German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2016-03 | DECEMBER 22, 2016

XOR-Folding for Bloom Filter-based Encryptions for Privacy-preserving Record Linkage

Rainer Schnell | Christian Borgs

XOR-Folding for hardening Bloom Filter-based Encryptions for Privacy-preserving Record Linkage

Rainer SCHNELL [a] and Christian BORGS [b]
[a] *City, University of London, London, United Kingdom*
[b] *University of Duisburg-Essen, Duisburg, Germany*

Abstract. In chemo-informatics, XOR-folding of fingerprint bit vectors is a prominent technique for speeding up searches in databases [7]. For Privacy-preserving Record Linkage (PPRL), folding techniques have been applied using Multibit Trees for rapid linking [4,9].

In this paper, we propose using this technique on Bloom Filters as used in PPRL [2], to prevent bit pattern attacks [11]. We test the method in terms of linkage quality and decoding rate as defined in [12,11] using real-word data from a German phone book.

Our results show no reduction in linkage quality using single XOR folding. However, the folding prevents bit pattern attacks completely. Another advantageous property of the folded Bloom Filters is the reduced length, which leads to a 50% decrease in memory consumption and a reduction in the time required for blocking and linking the data. Therefore, we consider the technique demonstrated in this paper as a promising candidate to be examined in further systematic tests of the cryptographic properties of hardened Bloom Filters for PPRL applications.

Keywords. Entity resolution, PPRL, bit vectors, cryptographic attack.

1. Introduction

Larger bit vectors (with a typical length of 512 to 1024 bits) are widely used in chemometrics [1] as well as in Privacy-preserving Record Linkage (PPRL [2]) to represent information on the characteristics of interest. Efficient searching of data bases containing up to 10^7 records of binary vectors in chemo-metrics and up to 10^8 records in PPRL for health informatics [3] is an important research topic in many fields [4,5,6]. One of the available techniques for speeding up searches is bit vector folding [7].

In chemo-informatics, the folding of a bit vector is used to speed up searching chemical data bases [1]. For the intended application, privacy considerations are irrelevant. However, the same idea of folding can be used in the context of Privacy-preserving Record Linkage [8], especially for rapid linking of binary data [9]. We here propose the use of bit vector folding to increase the security of Bloom Filters [10]. The folded bit vectors will be more resilient against bit pattern attacks as described in [11].

We first describe the folding process. Then we will demonstrate the resulting cryptographic properties with respect to the attack described by [11,12]. Finally, linkage quality will be discussed briefly.

2. Vector folding

We denote the original binary vector of length n (typically $500 \leq n \leq 1000$) as $\vec{A}$. The resulting vector $\vec{a}$ is constructed by splitting the original vector $\vec{A}$ in two halves ($\vec{a}_1$ and $\vec{a}_2$) of length $\frac{n}{2}$:

$$\vec{a}_1 = \vec{A}_i,\ i \in \{1, \ldots, \frac{n}{2}\}$$

$$\vec{a}_2 = \vec{A}_i,\ i \in \{\frac{n}{2} + 1, \ldots, n\}$$

Finally, both halves are combined by the XOR operator:

$$\vec{a} = \vec{a}_1 \oplus \vec{a}_2$$

Of course, $\vec{a}$ contains a 1 if, and only if exactly one of the vector halves contains a one at the corresponding bit position. Figure 1 demonstrates the construction.

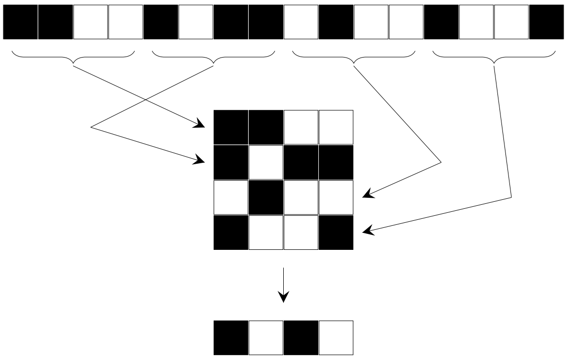

Figure 1. XOR-Folding using $n = 16$ bits set to zero (white) and one (black). The vector is folded four times, resulting in a vector of the length $n = 4$. The idea of the figure is taken from [7, p. 1369].

3. Decoding rates using XOR-folding

To attack standard composite Bloom Filters (CLKs) and folded Bloom Filters, the attack by [12] was used. It uses clear text training data to build a bigram frequency table, which is used to guess the bigrams of Bloom Filters according to their bit pattern frequencies in the attacked data set. Two frequency matrices D and E are created, where D contains the frequencies of the training data bigrams and E is the matrix with the bit pattern frequencies of the Bloom Filter data set. The similarity of the two matrices is minimized using an automated algorithm, which guesses the bigrams corresponding to a bit pattern according to their frequency similarity. If the resulting clear text matches the clear text proposed by the attack, a record is considered to be successfully decoded.

For standard CLKs, the decoding rate was 79%. If dates of birth are included as an identifier, only 49% of the CLKs can be decrypted. However, using the attack on folded Bloom Filters resulted in a decoding rate of zero. The deletion of ones by the XOR-operator changes the bit patterns in a way, which can no longer be re-identified by a systematic search. Therefore, we are tempted to assume that folded Bloom Filters can not be attacked by the original bit pattern attack as implemented by [11,12]. This warrants further research.

4. Linkage quality using XOR-folding

To assess the linkage quality in terms of precision:

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{false positives}}$$

and recall:

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{false negatives}}$$

$n = 100.000$ records were sampled from a German phone book. In a copy of this file, we simulated errors for 20% of all rows. Data corruption types included swapping, deletion, insertion and replacement with probabilities for each error type according to FEBRL [13]. In addition, for 3% of the records, first and last name were interchanged.

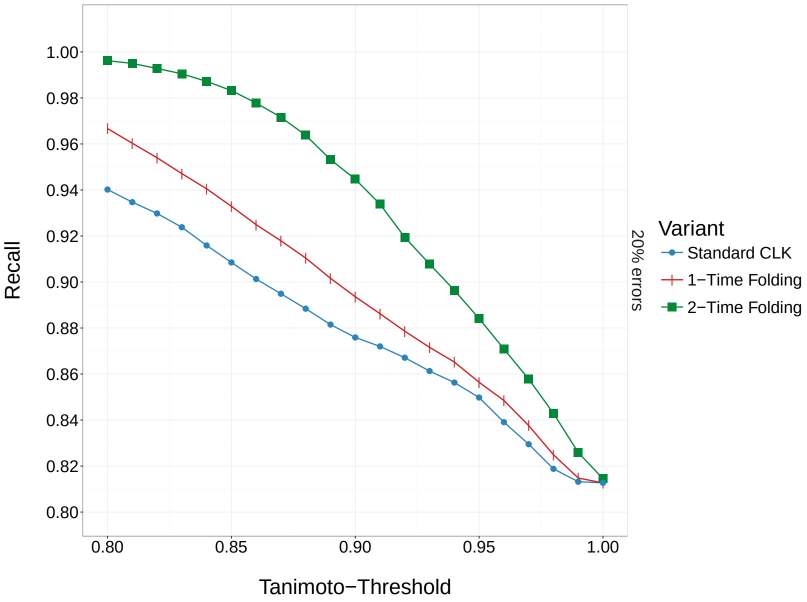

Figure 2. Recall for linking German phone data with 20% errors. XOR-Folding was done once and twice on standard CLKs of the length $n = 1000$ using $k = 20$ hash functions.

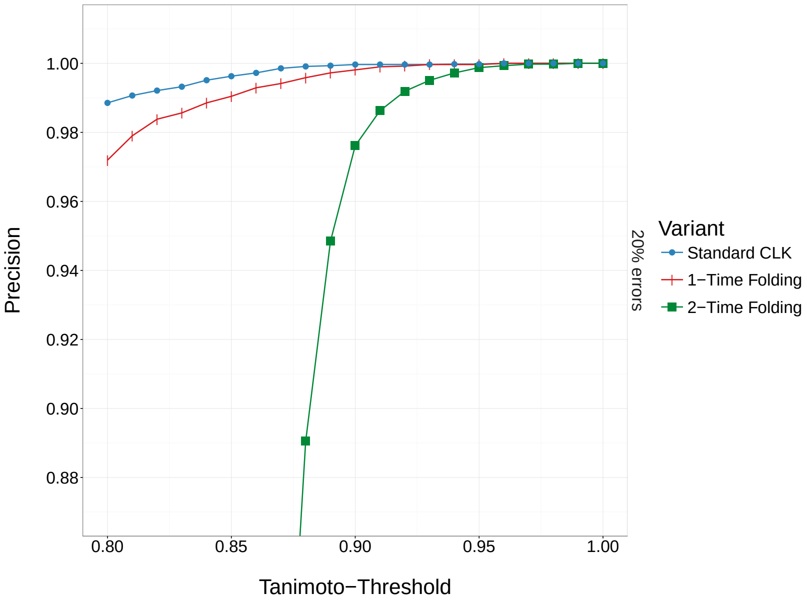

Figure 3. Precision for linking German phone data with 20% errors. XOR-Folding was done once and twice on standard CLKs of the length $n = 1000$ using $k = 20$ hash functions.

Figures 2 and 3 show the impact of folding on the linkage quality in terms of recall and precision. Folding the Bloom Filter once shows no decrease in precision, while a second folding leads to a markedly decreased performance. This reduced precision is due to the fact that more false positives are found when only $\frac{1}{4}$ of the original length is retained. Finally, the recall is stable using any method, even slightly outperforming unfolded Bloom Filters.

5. Systematic evaluation

To better assess the impact on folding on linkage quality, a test was devised using more encryption parameters and data sources. Figures 4, 5 and 6, show recall, precision and f-measure (calculated here as the unweighed mean of precision and recall) for CLKs using $k = 10$ to $k = 30$ hash functions with FEBRL-generated data, NC Voter data and German telephone data (the data source of the first test).

While the recall is consistently better using folding (figure 4), the precision drops considerably (figure 5), leading to a mixed view when averaging both measures (figure 6): Using German telephone data, folding looks to be slightly superior to standard CLKs. However, looking at the other data sources shows virtually no difference in F-Score.

To sum up, folding a Bloom Filter exactly once using the method described above yields virtually no reduction in linkage quality, while completely preventing all currently known attacks.

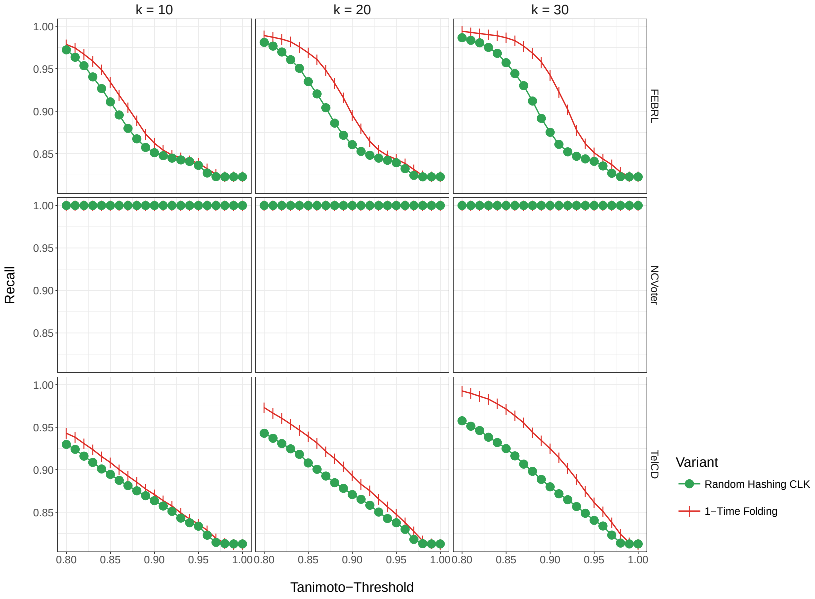

Figure 4. Recall for linking German phone book/FEBRL (20% errors) NC Voter (0% errors) data. XOR-Folding was done once on standard CLKs of the length $n = 1000$ using $k = 10$ to $k = 30$ hash functions.

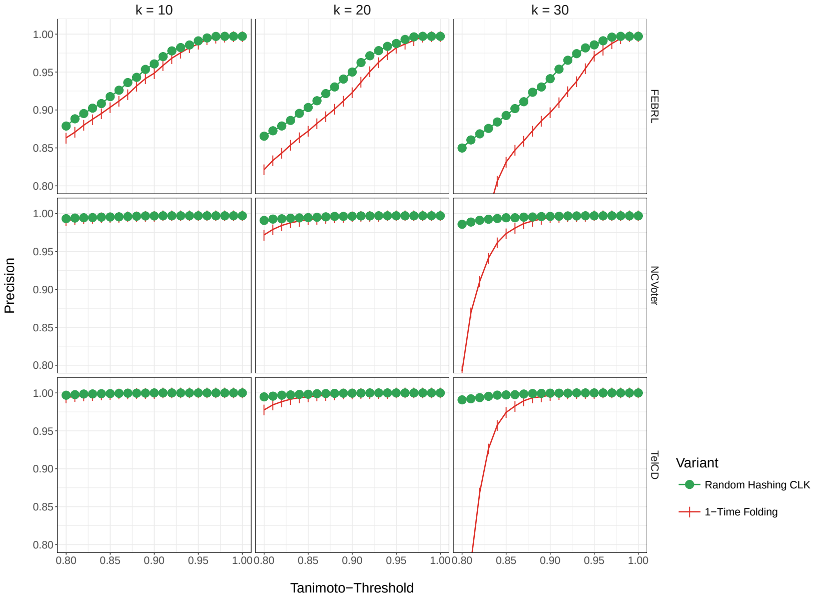

Figure 5. Precision for linking German phone book/FEBRL (20% errors) NC Voter (0% errors) data. XOR–Folding was done once on standard CLKs of the length $n = 1000$ using $k = 10$ to $k = 30$ hash functions.

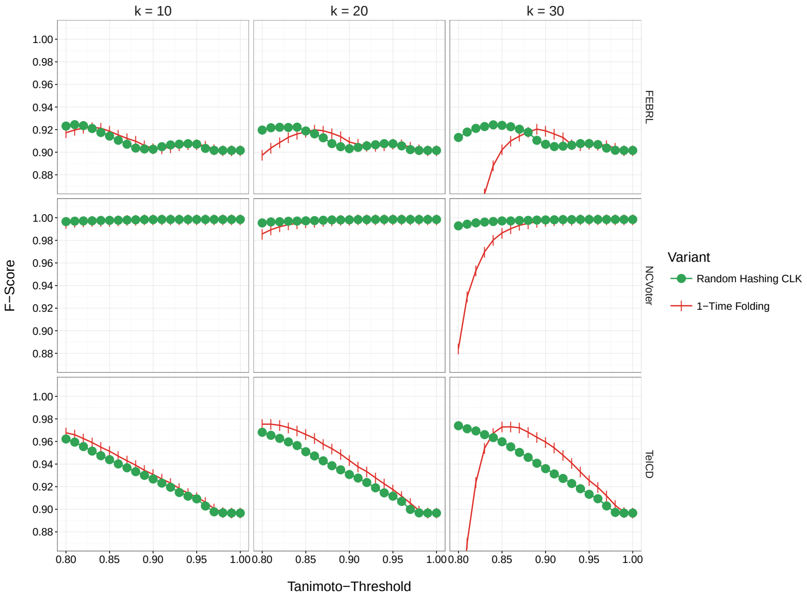

Figure 6. F-Score for linking German phone book/FEBRL (20% errors) NC Voter (0% errors) data. XOR–Folding was done once on standard CLKs of the length $n = 1000$ using $k = 10$ to $k = 30$ hash functions.

6. Discussion

We here suggested a new hardening technique for Bloom Filters for PPRL. Up to now, only a few other techniques have been suggested. For example, replacing double hashing with random hashing [11], salting with stable identifiers [11], combining several Bloom Filters into one [14], sampling and shuffling bits from several Bloom Filters [15], rehashing [8], randomized-response bit flipping [16] and balancing Bloom Filters [16].

The results from our preliminary experimentation show no signs of a decreased linkage quality in terms of precision and recall using one-time folding on Bloom Filters, while completely preventing the attack proposed by [11]. It has to be noted, that the reduced size of the Bloom Filters after folding also reduces the memory required, while speeding up the linkage, because fewer bits have to be compared.

Therefore, we consider the technique demonstrated in this paper as a promising candidate to be examined in further systematic tests of the cryptographic properties of hardened Bloom Filters for PPRL applications.

References

[1] Chen, J., S. J. Swamidass, Y. Dou, J. Bruand, P. Baldi: A public database of small molecules and related chemoinformatics resources. In: *Bioinformatics*, Vol. 21, No. 22 (2005), pp. 4133–4139.

[2] Schnell, R., T. Bachteler, J. Reiher: Privacy-preserving record linkage using Bloom filters. In: *BMC Medical Informatics and Decision Making*, Vol. 9, No. 41 (2009).

[3] Randall, S. M., A. M. Ferrante, J. H. Boyd, J. K. Bauer, J. B. Semmens: Privacy-preserving record linkage on large real world datasets. In: *Journal of Biomedical Informatics*, Vol. 50 (2014), pp. 205–212.

[4] Schnell, R.: An Efficient Privacy-Preserving Record Linkage Technique for Administrative Data and Censuses; in: *Statistical Journal of the IAOS* Vol. 30 No. 3 (2014), pp. 263–270.

[5] McCallum, A., K. Nigam, L. H. Ungar: Efficient clustering of highdimensional data sets with application to reference matching. In: Ramakrishnan, R., Stolfo, S., Bayardo, R., and Parsa, I., editors, *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2000), pp. 169–178.

[6] Christen, P. (2012): A survey of indexing techniques for scalable record linkage and deduplication. In: *IEEE Transactions on Knowledge and Data Engineering*. Vol. 24, No. 9 (2012), pp. 1537–1555.

[7] Baldi, P., D. S. Hirschberg, R. J. Nasr: Speeding Up Chemical Database Searches Using a Proximity Filter Based on the Logical Exclusive OR. In: *Journal of Chemical Information and Modeling*, Vol. 48, No. 7 (2008), pp. 1367–1378.

[8] Schnell, R. : Privacy-preserving Record Linkage. In: Harron, K., Goldstein, H. and Dibben, C. (Eds.): *Methodological Developments in Data Linkage*, Chichester: Wiley (2016), pp. 201–225.

[9] Kristensen, T. G., J. Nielsen, C.N.S. Pedersen: A tree-based method for the rapid screening of chemical fingerprints. *Algorithms in Molecular Biology* Vol. 5 (2010), pp. 9–19.

[10] Bloom, H. : Space/time trade-offs in hash coding with allowable errors. In: *Magazine Communications of the ACM*, Vol. 13 (1970), pp. 422–426.

[11] Niedermeyer, F., S. Steinmetzer, M. Kroll, R. Schnell: Cryptanalysis of basic bloom filters used for privacy preserving record linkage. In: *Journal of Privacy and Confidentiality*, Vol. 6, No. 2 (2014), pp. 59–79.

[12] Kroll, M., S. Steinmetzer: Who Is 1011011111...1110110010? Automated Cryptanalysis of Bloom Filter Encryptions of Databases with Several Personal Identifiers. In: *Biomedical Engineering Systems and Technologies*, Vol. 574 (2016), pp. 341–356.

[13] Christen, P.: Febrl – An Open Source Data Cleaning, Deduplication and Record Linkage System with a Graphical User Interface. *Knowledge Discovery and Data Mining Conference (KDD'08)*, August 24–27, 2008, Las Vegas, Nevada, USA.

[14] Schnell, R., T. Bachteler, J. Reiher: A Novel Error-Tolerant Anonymous Linking Code. In: *German Record Linkage Center Working Paper Series* Working Paper No. WP-GRLC-2011-02 (2011).

[15] Durham,E. A.: A Framework for Accurate, Efficient Private Record Linkage. PhD Thesis (2012).

[16] Schnell, R., C. Borgs: Randomized Response and Balanced Bloom Filters for Privacy Preserving Record Linkage. In: *IEEE International Conference on Data Mining (ICDM)* (2016).

IMPRINT

Publisher

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

Editors

Stefan Bender, Rainer Schnell

Template layout

Christine Weidmann

Download

www.record-linkage.de

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