

Recent developments in Privacy Preserving Record Linkage

Prof. Dr. Rainer Schnell

University of Duisburg-Essen
German Record Linkage Center

ADLS Conference

Major innovations in the field of administrative data linkage and analysis
Friday 16th November, London

UNIVERSITÄT
DUISBURG
ESSEN

German
RLC

The Record Linkage Problem

- ▶ Record linkage identifies matching record pairs in two separate data files.
- ▶ The record linkage results in a classification of pairs of records as *links* and *non links*.
- ▶ Pairs of records which represent identical observational units are called *match*.

		Reality	
		match	non match
Classification	link	TRUE POSITIVES	FALSE POSITIVES
	non link	FALSE NEGATIVES	TRUE NEGATIVES

Record Linkage Techniques

- ▶ Deterministic Record Linkage
 - ▶ Exact one-to-one character matching on the chosen identifiers.
 - ▶ E.g. surname, date of birth, month of birth, ZIP code
- ▶ Distance-based Record Linkage
 - ▶ Exact match requirement is given up.
 - ▶ String similarity functions may be used.
- ▶ Probabilistic Record Linkage
 - ▶ Current standard method with many variants.
 - ▶ Probabilistic record linkage can also be used with string similarity functions.

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

Deterministic Record Linkage

- ▶ From a technical point of view, linking with unique personal identification numbers is ideal.
- ▶ Advantage: distinct, (nearly) error free
- ▶ In Europe: Belgium, Sweden, Norway, Denmark, Finland
- ▶ Not available in Germany, as well as in many other countries.
- ▶ Instead: Personal identifiers such as name, date of birth or address are used as linkage variables.
- ▶ Disadvantage: Taken individually, the identifiers are not unique and must be used in combination.
- ▶ Main problem: Errors in identifiers. Exact matching therefore yields linking errors. Up to 20% errors in surnames is a common experience.

Record Linkage with String Similarity Functions

- ▶ In general, distance-based and probabilistic record linkage use the similarity of two strings for the identification of record pairs.
- ▶ There are many ways to compute the similarity of two strings.
- ▶ The following discussion will concentrate on n-grams.

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

n-grams

- ▶ *n*-grams are substrings of a string with length of *n*. For example bigrams are *n*-grams with *n* = 2.
- ▶ Strings which share many *n*-grams have a higher *n*-gram-similarity.
- ▶ Using the Dice coefficient, the similarity can be determined as:

$$D_{a,b} = \frac{2h}{(|a| + |b|)}, \quad (1)$$

where *h* is the number of shared *n*-grams and *|a|*, *|b|* is the number of *n*-grams in strings *a*, *b*.

Standard Setting for Statistical Data Analysis Problems

- ▶ Two data holders A and B.
- ▶ A research group wants to link the datasets D_a and D_b as micro datasets.
- ▶ A and B have no common personal identification number.
- ▶ Only the name and demographic variables are available for the use as identifiers.
- ▶ Legal constraints limit the use of unencrypted identifiers.
- ▶ The linkage should be tolerant for errors in the identifiers.
- ▶ Data traffic between the involved parties should be as low as possible.

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

Possible Solutions to the Problem

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

Trustee High organisational demands, requires a trustworthy institution with access to plain text identifiers.

Secure Multi-party Computationally intensive, network access is necessary, typically not apt for the development of a statistical model.

Encrypted Phonetic Codes Many false positives, only limited error-tolerance.

Privacy Preserving Record Linkage Several protocols suggested, but most of them are not applicable for the given problem.

Privacy Preserving Record Linkage

- ▶ Privacy Preserving Record Linkage tries to link micro data without access to unencrypted identifiers.
- ▶ The central problem: How to calculate string similarities between two names without exposing these names?
 - ▶ The first published approach was suggested by Churches & Christen (2004).
 - ▶ This protocol is very inefficient and was therefore never actually used.
 - ▶ Nevertheless, the paper was the starting point for the field of research in Privacy Preserving Record Linkage (PPRL).

Churches, T. & P. Christen, 2004: Some Methods for Blindfolded Record Linkage. BMC Medical Informatics and Decision Making 4 (9). Published Online 28.6.2004.

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

Excellent Recent Review of PPRL

Christen, P./Verykios, V. (2012):
A Tutorial on Privacy-Preserving Record Linkage

Workshop Tutorial for PAKDD-2012
(Pacific-Asia Conference on Knowledge Discovery and Data
Mining, Kuala Lumpur, 29 May–1 June 2012)

Slides (96 pages) are available at:

[http://cs.anu.edu.au/people/Peter.Christen/
pakdd2012-pprl-tutorial.html](http://cs.anu.edu.au/people/Peter.Christen/pakdd2012-pprl-tutorial.html)

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

PPRL with Cryptographic Bloom Filters

- ▶ Schnell et al. (2009) suggested a new method for the calculation of similarity between two encrypted strings for the use in record linkage procedures.
- ▶ The method (SAFELINK) is based on the idea of splitting an identifier into q-grams and hashing the q-gram set with **several different** keyed HMACs (MD5, SHA-1) in a bit vector, a so called Bloom filter (Bloom 1970).
- ▶ Given the Bloom filters, the initial string can not be reconstructed.
- ▶ Only the Bloom filters are used for the linkage.
- ▶ The similarity between two strings is approximated by the Dice-coefficient of their Bloom filters.

Schnell, R. & T. Bachteler & J. Reiher, 2009: Privacy-preserving Record Linkage Using Bloom Filters. BMC Medical Informatics and Decision Making 9 (41).

Bloom, B.H., 1970: Space/Time Trade-offs in Hash Coding with Allowable Errors. Communications of the ACM 13 (7):422-426.

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

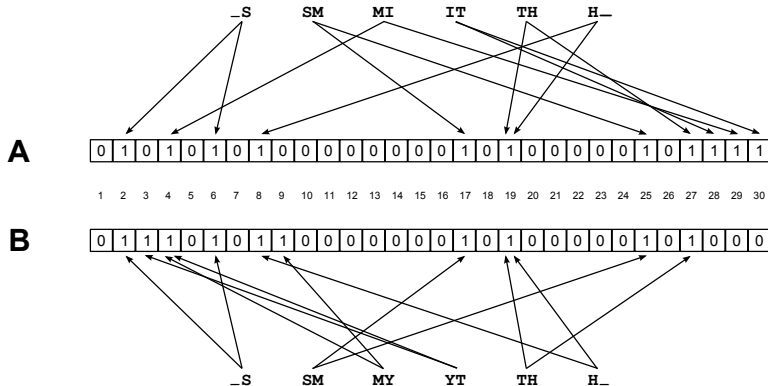
Large Datasets in
Practice

Conclusion

Contact

A SAFELINK example

Two Bloom filters A, B with a length of 30 for "Smith" and "Smyth" and two HMACs.



Recent developments in Privacy Preserving Record Linkage

Rainer Schnell

Record Linkage Techniques

Privacy Preserving Record Linkage SAFELINK CLK Security aspects

Large Datasets in Practice

Conclusion

Contact

The Safelink Procedure: Illustration

- ▶ In both Bloom filters 20 identical bits are set to 1.
- ▶ Overall $30 + 30 = 60$ bits are set to 1.
- ▶ Using the Dice coefficient, the similarity of the two Bloom filters can be determined as $\frac{2 \cdot 20}{60} \approx .67$.
- ▶ The Bloom similarity of two completely different names, such as SMITH and BLACK, is much closer to zero (using the parameters of this example: 0.14).
- ▶ In general, the similarity between two names can be approximated by using the Bloom filters only.
- ▶ The Safelink procedure allows the computation of string similarity with encrypted identifiers.

Simulation

- ▶ Dataset A with 500.000 simulated records.
- ▶ Dataset B with 500.000 simulated records, 125.000 containing errors.
- ▶ Trigrams, Bloom filter with 1000 bits, number of HMACs between 5 and 50.

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

Performance: SAFELINK

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

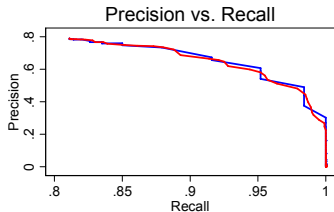
Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

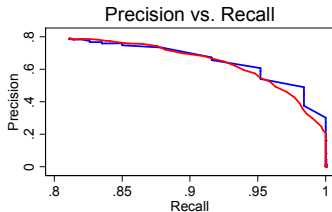
Large Datasets in
Practice

Conclusion

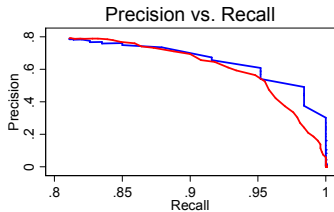
Contact



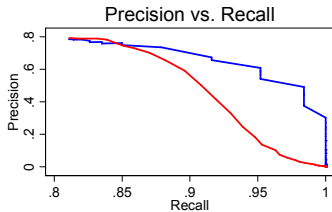
s001, Dice, 1000 bits, 5 Hashfunctions, Bloom: red



s001, Dice, 1000 bits, 25 Hashfunctions, Bloom: red



s001, Dice, 1000 bits, 50 Hashfunctions, Bloom: red



s001, Dice, 1000 bits, 100 Hashfunctions, Bloom: red

Cryptographic Long-term Key (CLK)

- ▶ Due to legal constraints, in some applications in some countries only the use of one single key is allowed.
- ▶ So far, all of the solutions proposed suffer from many false negatives.
- ▶ Schnell et al. (2011) therefore suggested encrypting all identifiers in one single Bloom filter.
- ▶ The results produced by the CLK are only slightly inferior to those of Safelink, but even more secure.

Schnell, R. & T. Bachteler & J. Reiher, 2011: Bloom Filter Based Cryptographic Personal Identification Keys for Longitudinal Research, ASA Spring Methodology Conference at Tillburg University, 19.5.2011.

Schnell, R. & T. Bachteler & J. Reiher, 2011: A Novel Error-tolerant Anonymous Linking Code, German Record Linkage Center, Working Paper Series No. 2.

Wjst, M. 2005: Anonymizing personal identifiers in genetic epidemiologic studies. *Epidemiology*, 16(1):131

Jaquet-Chiffelle, D. et al. 2001: How to protect the rights of patients to medical secrecy in official statistics. *Information Security Bulletin*, 6(8):41-44

Karmel, R. 2005: Data linkage protocols using a statistical linkage key. Technical report, Canberra: AIHW.

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

Example for the Construction of the CLK

1. A Bloom filter of length 1000 is set to 0.
2. First name is split into bigrams and stored in the Bloom filter using 10 HMACs with Key K_1 .
3. Last name is split into bigrams and stored in the Bloom filter using 10 HMACs with Key K_2 .
4. Day of birth is split into bigrams and stored in the Bloom filter using 10 HMACs with Key K_3 .
5. Month of birth is split into bigrams and stored in the Bloom filter using 10 HMACs with Key K_4 .
6. Year of birth is split into bigrams and stored in the Bloom filter using 10 HMACs with Key K_5 .
7. Sex is stored in the Bloom filter using 10 HMACs with Key K_6 .

Simulation of the CLK

- ▶ Forename and surname from a German phone book, simulated date of birth according to a large administrative data set.
- ▶ Test data set A consisting of 2.500 records.
- ▶ Errors were intentionally introduced in 2.000 of these records. These were then stored in a second data file B and supplemented with 8.000 new records.
- ▶ This results in 25 million pairs, with 2.000 being true matches.

Method	FN	FP	TP	% TP
Basic ALC	735	0	1,265	65.0
Swiss ALC	479	0	1,521	78.2
Encrypted SLK	420	0	1,580	81.2
CLK	47	50	1,953	100.4
Plain Text	55	22	1,945	

Further Reduction of False Positives

- ▶ The previous table is based on only three identifiers.
- ▶ If more identifiers are available (name at birth, place of birth, depending on the application: place of residence, medical birth parameters), the number of false positives can be reduced.
- ▶ If the number of hash functions used depends on the quality of the identifier, the number of FPs can be reduced further.
- ▶ Finally, the number of FPs can be minimized by modifying the threshold for classification.

Further Improvements on CLK variants

- ▶ We have tried to improve the performance of the basic CLK.
- ▶ For numerical identifiers, we now prefer a special type of *q - grams*.
- ▶ After many simulations, we now weight the identifiers proportionally to their estimated entropy.
- ▶ A number of methods for handling large data sets with CLKs have been tested.
- ▶ By using one of these techniques, we are now able to link 200.000 records * 200.000 records in 16 minutes without the use of additional techniques. Pairs of 1 Million records can be linked in 5h, pairs of 2 million records in 25h.
- ▶ Beyond that, either special hardware or traditional blocking have to be used.

Security aspects of SAFELINK and CLK

- ▶ Until now three papers on the security of these procedures have been published:
 - ▶ Kuzu, M. et al., 2011: A Constraint Satisfaction Cryptanalysis of Bloom Filters in Private Record Linkage. S. 226-245 in S. Fischer-Hübner & N. Hopper (eds.), Privacy Enhancing Technologies. Berlin: Springer.
 - ▶ Durham, E. et al., 2012: Quantifying the Correctness, Computational Complexity, and Security of Privacy-preserving String Comparators for Record Linkage. Information Fusion 13(4): 245-259.
 - ▶ Kuzu, M. et al., 2012: A Practical Approach to Achieve Private Medical Record Linkage in Light of Public Resources. Journal of the American Medical Informatics Association. Published Online First 30 July 2012.

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

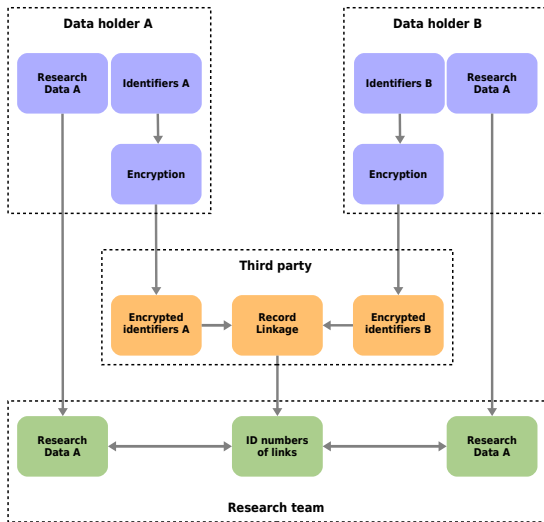
Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

General Setting: RL with a third party



Recent developments in Privacy Preserving Record Linkage

Rainer Schnell

Record Linkage Techniques

Privacy Preserving Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in Practice

Conclusion

Contact

Kuzu 2011, CSP-Experiment 1

- ▶ Kuzu (2011): SAFELINK, $k=15$, $m=500$, $q=2$, $n=20.000$ first name, 3500 unique, start=400 most common first names.
- ▶ The attack is done by a Constraint Satisfaction Problem (CSP) Solver.
- ▶ With a running time of several days, 11% of the first names can be assigned correctly, though wrong assignments were also being observed.
- ▶ Kuzu et al. (2011): „Parameters of the BFE protocol can be configured to make it relatively resilient to the proposed attack without significant reduction in record linkage performance.“

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

Kuzu 2012, CSP-Experiment 2

- ▶ Kuzu (2012): CLK-version without collision, $q=2$, $m=500$, 20.264 unique forenames, 30.217 unique surnames, 129 unique ZIP codes, 131 unique cities, k not reported.
- ▶ After reduction to the most common 20 surnames, 4 correctly identified surnames, meaning 16 false assignments.
- ▶ Kuzu (2012): „(...) when patient identifiers are not a proper random sample of a resource available to an attacker (eg, voter list), cryptanalysis is less likely to succeed.“
- ▶ Kuzu (2012): „Performance of cryptanalysis against BFEs based on patient data is significantly lower than theoretical estimates. The proposed countermeasure [CLK, RS] makes BFE's resistant to known attacks.“

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

Additional Security Measures

- ▶ Insertion of dummy-records, which do not represent actual cases, before sending the set of Bloom filters to the trustee (Fake-Injections, Karakasidis et al. 2012).
- ▶ Insertion of random bits into the Bloom filters (Schnell et al. 2011). This barely affects the calculation of similarities.
- ▶ Usage of CLKs instead of several Bloom filters.
- ▶ Limit the of number of bits per identifier in a CLK (Durham 2012). Reduces the accuracy, but makes a CSP-attack harder.

Karakasidis, A. & V. S. Verykios & P. Christen, 2012: Fake Injection Strategies for Private Phonetic Matching. S. 9-24 in J. Garcia-Alfaro et al. (Eds.), Data Privacy Management and Autonomous Spontaneous Security. Berlin: Springer.

Durham, E.A., 2012: A Framework for Accurate, Efficient Private Record Linkage, Dissertation, Vanderbilt University.

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

Applications for Large Data Sets in Practice

- ▶ Santos et al. (2011) are using our method for the record linkage of Brazilian health records on remote military owned computers.
- ▶ The Swiss Cohort (University of Bern) research group has adapted our method for linkage applications in Switzerland (www.swissnationalcohort.ch).

Santos, L. et al., 2011: Peso ao nascer entre crianças de famílias de baixa renda beneficiárias e não beneficiárias do Programa Bolsa Família da Região Nordeste. S. 271-293 in: Ministério da Saúde (Eds.), Saúde Brasil 2010. Brasília.

Kuehni, C. et al., 2011: Cohort Profile: the Swiss Childhood Cancer Survivor Study. International Journal of Epidemiology.

Recent
developments in
Privacy
Preserving
Record Linkage

Rainer Schnell

Record Linkage
Techniques

Privacy
Preserving
Record Linkage
SAFELINK
CLK
Security aspects

Large Datasets in
Practice

Conclusion

Contact

Conclusion

- ▶ PPRL allows linking and blocking of data sets without exposing the identifiers.
- ▶ The requirements even for a limited attack on Safelink and CLK are typically not given in practice.
- ▶ The performance of these methods regarding precision and recall seem to exceed all published alternatives like phonetic codes or embeddings.
- ▶ Using Safelink or CLK, error tolerant privacy preserving record linkage for large files is possible in practice.

Contact: www.record-linkage.de

The screenshot shows the homepage of the German Record Linkage Center (GermanRLC). At the top left is the logo with the text "German RLC" and a green circle. Below the logo is a navigation menu with buttons for "Home", "RL Resources", "Services", "Research", "Cooperations", "Projects", "Publications", "Downloads", and "Contact". The "Home" button is selected. The main content area has a heading "German Record Linkage Center" followed by a paragraph describing the center's mission and services. Below this is a paragraph about funding from the German Research Foundation. A "Directors:" section lists two individuals. At the bottom, there are logos for the University of Duisburg-Essen, the Research Data Centre (FDZ), and the German Research Foundation (DFG).

German RLC

Home RL Resources Services Research Cooperations Projects Publications Downloads Contact

Home

German Record Linkage Center

The German Record Linkage Center (GermanRLC) was established in 2011 to promote research on record linkage and to facilitate practical applications in Germany. The Center will provide several [services](#) related to record linkage applications as well as conduct [research](#) on central topics of the field. The services of the GermanRLC are open to all academic disciplines.

The German Research Foundation funds the Center within the funding programme 'Scientific Library Services and Information Systems'. A summary of the grant proposal can be found [here](#).

Directors:

- Prof. Dr. Rainer Schnell, University of Duisburg-Essen
- Stefan Bender, Research Data Centre of the Federal Employment Agency at the Institute for Employment Research.

The GermanRLC is a joint project of:

UNIVERSITÄT
DUISBURG
ESSEN

Forschungsdatenzentrum
der Bundesagentur für Arbeit
zur Analyse der Arbeitsmarkt-
und Beschäftigung

FDZ

The GermanRLC is funded by:

DFG

rainer.schnell@uni-due.de