

$$D(A, B) = \frac{2 \cdot |A \cap B|}{|A| + |B|} = \frac{2 \cdot 3}{4 + 4} = 0.75$$

- For the Bloom filters, both show 7 bits set to one, while having 5 common bit positions. This way, the Dice coefficient can be estimated as:

$$D(A, B) = \frac{2 \cdot |A \cap B|}{|A| + |B|} = \frac{2 \cdot 5}{7 + 7} \approx 0.71$$

- This allows estimating the true n-gram similarity using encrypted identifiers.

Security of PPRL methods

- Since similarities are preserved, they can be exploited as a method to attack Bloom filters.
- Both ESLs and Bloom filters are, in theory, attackable using frequency attacks.
- Currently, three attacks with incrementally better results have been published (for an overview, see the newest attack by Christen et al. (2017)).
- To counteract all current attacks, several hardening methods have been suggested.
- Using state of the art modifications as described in Schnell (2015) and Schnell/Borgs (2018), so far, no published attack was successful.

PPRL for large-scale data

- For linking very large databases, it is computationally infeasible to perform a similarity calculation for all record pairs.
- Therefore, blocks which most likely contain correct pairs are formed before the actual merging.
- For example, the zip code is often used.
- CLKs, however, make blocking possible with more sophisticated methods.

Multibit trees

- Multibit trees are rooted in chemo-informatics (Kristensen/Nielsen/Pedersen 2010).
- Using them for blocking in PPRL was proposed by Bachteler/Reiher/Schnell (2013).
- First, an index tree is built with the larger data set. After this, the smaller data set is then searched sequentially.
- Pair combinations above a user-set similarity threshold (Tanimoto coefficient) are used as a block.

Computational speed

- Linking with Multibit Trees is feasible on standard office hardware.
- For example, compare the following published timings for linking administrative data using Bloom filters (Schnell (2012), Brown et al. (2016) and Schnell/Borgs (2016)), with a re-test done in 2018:

2012:	Two files with 1 million records each:	~5 hours.
2016:	Two files with 1 million records each:	~2 hours.
2018:	Two files with 1 million records each:	38 minutes.
2012:	Two files with 2 million records each:	25 hours.
2016:	Two files with 5 million records each:	22 hours.
2018:	Two files with 5 million records each:	~21 hours.

- In the long run, better hardware and optimizations will further reduce the computing time required.

PPRL using R

- We wrote an R-package to implement most current PPRL techniques in R.
- It can be found on CRAN: `https://cran.biodisk.org/web/packages/PPRL/index.html`
- We will demonstrate the complete process using a small example.

PPRL using R

```
library(PPRL)

# Load test data
testFile <- file.path(path.package("PPRL"),
    "extdata/testdata.csv")
testData <-read.csv(testFile, head = FALSE, sep = "\t",
    colClasses = "character")
```

V1	V2	V3	V4	V5	V6	V7	V8	V9
12345	M	1964	01	01	Here Town	John	Doe	01.01.1964
12346	F	1970	02	31	There City	Jane	Doe	31.02.1970

```
## Encode data
CLK <- CreateCLK(ID = testData$V1,
    data = testData[, c(2, 3, 7, 8)],
    padding = c(0, 0, 1, 1), q = c(1, 1, 2, 2),
    k = 20, l = 1000,
    password = c("HUh4q", "lkjg", "klh", "Klk5"))
```

ID	CLKs
12345	000001100000.....01100
12346	000001100010.....00000

PPRL using R

```
# Create erroneous copy
testData2 <- testData
testData2$V7[testData2$V7 == "Jane"] <- "Janey"

# Encrypt the same way
CLK2 <- CreateCLK(ID = testData2$V1,
data = testData2[, c(2, 3, 7, 8)],
padding = c(0, 0, 1, 1), q = c(1, 1, 2, 2),
k = 20, l = 1000,
password = c("HUh4q", "lkjg", "klh", "Klk5"))
```

V1	V2	V3	V4	V5	V6	V7	V8	V9
12345	M	1964	01	01	Here Town	John	Doe	01.01.1964
12346	F	1970	02	31	There City	Janey	Doe	31.02.1970

PPRL using R

```
# Define Bloom filter column in data and
# select similarity function and threshold
# mtan: Tanimoto-similarity using Multibit trees
lBF <- SelectSimilarityFunctionBF("CLKs","CLKs",
method = "mtan", threshold = 0.85)

# Calculate result
linked <- BloomFilterLinkage(CLK$ID, CLK, CLK2$ID, CLK2,
blocking = NULL, similarity = lBF)
```

ID1	ID2	similarity
12345	12345	1.00
12346	12346	0.9212296

Conclusion

- Current PPRL techniques allow linking of even very large data sets and most big data sources containing natural persons in reasonable time.
- State of the art encryption methods are currently hard to break.
- We plan on doing a systematic evaluation of
  - the speed of all current blocking methods and
  - the security of all current encryption methods.
- We expect further improvements in security, practicability and computational speed as a result.

Best-practice suggestions for Bloom filter-based PPRL for large-scale data

- For linkage in general, use as many stable identifiers as possible.
- For increased security, use CLKs instead of Bloom filters.
- Random number generators instead of hash functions are strongly recommended.
- For each identifier, a different seed for the generator should be used.
- Further increases in security can be achieved by using a stable block variable (such as year of birth) as a salt.
- If privacy is a primary concern, several hardening techniques can be used or combined.
- Currently, we consider a balanced, salted CLK as resilient to all known attacks.
- Fastest linkage method for PPRL of inexact identifiers so far seem to be Multibit trees.

References

Bachteler, Tobias/Jörg Reiher/Rainer Schnell (2013): Similarity Filtering with Multibit Trees for Record Linkage. Nürnberg: German Record Linkage Center. WP-GRLC-2013-01.

Bernstein, D. J. (2008): The Salsa20 Family of Stream Ciphers. In: *New Stream Cipher Designs*. Hrsg. von M. Robshaw/O. Billet. Berlin: Springer: 84–97.

Brown, Adrian/Christian Borgs/Sean Randall/Rainer Schnell (2016): *High quality linkage using multibit trees for privacy-preserving blocking*. International Population Data Linkage Conference (IPDLN2016): 24.08-26.08.2016; Swansea.

Christen, Peter/Rainer Schnell/Dinusha Vatsalan/Thilina Ranbaduge (2017): Efficient Cryptanalysis of Bloom Filters for Privacy-Preserving Record Linkage. In: *PAKDD, May 23-26, 2017, South Korea*. Cham: Springer: 628–640.

References

Kristensen, T. G./J. Nielsen/C. N. Pedersen (2010): A Tree-based Method for the Rapid Screening of Chemical Fingerprints. In: *Algorithms for Molecular Biology* 5 (9).

Randall, Sean/Anna Ferrante/James Boyd/Adrian Brown/James Semmens (2016): Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581? In: *Health Information Management Journal* 45 (2): 71–79.

Schnell, Rainer (2012): *Recent Developments in Privacy Preserving Record Linkage*. MEA Workshop 'Linking Survey and Survey Data': 19.11.-20.11.2012; Berlin, 19.11.2012.

Schnell, Rainer (2014): An efficient Privacy-Preserving Record Linkage Technique for Administrative Data and Censuses. In: *Journal of the International Association for Official Statistics* 30 (3): 263–270.

References

Schnell, Rainer (2015): „Privacy Preserving Record Linkage“. In: *Methodological Developments in Data Linkage*. Hrsg. von Katie Harron/Harvey Goldstein/Chris Dibben. Chichester: Wiley: 201–225.

Schnell, Rainer/Tobias Bachteler/Jörg Reiher (2009): Privacy-preserving record linkage using Bloom filters. In: *BMC Medical Informatics and Decision Making* 9 (41).

Schnell, Rainer/Christian Borgs (2016): Available Methods for Privacy Preserving Record Linkage on Census Scale Data. In: *European Conference on Quality in Official Statistics (Q2016), June 1st*. Madrid.

Schnell, Rainer/Christian Borgs (2018): Protecting Record Linkage Identifiers Using a Language Model for Patient Names. In: *Stud Health Technol Inform.* 253: 91–95.

References

Vatsalan, Dinusha/Peter Christen/Vassilios S. Verykios (2013): A Taxonomy of Privacy-preserving Record Linkage Techniques. In: *Information Systems* 38 (6): 946–969.