Available Methods for Privacy Preserving Record Linkage on Census Scale Data

Rainer Schnell

Centre for Comparative Social Surveys
City University London
London, United Kingdom
Email: Rainer.Schnell@city.ac.uk

Christian Borgs

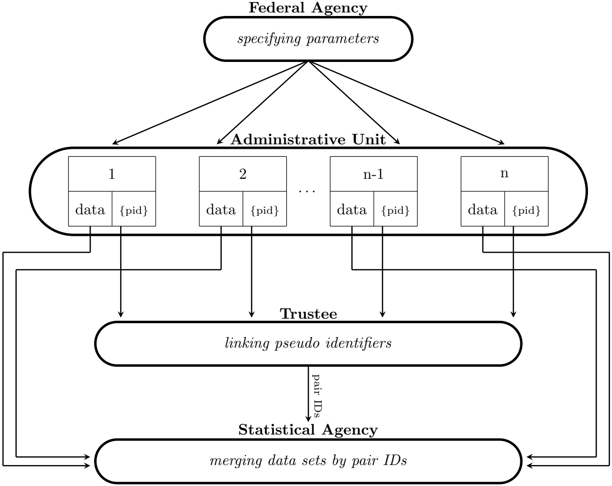
German Record Linkage Center
University of Duisburg-Essen
Duisburg, Germany
Email: christian.borgs@uni-due.de

European Conference on Quality in Official Statistics (Q2016)
Madrid, 1 June 2016

Background

- Register-based censuses are becoming more and more common in Europe (Valente 2010).
- In countries where unique personal identifiers (PIDs) are not available, linking real-world entities across administrative data requires the use of identifiers such as names or birth dates (Abbott et al. 2016; Office for National Statistics 2013).
- Since these identifiers are prone to error, they can lead to non-linked pairs, which may imply biased estimates (Harron et al. 2014; Bohensky 2016).
- If the jurisdiction does not allow the use of unencrypted identifiers for record linkage, Privacy Preserving Record Linkage (PPRL) methods have to be used.

Intended PPRL Setting



Dimensions of PPRL implementations in practice

1. Linkage quality (precision and recall)
2. Security against cryptographic attacks
3. Scalability (able to handle large datasets)

Criteria for linkage quality

Precision is defined as the number of correctly classified pairs (true positive classifications tp) divided by the number of all classified pairs (tp and false positives fp):

$$\text{Precision} = \frac{tp}{tp + fp}$$

Recall is defined as the number of true positive matches divided by the number of factual pairs, including pairs falsely classified as non-matches (false negatives fn) by the linkage algorithm:

$$\text{Recall} = \frac{tp}{tp + fn}$$

Finally, F-score is defined as the harmonic mean of recall and precision:

$$\text{F-score} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

PPRL approaches to census scale data

- Phonetic codes, subsamples of name elements and Bloom Filters have been used on census scale data.
- Since phonetic codes and subsamples of name elements suffer from their inability to account for small variance in the identifiers, their performance on real data is often disappointing.
- Therefore, Bloom Filter approaches have attracted interest.
- In general, higher recall compared to phonetic codes and comparable precision can be attained. Compared to unencrypted identifiers, performance of all PPRL approaches is usually reduced.
- However, Bloom Filter approaches have been used successfully in practical applications (Randall et al. 2013; Schmidlin et al. 2015; Vatsalan/Christen 2016; Schnell et al. 2014).
- Because other PPRL techniques (Vatsalan et al. 2013) require repeated internet access, don't scale well or demonstrate inferior linkage quality, we will concentrate on Bloom Filters here.

Bloom filter encryption

- We (Schnell et al. 2009) suggested the use of Bloom filters (Bloom 1970) to encrypt identifiers for PPRL.
- Initially, all Bloom filters are bit arrays length L initialised to 0.
- To encrypt a set of identifiers into separate Bloom filters, each identifier is split into a set of bigrams (for string-based identifiers) or unigrams (for numeric identifiers).
- Each n-gram is encoded by the sum of the numeric representation of MD5 and SHA1 hashes.
- This construction of hash-functions is called "double-hashing" by Kirsch/Mitzenmacher (2006).

Cryptographic Long-term Keys (CLKs)

- Basic Bloom Filters as described here so far, can be attacked by simple frequency attacks (Durham 2012).
- Therefore, we suggested using "Cryptographic Long-term Keys" (*CLKs* (Schnell et al. 2011)).
- A CLK is a common bit array for all separate Bloom filters.
- CLKs are more difficult to attack by frequency attacks than Bloom Filters.
- Further computational measures (Schnell 2016) can be used to protect Bloom filters against frequency attacks, for example 'salting'.
- Salting is simply the use of different hash-functions for an identifier given the value of a different identifier.

Example: Hardening Bloom filters with random hashing

- We (Niedermeyer et al. 2014) showed that the double hashing scheme is vulnerable to cryptographic attacks on bit patterns resulting from bigrams.
- We also showed that this kind of attack can be prevented in total by replacing the double-hashing scheme with random hashing.
- Random hashing is implemented using a pseudo-random number generator (Stallings 2014) to generate a sequence X with the length k for each n-gram:

$$X_{n+1} = (a * X_n + c) \bmod L.$$

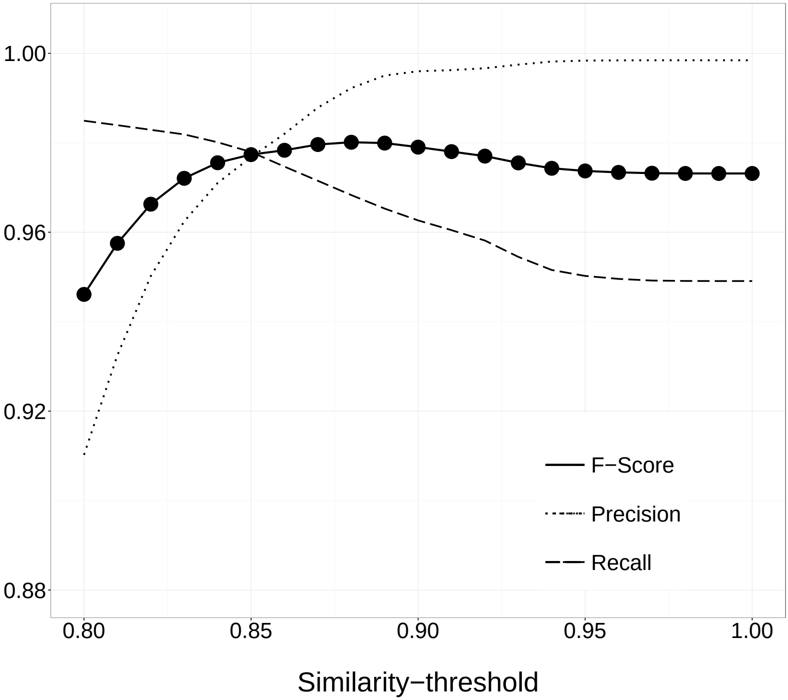
- There is no known attack on Bloom filters using random hashing (Schnell et al. 2016).

Scalability: Linking large databases with CLKs

- Calculating similarity is computationally expensive.
- The number of pairwise comparisons for census-scale data have to be reduced.
- Therefore, special techniques (blocking) for finding nearest neighbours have to be used.
- In practical use for PPRL are:
  - Canopy Clustering (CC, McCallum et al. (2000)) and
  - Sorted nearest neighbourhood blocking (SNN, Hernandez/Stolfo (1998).
- We (Bachteler et al. 2013) suggested the use of Multibit trees (Kristensen et al. 2010).

Empirical applications of Multibit trees

- Using two data sets with more than 6 million records each, 97% of all true matches were found, while keeping the amount of false positives under 5% (Brown et al. 2016).
- This was achieved *without* any blocking.
  - This requires up to 100 hours for 10 million by 10 million records and 64Gbyte RAM.
  - Smaller blocks (1 Mio by 1 Mio) require about 2 hours.
- However, in general performance of linking CLKs is highly dependent on the parameters of the linkage process.
- An example using German cancer registry data ($n_1 = 138131$, $n_2 = 73184$) is shown in the next slide.


1.00
0.96
0.92
0.88
0.80
0.85
0.90
0.95
1.00
Similarity−threshold
F−Score
Precision
Recall

Conclusions

- Compared to unencrypted identifiers, performance of all PPRL approaches is usually reduced.
- Using optimal parameters for the encoding procedure and similarity thresholds will find most true links despite missing or misspelled names.
- Currently, no attacks against salted random hash CLKs are known.
- Including additional (correct) identifiers will reduce false positive links (for example, carefully preprocessed 'Place of Birth', Schnell/Borgs 2015).
- The performance of Bloom filter-based PPRL strongly depends on the parameters chosen.
- Using birth year as external block, PPRL on a European Census can be done in less than a week.

Ongoing research

- We plan to release an R package this year.
- We are investigating the automatic choice of optimal parameters for Bloom filter-based PPRL.
- Using very recent optimizations, the time required to link large data sets will be roughly reduced by 40%.
- In general, we expect higher precision and recall by using more elaborate preprocessing (Abbott et al. 2016).
- Census applications for RL will require the use of additional information, for example information on relationships among persons (Abbott et al. 2016).
- Using this additional information will make privacy protection even more challenging.

References

Abbott, Owen/Peter Jones/Martin Ralphs (2016): "Large-scale linkage for total populations in official statistics". In: *Methodological Developments in Data Linkage*. Ed. by Katie Harron/Harvey Goldstein/Chris Dibben. Chichester: Wiley: 170–200.

Bachteler, T./J. Reiher/R. Schnell (2013): Similarity Filtering with Multibit Trees for Record Linkage. Working Paper. German Record Linkage Center. WP-GRLC-2013-02.

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

References

Bohensky, Megan (2016): "Bias in data linkage studies". In: *Methodological Developments in Data Linkage*. Ed. by Katie Harron/Harvey Goldstein/Chris Dibben. Chichester: Wiley: 63–82.

Brown, Adrian/Christian Borgs/Sean Randall/Rainer Schnell (2016): *High quality linkage using Multibit Trees for privacy-preserving blocking, IPDLN Conference 2016, accepted presentation.*

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

Durham, Elisabeth Ashley (2012): *A Framework for Accurate, Efficient Private Record Linkage*. Generic.

References

Harron, Katie/Angie Wade/Ruth Gilbert/Berit Muller-Pebody/Harvey Goldstein (2014): Evaluating bias due to data linkage error in electronic healthcare records. In: *BMC Medical Research Methodology* 14 (1): 36.

Hernandez, Mauricio A./Salvatore S. Stolfo (1998): Real-world data is dirty: data cleansing and the merge/purge problem. In: *Data Mining and Knowledge Discovery* 2 (1): 9–37.

Jones, Pete (2015): *Development of pseudonymised matching methods for linking administrative datasets. Presentation for the S4CP3 Conference.*

Kirsch, Adam/Michael Mitzenmacher (2006): Less hashing same performance: building a better Bloom filter. In: *Algorithms-ESA 2006, Proceedings of the 14th Annual European Symposium*. Springer: 456–467.

References

Kristensen, Thomas G./Jesper Nielsen/Christian N.S. Pedersen (2010): A Tree-based Method for the Rapid Screening of Chemical Fingerprints. In: *Algorithms for Molecular Biology* 5 (1): 9–20.

McCallum, A./K. Nigam/L. H. Ungar (2000): Efficient clustering of high-dimensional data sets with application to reference matching. In: *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM: 169–178.

Niedermeyer, Frank/Simone Steinmetzer/Martin Kroll/Rainer Schnell (2014): Cryptanalysis of Basic Bloom Filters Used for Privacy Preserving Record Linkage. In: *Journal of Privacy and Confidentiality* 6 (2): 59–79.

References

Office for National Statistics (2013): Beyond 2011: Matching Anonymous Data. Office for National Statistics. Methods & Policies Report M9.

Randall, Sean M./Anna M. Ferrante/James H. Boyd/Jacqueline K. Bauer/James B. Semmens (2013): Privacy-preserving Record Linkage on Large Real World Datasets. In: *Journal of Biomedical Informatics*: 205–212.

Schmidlin, Kurt/Kerri M. Clough-Gorr/Adrian Spoerri (2015): Privacy Preserving Probabilistic Record Linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. In: *BMC Medical Research Methodology* 16 (1): 46–56.

Schnell, R./M. Thürling/Christian Borgs (2016): “Explorations of Protective Measures against Cryptanalysis of Bloom filter-based Privacy-preserving Record Linkage”. Unpublished.

References

Schnell, Rainer (2016): "Privacy Preserving Record Linkage". In: *Methodological Developments in Data Linkage*. Ed. by Katie Harron/Harvey Goldstein/Chris Dibben. Chichester: Wiley: 201–225.

Schnell, Rainer/Christian Borgs (2015): Building a national perinatal database without the use of unique personal identifiers. In: *2015 IEEE 15th International Conference on Data Mining Workshops*. IEEE: 232–239.

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/Tobias Bachteler/Jörg Reiher (2011): A Novel Error-Tolerant Anonymous Linking Code. Report. German Record Linkage Center. WP-GRLC-2011-02.

References

Schnell, Rainer/Anke Richter/Christian Borgs (2014): Performance of different methods for privacy preserving record linkage with large scale medical data sets. In: *2014 International Health Data Linkage Conference: 28.04.-30.04.2014, Vancouver*.

Stallings, William (2014): *Cryptography and Network Security: Principles and Practice*. 6th ed. New Jersey: Pearson.

Valente, Paolo (2010): Census taking in europe: how are populations counted in 2010? In: *Population and Societies* 467: 1–4.

Vatsalan, Dinusha/Peter Christen (2016): Privacy-preserving matching of similar patients. In: *Journal of Biomedical Informatics* 59: 285–298.

References

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