# Efficient private record linkage of very large datasets

Rainer Schnell

Universität Duisburg-Essen & German Record Linkage Center

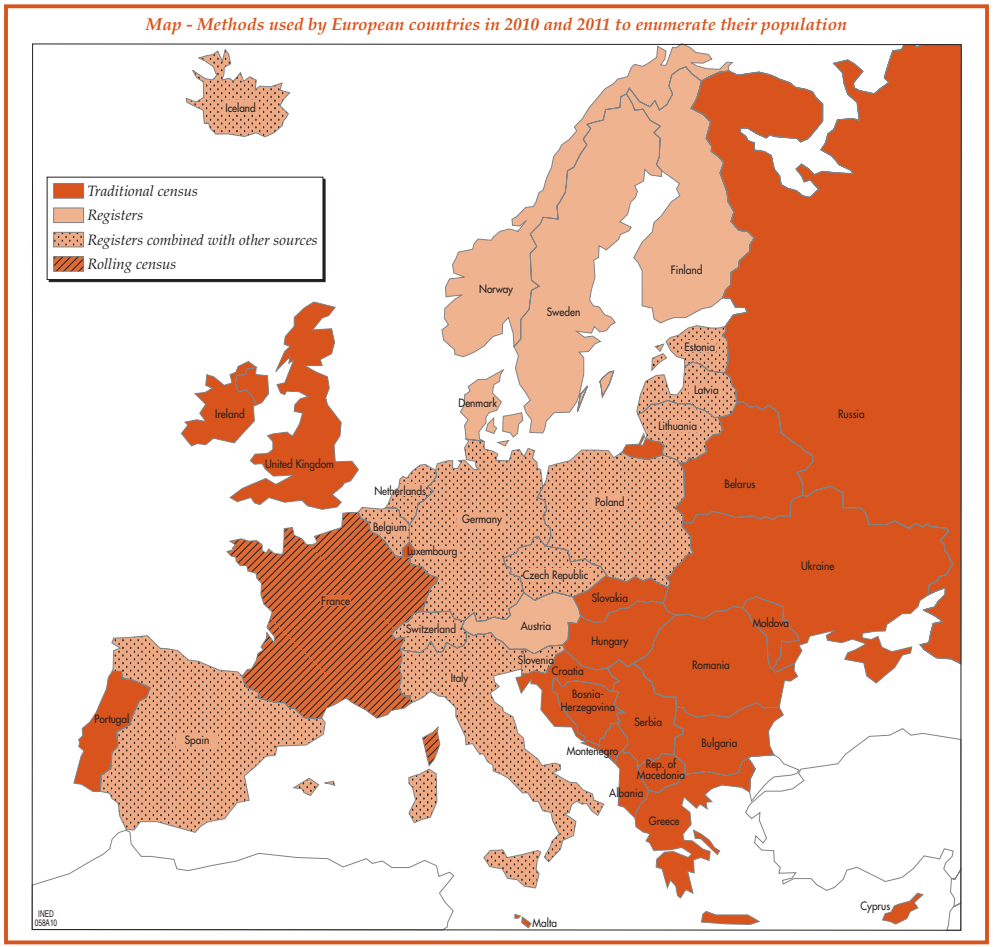


59th World Statistics Congress
Hongkong
29. August 2013

# Introduction

- Linking administrative data for research purposes is increasingly common.
- For many applications, these databases contain millions of records.
- Examples from official statistics are cancer registries and census operations.

Example: Census 2010/11: 21 of 40 European countries did a traditional census; 19 used registers or mixes




Valente,P. (2010): Census taking in Europe: how are populations counted in 2010? in: Population & Societies, 467, p. 2

# Record-Linkage

- Technically, linking with a universally available unique personal identification number (PIDs) is ideal.
- Such PIDs are available in Europe for example, for Denmark, Finland, Norway and Sweden. Under such conditions, linking data bases is technically trivial.
- In most other countries, personal identifiers like names or date of birth have to be used.
- Identifiers are not stable and are recorded with errors: Winkler (2009:362) reports that 25% of true matches in a census operation would have been missed by exact matching.

---

Winkler, W. E. (2009). Record linkage. In: Pfeffermann, D./Rao, C. (eds.): Handbook of Statistics Vol. 29A, p.351-380. Elsevier: Amsterdam.

# Methods for Record-Linkage

- Many different techniques for record-linkage (RL) have been suggested (Christen 2012).
- In official statistics, most often the Fellegi-Sunter-RL model (Herzog/Scheuren/Winkler 2007) is used.
- Most RL methods require unencrypted identifiers.
- Due to privacy concerns, there is an increasing pressure to use encrypted identifiers for record-linkage.
- This field is called private record linkage or privacy preserving record linkage (PPRL)

Christen, P. (2012): Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin.

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

# Approaches to PPRL

**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:** Only limited error-tolerance.

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

---

Vatsalan, D./Christen, P./Verykios, V. S. (2013): A taxonomy of privacy-preserving record linkage techniques; in: Information Systems, 38(6), p.946-969.

# 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 binary vector (a Bloom filter).
- Given the resulting binary vectors , the initial string can not be reconstructed.
- Only the binary vectors are used for the linkage.
- The similarity between two strings is approximated by the Dice-coefficient of the binary vectors.

---

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

# Examples for applications of Bloom-Filter-PPRL

- For real applications, Safelink is used with larger Bloomfilters (500 to 1000 bits) and 15–25 hash functions.
- Safelink is applied to each identifier separately.
- Therefore, Safelink can be implemented in standard RL software like GLINK.

Switzerland Kuehni, C. E. et al. (2012): Cohort profile: the Swiss Childhood Cancer Survivor Study. International Journal of Epidemiology, 41(6), p.1553-1564.

Brasilia Santos, L. et al., 2011: Peso ao nascer entre criancas de familias de baixa renda beneficiarias e nao beneficiarias do Programa Bolsa Familiada Regia o Nordeste. S. 271-293 in: Ministerio da Saude (Eds.), Saude Brasil 2010. Brasilia.

# 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.

# Basic set of identifiers

- To link persons across time and databases, a set of common identifiers is needed.
- In general, for most administrative databases a basic set of identifiers (BSID) is available:
  - first name
  - surname (at birth)
  - sex
  - date of birth
  - country of birth
  - place of birth

# Building Cryptographic long Term Keys (CLKs)

1. Standardization of the identifiers (uppercase, transforming special characters, removing titles, resolving double-names).
2. Constructing a set of unique n-grams of each identifier.
3. Encrypting each unique set of identifiers with a different number of hash-functions and a different password.
4. Mapping all hash-functions to the same Bloom filter.

# Problems of linking large databases with CLKs

- Finding similar CLKs is the problem of finding nearest neighbors in a high dimensional binary space.
- For a census: >100 million candidates with >500 binary variables have to be compared.
- Comparing all pairs of CLKs is impossible.
- Methods for reducing the number of comparisons are called „similarity filtering“ or „blocking methods“.
- A large number of blocking methods have been suggested in the literature (Christen 2012).
- The options for large scale high-dimensional binary data are very limited.

---

Christen, P. (2012): A survey of indexing techniques for scalable record linkage and deduplication. IEEE Transactions on Knowledge and Data Engineering, 24 (9), p.1537-1555.

# Blocking methods for CLKs

The current options for blocking large scale high-dimensional binary data are limited.

- Obvious candidates for blocking CLKs are:
  1. External blocking
  2. Sorted neighborhood
  3. Canopy clustering
  4. LSH-variants like bit-sampling.
- In our preliminary simulations LSH performed badly. The performance of LSH seem to decline with increasing file size.
- Our simulations reported here are restricted to 1-3.

# Similarity filtering with Multibit Trees

- Since we were not satisfied with the current approaches, we searched for new algorithms.
- Early this year, we suggested a new blocking method for record linkage: The use of Multibit Trees for similarity filtering (Bachteler/Reiher/Schnell 2013).
- In our working paper, we described the application of Multibit Trees for record linkage in general (by transforming all identifiers in a standard record linkage problem ($q$-gram-blocking) to one large bit array and applying the Multibit Tree to the CLKs.)
- Here, I will describe the method for blocking within a privacy preserving record linkage context.

---

Bachteler,T./Reiher,J./Schnell,R. (2013): Similarity Filtering with Multibit Trees for Record Linkage, German Record Linkage Center, Working Paper 5.3.2013

# Multibit Trees

- Kristensen/Nielsen/Pedersen (2010) introduced Multibit Trees to search databases of structural information about chemical molecules.
- The structural information on molecules is stored in bit-arrays.
- Therefore, the data structure is identical with a CLK.

---

Kristensen, T. G./Nielsen, J./Pedersen, C. N. S. (2010): A tree-based method for the rapid screening of chemical fingerprints, Algorithms for Molecular Biology, 5, 9

# Searching with Multibit Trees

- We search for all records above a similarity threshold $t$ for the Jaccard Similarity

$$S_j(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cap \mathbf{B}}{\mathbf{A} \cup \mathbf{B}} \tag{1}$$

  where $\mathbf{A} \cap \mathbf{B}$ is the number bits set to 1 in both arrays and $\mathbf{A} \cup \mathbf{B}$ is the number of bits set to 1 in $\mathbf{A}$ or $\mathbf{B}$.
- Multibit Trees work in three steps.

# 1. Partion step

The database is partitioned into groups:

- Vectors of the larger file are grouped into bins.
- Bins are formed by the $|\mathbf{B}|$, the number of bits set to 1 in $\mathbf{B}$.
- Therefore, bins satisfying

$$|\mathbf{B}| \leq \mathbf{t}\,|\mathbf{A}| \quad \text{or} \quad \mathbf{t}|\mathbf{B}| \geq |\mathbf{A}| \tag{2}$$

can be ignored in the searching step, because $\frac{min(|\mathbf{B}|,|\mathbf{A}|)}{max(|\mathbf{B}|,|\mathbf{A}|)}$ constitutes an upper bound of $S_J(\mathbf{A},\mathbf{B})$.

# 2. Treebuilding Step

Within each group an actual Multibit Tree is built:

- Vectors of equal size are stored in a binary tree structure for each bin.
- Two additional lists are stored at each node:
  - List $O$ contains all bit positions with constant value 0.
  - List $I$ contains all bit positions with constant value 1 in all remaining vectors below that node.

# 3. Search Step

Query vector **A** is searched in three phases

1. All bins satisfying equation 2 are eliminated
2. At each node of the remaining trees, the lists $O$ and $I$ allow the computation of an upper bound of the similarity. If this bound falls below the threshold $t$, all nodes below can be eliminated.
3. For all remaining nodes, the similarity between **A** and **B** is computed. If $S_j(\mathbf{A}, \mathbf{B_i}) \geq \mathbf{t}$, vector **A** is in the result set.

# Comparison of CLK blocking techniques

- Bachteler/Schnell/Reiher (2013) inter alia compared
  - Canopy Clustering
  - Sorted Neighborhood
  - Standard Blocking

  with Multibit Trees of a CLK data structure on different databases.
- In most situations, Multibit Trees on CLKs outperformed the methods performing best in other comparison studies. For details, we refer to the Working Paper.

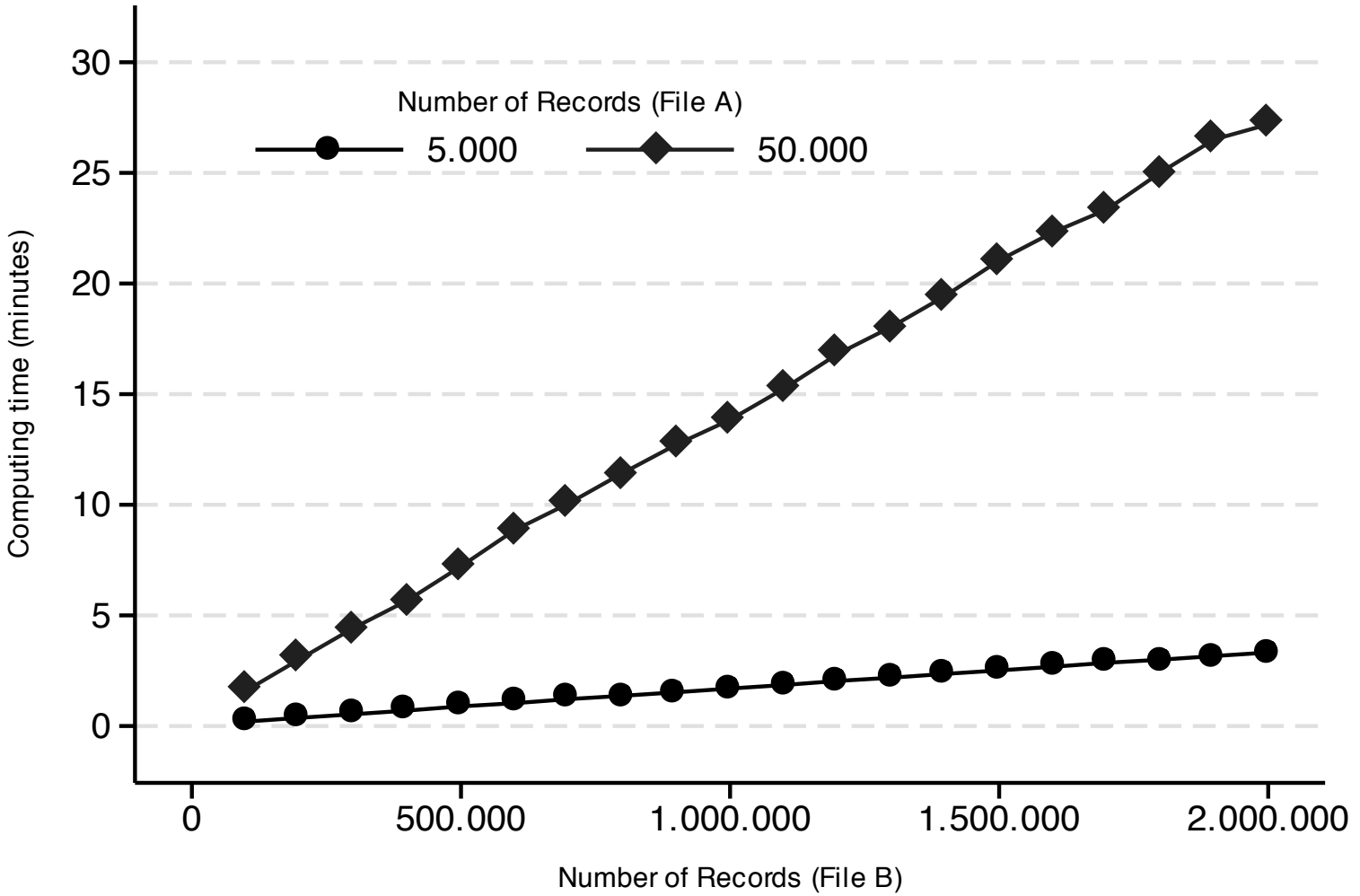
---

Bachteler,T./Reiher,J./Schnell,R. (2013): Similarity Filtering with Multibit Trees for Record Linkage, German Record Linkage Center, Working Paper 5.3.2013

## Scaling Properties for Linking CLKs with Multibit Trees

- In a second simulation, we used CLKs with 1320 bits and $k = 15$ hash functions per identificator field for each identifier in the set: Name, Surname, Sex, Day/Month/Year of Birth, Place of Birth, Country of Birth.
- We simulated with 5.000 and 50.000 records for the small file (A) and 100.000 to 2.000.000 records for the larger file (B).
- File A was a random sample of file B, but 10% of records in A were simulated with errors.

# Computing time for finding 5.000 and 50.000 CLKs




Similarity threshold 0.95.

# Results

**With exact matches** Even with 50.000 records in A and 2 million records in B, the match needed only slightly more than a minute on a standard server (16gb RAM, 2 * Quadcore CPU, 2.8ghz, Ubuntu 10.0.4).

**Similarity matches** With a similarity threshold of .95 the 5.000 records are still matched in less than 4 minutes, but for matching 50.000 records to 2 million records about 27 minutes are needed.

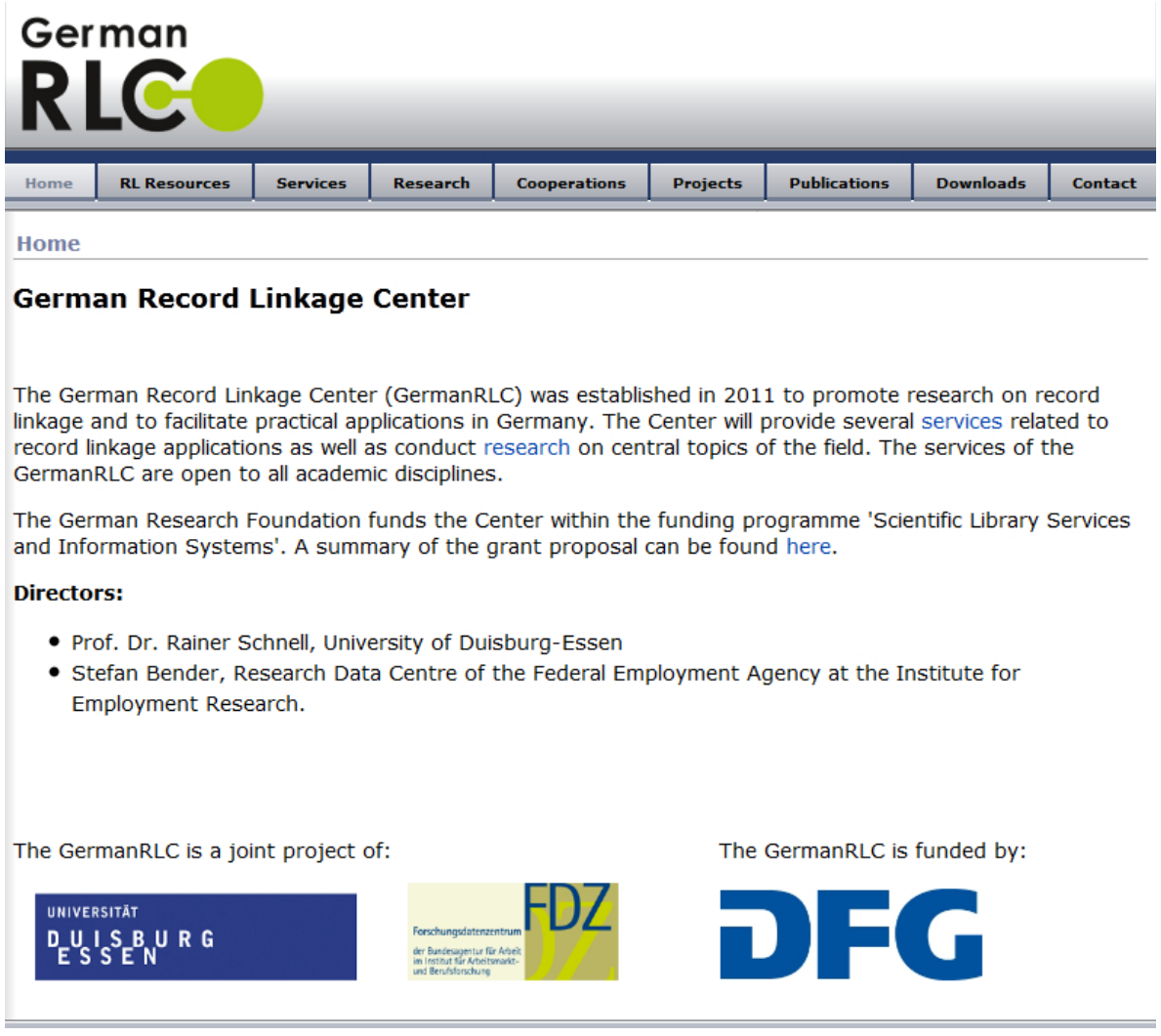
# Linking Larger Files

- The computing time seems to increase linear with the number of records in both files within the simulated range.
- The computing time will increase with decreasing similarity thresholds, since the number of pairwise comparisons will increase.
- However, the exceptional performance of Multibit Trees for this task even for large files is surprising.
- The running time for matching two files with 1 million CLKs each is less than 5 minutes for an exact match and for a similarity match (.95) less than 5:40 hours.
- Recently, we implemented Multibit trees (using C++) within R. This library performed 2-3 times faster.

# Conclusion

- By the application of Multibit Trees, private Record-Linkage using Bloom filters can be done on standard hardware with files up to 1 million records each within 2 hours.
- However, for European census size operations, we need to compare files with 100 million records each.
- Therefore, currently, we need to combine different techniques for such operations.
- An obvious solution is Standard Blocking with encrypted blocking variables, for example birth cohorts.
- Exploring such combinations is the current focus of our research.

# Contact: www.record-linkage.de




Rainer.Schnell@uni-due.de