International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I9P218
Volume 69 Issue 9, 146-152, September, 2021

Original Article

An Ensemble Approach for Privacy-Preserving Record Linkage

Vijay Maruti Shelake[1], Narendra M. Shekokar[2]

[1,2]*Department of Computer Engineering, D. J. Sanghvi College of Engineering, Mumbai, Maharashtra, India*

[1]vijay_sakec@yahoo.co.in, [2]narendra.shekokar@djsce.ac.in

Abstract *– In today's world, it is essential to collect and identify the information of the same individuals from multiple databases in a secure manner for record matching, linkage and integration. Thus, the privacy-preserving record linkage (PPRL) refers to identifying and comparing the same person's records across multiple databases in secure manner. In this paper, the various PPRL techniques are discussed. Among the different PPRL techniques, the Bloom filter encoding is suitable for secure and approximate record matching. However, most of the hardened Bloom filter encoding techniques provide privacy while compromising linkage accuracy. Hence, the ensemble approach is suggested to provide improved linkage accuracy than existing basic, balanced, and cellular automata Bloom filter-based PPRL.*

Keywords — *Data integration, matching, privacy, linkage, Bloom filter, similarity measures.*

I. INTRODUCTION

Nowadays, the number of records belong to the same person are generated and stored in multiple data sets. With the tremendous demand for accurate data integration and analytics, there is the necessity of matching and linking records referring to the same entity called record linkage. This task of record linkage is difficult since real-world data may contain misspelled names, errors, missing values, duplicate entries, etc. Hence, there is a need for approximate matching of records rather than in an exact way. More prominently, since the databases contain confidential information, it is important to identify and match the records from various data sets in an encoded form termed as privacy-preserving record linkage (PPRL). Thus, the records across various databases are obfuscated and sent for secure record matching in PPRL. It is useful in different applications, including census, banking, healthcare, e-commerce, fraud detection, and so on [1][2][3][4].

The PPRL techniques utilize similarity measures for approximate matching of the encoding records. The similarity measures or metrics adopted for the comparison of numerical and string values are referred to as privacy-preserving numerical values (PPNVs) and privacy-preserving string comparators (PPSCs), respectively[5][6][7]. Achieving a trade-off between privacy and accuracy can be difficult with the utilization of secure hardening approaches for PPRL. The hardened Bloom filter-based PPRL approaches are currently most relevant for approximate matching. More precisely, the similarity measures should be compatible with PPRL techniques in order to achieve the same level of accuracy as with traditional matches utilized in schema and data matching methods [2][3][4].

Therefore, accuracy is a prominent challenge for secure data linkage and integration. Hence, in this work, an ensemble approach for PPRL is suggested and analyzed in terms of linkage accuracy and compared with cellular automata, balanced and basic Bloom filter-based encryption techniques.

II. RELATED WORK

The PPRL techniques had gained significant importance for securely matching and linking records in multiple databases. The prominent PPRL techniques includes secure multi-party computation, secure hashing, phonetic encoding, embedded space, differential privacy, Bloom filter encryption and so on [8][9][10][11][12][13][14][15][16][17]. In cryptography-based PPRL, the record matching is performed on encrypted records. Due to the cryptographic operations, there is an impact on the accuracy of the linkage. The distance-based PPRL utilize triangular inequality to determine the distance between attribute values represented as vectors. As a result of distance approximation, it can affect linkage accuracy for PPRL. The anonymization-based techniques satisfy k-anonymity to provide privacy for PPRL. The perturbation and generalization operations are used for PPRL, and hence the accuracy level depends on them. In PPRL techniques considering noise mechanisms, the obfuscation is carried out by adding random or extra information to the original databases. However, the trade-off between security and accuracy in noise-based PPRL generally depends on the number of perturbed data. The phonetic encoding methods consist of generating codes and then perform matching between them. The phonetic-based PPRL methods may lead to false matches and can result in reduced accuracy. So, considering accuracy and privacy aspects, the Bloom filter-based PPRL utilizing cryptographic primitives is a useful technique. It involves encoding of records and then performing approximate matching for PPRL [12][18][19][20][21][22].

The use of hashing and hardened Bloom filter techniques have been suggested by researchers for exact or

approximate secure record matching respectively. The Bloom filter-based PPRL include standard/basic, balanced, salting, cellular automata-based techniques, and so on[23][24].
The Bloom filter encoding and hardening methods for secure record matching and linkage are discussed as follows:

A. Bloom Filter Encoding for PPRL

The Bloom filter is considered a probabilistic storage data structure. It can be used to encode attribute values (strings/numerical) in PPRL[9][25]. The cryptographic properties compatible with Bloom filters were adopted for PPRL. Generally, the string/numerical values in various data sources are split into sub-sequences/tokens known as q-grams. The q-gram tokens can result in variants like 2-grams, 3-grams, 4-grams, 5-grams, and so on. The q-grams can be chosen depending on various application scenarios and areas of concern for data integration. Then, the number of hash functions (k) from cryptography fields like HMAC; SHA variants are applied on each q-gram of respective string or numerical values. It results in a hash code for each q-gram. Further, the modulus using the length of Bloom filter (l) is computed for each hash, thereby resulting in a hash value. Lastly, the Bloom filter positions are set to 1 for the respective hash value; otherwise, marked as 0. This Bloom filter-based encoding technique considering 2-grams performs very well for PPRL [2][26]. The bloom filter encoding of string 'john' from one database is depicted as shown in figure 1.

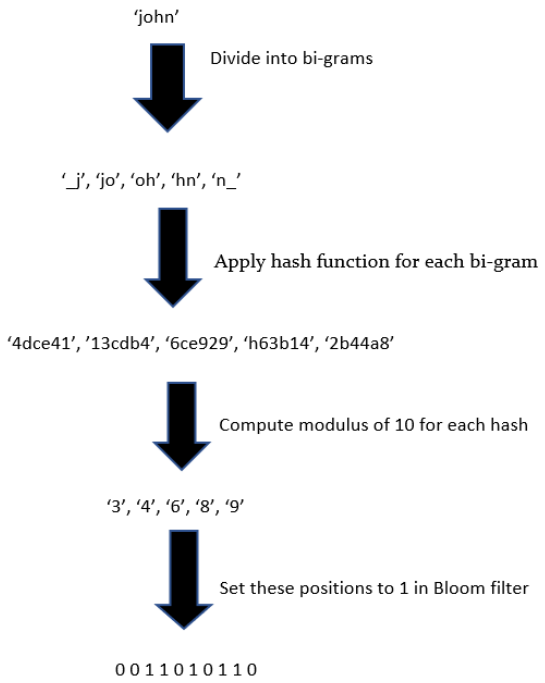

Fig. 1. Bloom filter Encoding for PPRL

Initially, as shown in figure 1, the string is divided into bi-grams. For each bigram, the hash function (say SHA-256) is applied to obtain the hash code. The modulus of 10 is computed for the hash code equivalent. For example, the 2-gram 'jo' from string S1= 'john' yields the value 3 for the hash function. The bit on position 3 is set to 1 in Bloom filter of length 10 representing string S1. This process is repeated for all bigrams for each string, and the corresponding bits are set to 1 into the Bloom filter.

Similarly, the string from another database is encoded and sent for PPRL. The similarity between Bloom filter encodings for attributes values S1 and S2 are calculated with metrics such as Dice Coefficient, which is given by:

$$sim(a, b) = 2 * (a \cap b)/(|a|+|b|) \quad (1)$$

where a and b are the Bloom filter encodings for strings S1 and S2 strings, respectively.

For instance, the matching between two Bloom filters a=0011010110 and b=0011001010 is performed using equation 1. The number of matching 1-bits in both encodings a, b (a ∩ b) is 3, the number of bits set in a(|a|) is 5, and b |b)| are 4, producing a similarity sim(a,b) as 0.67.

Thereafter, the similarity across the encoded data is generally calculated using metrics like Tanimoto coefficient, Jaccard coefficient, Dice coefficient, etc. The value obtained as a result of similarity calculation is compared to the user-defined threshold to consider the two strings as match or non-match.

The Bloom filter encoding for secure data linkage and integration achieves good accuracy and can be easily extended with less computation. Several variants of Bloom filter encoding were suggested for PPRL [22][23][24].

B. Bloom Filter Encoding Hardening Methods

The various Bloom filter based hardening methods were suggested for PPRL as discussed below:

a) Balanced Bloom Filter

Usually, the Hamming weights, i.e., the number of 1s in Bloom filter encodings, can be utilized to re-identify the encoded data. It may reduce the patterns required for the frequency-based attack. The Balanced Bloom filter can be constructed by combining encoding of Bloom filter length l with all its bits flipped. The resulting codes with constant Hamming weights of length 2l bits are called balanced codes. This method can be difficult to attack; however, the balanced codes leads to scalability issue and increased computational complexity[27][28][29].

b) Random Hashing

The random hashing includes the generation of k random numbers for each possible bigram. Initially, for every attribute value, all the possible q-grams are produced. For each possible q-gram, the random numbers are drawn within the range 1 and length of the Bloom filter l. It also includes replacements by a password mechanism with a Bloom filter marked at k bit positions as one. The hash functions are no longer required, and a pattern-based attack is unlikely since the search requires much more computational effort [2][24].

c) XOR-Folding

The XOR-folding approach considers the Bloom filter encoding of length l to be divided into two Bloom filter encodings of length l/2 each. Then, the bit-wise exclusive OR (XOR) is applied to combine the splitted parts of Bloom filters. It was found to be data independent hardening technique. This easy approach makes cryptanalysis attacks difficult due to the folding of Bloom filter encoded data a number of times. However, it resulted in a loss in precision and recall[24][30].

d) Salting

The salting method utilizes an additional value (salting key) generated from a suitable identifier. This key is concatenated with original attribute values prior to hashing into the Bloom filter. Hence, the cryptanalysis attack will be difficult without knowing the salting key. The security measures during record linkage can be effectively built with a salting mechanism. However, the salting key containing short attribute values (e.g., date of birth) may contain errors. Also, the salting keys, if they are not equivalent, could create a problem for PPRL. Additionally, the attributes compatible for salting may not be available in the data [2][24].

e) Random Noise

Random noise is a data perturbation technique. It includes inserting extra or fake records into the data to be linked for overcoming the problem of cryptanalysis attacks during PPRL. However, the addition of random noise could considerably increase imbalanced privacy and linkage accuracy[2][23].

f) Bloom and Flip

The Bloom and flip (BLIP) method consider arbitrarily flipping the values of bit positions of Bloom filter encryption using permanent randomized response for PPRL. Each bit position in the Bloom filter is treated with a randomized response resulting in a new Bloom filter with the flipped bit value with some bias. It includes the use of differential privacy. The bit flipping probability f and the privacy parameter epsilon are co-related. A sufficient level of privacy through differential privacy can be achieved with lower values of privacy parameter epsilon. The increased privacy level with differential privacy in the BLIP mechanism makes the deterministic attacks practically impossible [24][31].

g) Cellular Automata

The PPRL approach based on cellular automata adopts the wolfram rule 90 to harden the Bloom filter encoding. It considers the replacement rules for transformating the bits in the Bloom filter encoded records [24][32]. The middle bit is transformed for every set of three inputs of Bloom filter encryption using following replacement rules:

{'111': 0, '110': 1, '101': 0, '100': 1, '011': 1, '010': 0, '001': 1, '000': 0}

Thus, the Bloom filter encoding hardening techniques enforce increased privacy against cryptanalysis attacks while affecting the linkage accuracy.

Thus, in this article, the various hardening techniques for PPRL are identified and discussed. The cellular automata-based technique can provide better security in PPRL. However, it can result in reduced linkage accuracy. The following section introduces the ensemble approach for PPRL utilizing phonetic encoding and cellular automata to achieve acceptable accuracy in PPRL.

III. ENSEMBLE APPROACH FOR PPRL

The hardening of Bloom filters will impact the linkage accuracy in PPRL. In this paper, an ensemble approach for secure record matching and linkage is introduced to further increase linkage accuracy for hardened Bloom filters. It consists of two-factor encoding, which employs phonetic codes and hardened Bloom filter encoding for record matching in an approximate manner.

In this approach, the two parties choose the parameters for Bloom filter-based PPRL. These parameters consist of the q-grams value, the number of hash functions, length of Bloom filter, similarity measures, etc. The common attribute values across the databases are identified for the encoding process in PPRL. Initially, the phonetic method (e.g., Soundex) is applied to person names resulting in phonetic codes, and these codes are converted into q-grams. The hash functions such as SHA3-384, SHA-512 are applied on the generated q-grams to obtain the hash code. The modulus of Bloom filter length is performed on the hash to obtain the hash value. The bit positions across the Bloom filters of respective attribute values are set to 1 as per the obtained hash value. Later, the resultant Bloom filters are compared using similarity metrics like the Dice coefficient. The similarity values are then checked with the threshold decided by the user to know the matched and non-matched records.

The algorithm for the ensemble PPRL approach is discussed as follows:

Input: Database Di containing identifiers Ii, i<{1,2,…..N} with records ri and database DSj consisting of identifiers Ij, j<{1,2,…..N} with records rj

Output: Records matched between ri, rj

Steps:

Begin

Step 1: Each party agree on similar identifiers IiεDi and IjεDj and number of records riεIi and rjεIj appropriately in PPRL.

Step 2: Every party encodes records with hardened PPRL technique as:

- For every identifier IiεDi and IjεDj, encode ri and rj to create phonetic codes PCi and PCj respectively with phonetic technique.
- For every phonetic code PCi and PCj, create q-grams qi and qj, where q=1, 2,..,n.
- For every qiεPCi and qjεPCj, generate cellular automata(CA) Bloom filter encodings CABFEi and CABFEj as follows:

$$\text{CABFEi} = (\text{hi}(\text{qi}) + \text{i}\,\text{hj}(\text{qi}))\bmod \text{l} \quad (2)$$

$$CABFEj = (hi(qj) + j\ hj(qj)) \bmod l \quad (3)$$

Step 3: The resultant hardened encodings CABFEiεri and CABFEjεrj are then sent for record matching to a trusted party.

Step 4: Comparison of encodings CABFEiεri and CABFEjεrj by utilizing Dice coefficient to obtain similarity value sim(ri, rj) as

$$sim(ri,rj)=2*m/(|ri|+|rj|) \quad (4)$$

where |ri| and |rj| are the count of 1-bits in Bloom filter encodings CABFEi and CABFEj for records ri and rj respectively and m is the number of common 1-bits between CABFEi and CABFEj.

When sim(ri, rj)> threshold θ,

ri, rj=match,

then

ri, rj=non-match.

Step 5: The status of approximately matched records ri, rj are communicated to parties participated in secure record linkage.

End

The ensemble approach adapts the hardened Bloom filters and inherent property of phonetic encoding to achieve increased linkage accuracy for PPRL.

IV. RESULTS AND DISCUSSION

The PPRL technique is essential to protect personal identifiers during record matching. The Bloom filter-based encryption techniques are essential to compare the encoded records in PPRL. In this work, the basic, balanced Bloom filter and cellular automata techniques are implemented for PPRL. To improve the accuracy, the ensemble approach is suggested and compared with existing Bloom filter encryption techniques for PPRL.

The parameters considered for the Bloom based PPRL are:

Bloom filter length l=30

q-grams=2

Hash functions=2 (SHA 3-512, SHA 3-384)

Padding=Yes

Dice Coefficient threshold value=0.85

The voter registration data set (NCVR) containing 3 identifiers, Lastname, first name, and middle name, are considered in secure record matching and linkage. The initial experimentation for PPRL contains 200 records among the two databases. The sample outcome of encoding string ‘john’ and ‘jon’ using above mentioned parameters for basic Bloom filter based PPRL is shown as follows:

- **Encoding of string ‘john’ in data set 1**

gram:_j
hash code :5bb02e
Mod Value :29
Sha3-512 :000000000000000000000000000010
gram:_j
hash code :f3b538
Mod Value :1
Sha3-384 :100000000000000000000000000010
gram:jo
hash code :e1a45c
Mod Value :17
Sha3-512 :100000000000000010000000000010
gram:jo
hash code :db71c4
Mod Value :19
Sha3-384 :100000000000000010100000000010
gram:oh
hash code :6cee03
Mod Value :20
Sha3-512 :100000000000000010110000000010
gram:oh
hash code :345107
Mod Value :6
Sha3-384 :100001000000000010110000000010
gram:hn
hash code :79c42a
Mod Value :15
Sha3-512 :100001000000001010110000000010
gram:hn
hash code :0bbdd3
Mod Value :22
Sha3-384 :100001000000001010110100000010
gram:n_
hash code :4135bc
Mod Value :7
Sha3-512 :100001100000001010110100000010
gram:n_
hash code :d34266
Mod Value :5
Sha3-384 :100011100000001010110100000010

- **Encoding of string ‘jon’ in data set 2**

gram:_j
hash code :5bb02e
Mod Value :29
Sha3-512 :000000000000000000000000000010
gram:_j
hash code :f3b538
Mod Value :1
Sha3-384 :100000000000000000000000000010

gram:jo
hash code :e1a45c
Mod Value :17
Sha3-512 :1000000000000000010000000000010
gram:jo
hash code :db71c4
Mod Value :19
Sha3-384 :1000000000000000010100000000010
gram:on
hash code :028e3a
Mod Value :23
Sha3-512 :1000000000000000010100010000010
gram:on
hash code :2f7139
Mod Value :8
Sha3-384 :1000000100000000010100010000010
gram:n_
hash code :4135bc
Mod Value :7
Sha3-512 :1000001100000000010100010000010
gram:n_
hash code :d34266
Mod Value :5
Sha3-384 :1000101100000000010100010000010

Thus, the final encoding of given two strings 'john' and jon are 1000111000000010101101000000010 and 1000101100000000010100010000010. The similarity value between these encodings was found to be 0.67 as calculated through the Dice coefficient, which is greater than the user-defined threshold. Hence, the encodings representing the given two strings are considered to be a match.

Importantly, the ensemble, cellular automata, basic and balanced Bloom filter secure record linkage techniques are examined with respect to f-measure, recall, and precision as depicted in table I. The f-measure (F), precision (P), and recall(R) are calculated using false positives (FP), false negatives (FN), and true positives (TP) as:

$$P = TP/(TP + FP) \quad (5)$$

$$R = TP/(TP + FN) \quad (6)$$

$$F = 2 * (P * R)/(P + R) \quad (7)$$

Table I. Analysis of secure record linkage techniques

Factors\ Techniques	TP	FP	FN	Precision	Recall	F-measure
Basic Bloom PPRL	9	0	14	100	39.13	56.25
Balanced PPRL	12	64	11	15.79	52.17	24.24
CA PPRL	9	1	14	90	39.13	54.54
Ensemble PPRL	13	3	10	81.25	56.52	66.67

Table I indicates that the precision for basic, cellular automata (CA) based techniques is higher than balanced and ensemble PPRL for Dice coefficient (DC) threshold value of 0.85. There are no false positives observed for basic and CA-based, thereby leading to higher precision. The false positives are more likely for balanced-based PPRL. The ensemble approach provides matched records with a greater number of true positives and better recall than existing CA, balanced and basic Bloom PPRL techniques.

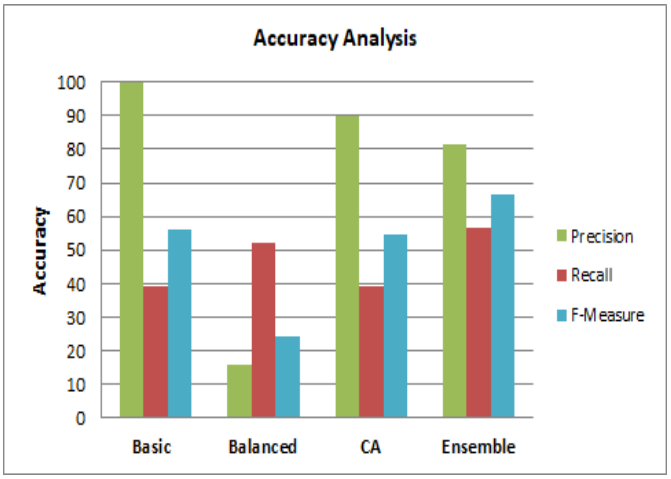

Fig. 2. F-measure, Recall, and Precision for PPRL Techniques

Figure 2 depicts that the basic and CA-based PPRL has better precision than balanced and ensemble techniques. This precision is achieved due to the fact that there are no false-positive outcomes for CA and basic techniques for PPRL. The highest recall is observed for ensemble PPRL, which indicates that there is a better percentage of matched records for it than existing balanced, basic, and CA-based PPRL techniques.

Moreover, the percentage of true positives or exact matches is greater for ensemble-based PPRL, as shown in figure 3.

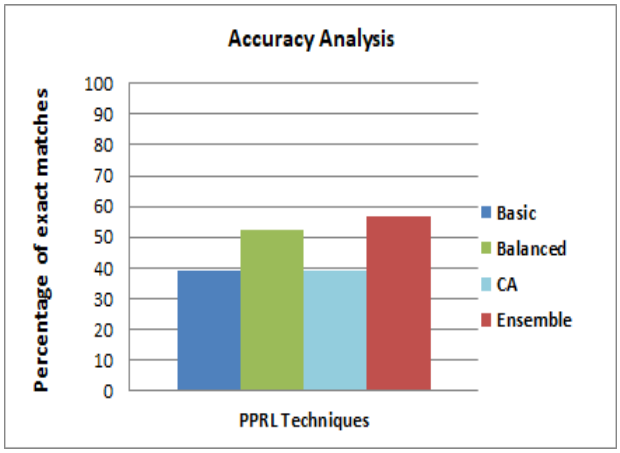

Fig.3.Percentage of true positives/exact matches for ensemble, CA, balanced, and Bloom filter based PPRL

Figure 3 indicates that the ensemble technique results in an increased number of true positives as compared with CA, balanced and basic Bloom-based PPRL.

The Dice coefficient value ranging from 0.85 to 1, as shown in figure 4, has an impact on the accuracy of PPRL techniques.

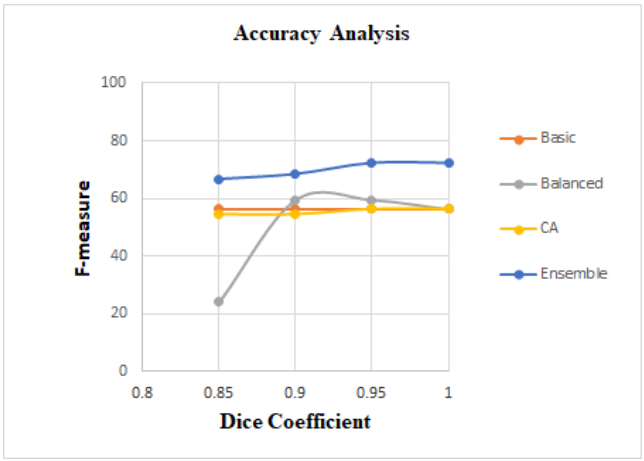

Fig. 4. Variation of Dice coefficient for PPRL techniques

It can be seen from figure 4 that the ensemble-based PPRL results in improved f-measure for higher thresholds of 0.95 and 1. Therefore, the ensemble approach performs better than balanced, basic, and CA-based secure record linkage techniques.

V. CONCLUSION

There is a need for secure approximate matching due to the presence of erroneous and confidential information across different databases. Hence, the Bloom filter encoding hardening methods have gained significance for approximate matching and providing security during PPRL. But there is a significant impact on linkage accuracy due to the hardening of Bloom filter-based PPRL techniques. In this research, the ensemble technique is proposed to obtain the increased linkage accuracy for PPRL. Further, this technique can be combined with anonymization mechanisms for achieving better privacy and accuracy for data integration and publishing.

REFERENCES

[1] Christen, P. Febrl – A Freely Available Record Linkage System with a Graphical User Interface. in HDKM '08 Proceedings of the Second Australasian Workshop on Health Data and Knowledge Management, Darlinghurst, Australia: Australian Computer Society, (2008) 17–25.

[2] Christen, P., Vatsalan, D. and Verykios, V. S. A Taxonomy of Privacy-Preserving Record Linkage Techniques. In Journal of Information Systems (Elsevier), 38(6) (2013) 946-969.

[3] Bernstein, P. A., Madhavan, J. and Rahm, E. Generic Schema Matching, Ten Years Later, PVLDB, 4(11) (2011) 695-701.

[4] Christen, P., Vatsalan, D. and Verykios, V. S. Challenges for Privacy Preservation in Data Integration. ACM Journal of Data and Information Quality, 5(1-2) (2014) 1-3.

[5] Durham, E., Xue, Y., Kantarcioglu, M. and Malin, B. A. Quantifying the Correctness, Computational Complexity, and Security of Privacy-Preserving String Comparators for Record Linkage. Information Fusion, 13(4), Elsevier, (2012) 245-259.

[6] Navarro-Arribas, G. and Torra, V. Information Fusion in Data Privacy: A Survey. Information Fusion, 13(4), Elsevier, (2012) 235-244.

[7] Vatsalan, D. and Christen, P. Privacy-preserving Matching of Similar Patients. Journal of Biomedical informatics, 59, Elsevier, (2016) 285-298.

[8] Vatsalan, D. and Christen, P. Scalable Privacy-Preserving Record Linkage for Multiple Databases. In CIKM '14: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, (2014) 1795–1798.

[9] Schnell, R., Bachteler,T. and Reiher, J. Privacy-Preserving Record Linkage Using Bloom filters. BMC Medical Informatics and Decision Making, 9(1) (2009).

[10] Shelake, V. M. and Shekokar, N. A Survey of Privacy Preserving Data Integration. In 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Mysuru, (2017) 59-70.

[11] Christen, P., Schnell, R., Vatsalan D., Ranbaduge T. Efficient Cryptanalysis of Bloom Filters for Privacy-Preserving Record Linkage. In: Kim, J., Shim, K., Cao, L., Lee, J.-G., Lin, X., Moon, Y.-S. (eds.) Advances in Knowledge Discovery and Data Mining. PAKDD 2017. Lecture Notes in Computer Science, 10234, Springer, Cham, (2017) 628-640.

[12] Christen, P., Ranbaduge, T., Vatsalan, D. and Schnell, R. Precise and Fast Cryptanalysis for Bloom Filter Based Privacy-Preserving Record Linkage. In IEEE Transactions on Knowledge and Data Engineering, 31(11) (2019) 2164-2177.

[13] Randall, S. M., Ferrante, A. M., Boyd, J. H., Bauer, J. K. and Semmens, J. B. Privacy-Preserving Record Linkage on Large Real World Datasets. Journal of Biomedical Informatics, 50, Elsevier, (2014) 205-212.

[14] Russell, R. C. US Patent No 1,261,167.,(1922).

[15] Bouzelat H, Quantin C, Dusserre L. Extraction and Anonymity Protocol of Medical File. In Proc. AMIA Fall Symposium, 1996, pp.323-327.

[16] Quantin, C., Bouzelat, H., Allaert, F.A. A, Benhamiche A-M., Faivre, J. and Dusserre, L. How to Ensure Data Security of an Epidemiological Follow-Up: Quality Assessment of an Anonymous Record Linkage Procedure. International Journal of Medical Informatics, 49(1), Elsevier, (1998) 117-22.

[17] Karakasidis A., Verykios, V. S. Privacy Preserving Record Linkage Using Phonetic Codes. In 2009 Fourth Balkan Conference in Informatics, Thessaloniki, (2009) 101-106.

[18] Karakasidis, A., Koloniari, G. Private Entity Resolution for Big Data on Apache Spark Using Multiple Phonetic Codes. Big Data Recommender Systems - Volume 1: Algorithms, Architectures, Big Data, Security and Trust, Chap. 13, IET Digital Library, (2019) 283-301.

[19] Karakasidis A., Verykios, V. S. and Christen, P. Fake Injection Strategies for Private Phonetic Matching. In: Garcia-Alfaro J., Navarro-Arribas G., Cuppens-Boulahia N., de Capitani di Vimercati S. (eds) Data Privacy Management and Autonomous Spontaneus Security, DPM 2011, SETOP 2011, Lecture Notes in Computer Science, 7122, Berlin: Springer, (2011) 9-24

[20] Abir Bin Ayub Khan, Mohammad Sheikh Ghazanfar, Shahidul Islam Khan. Application of Phonetic Encoding for Analyzing Similarity of Patient's Data: Bangladesh Perspective. 10 Humanitarian Technology Conference (R10-HTC), (2017) 664-667, IEEE.

[21] Koneru, K. and Varol, C. Privacy Preserving Record Linkage Using

MetaSoundex Algorithm. In 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, (2017) 443-447.

[22] Brown, A. P., Borgs, C., Randall, S. M. and Schnell, R. Evaluating Privacy-Preserving Record Linkage using Cryptographic Long-term Keys and Multibit Trees on Large Medical Datasets. BMC Medical Informatics and Decision Making, 17(83) (2017).

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

[24] Christen, P., Ranbaduge, T. and Schnell, R. Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing. Springer Science and Business Media LLC, 2020.

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

[26] Manning, C., Raghavan, P. and Schuetze, H. Introduction to Information Retrieval. 39. Cambridge University Press, 2009.

[27] Schnell, R. and Borgs, C. Randomized Response and Balanced Bloom Filters for Privacy Preserving Record Linkage. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, (2016) 218-224.

[28] Knuth, D. E. Efficient Balanced Codes. IEEE Transactions on Information Theory, 32(1) (1986) 51–53.

[29] Berger, J. M. A note on error detection codes for asymmetric channels, Information and Control, 4(1) (1961) 68–73.

[30] Schnell, R. and Borgs, C. XOR-Folding for Bloom Filter-based Encryptions for Privacy-preserving Record Linkage. German Record Linkage Center, NO. WP-GRLC-2016-03, SSRN, (2016).

[31] Alaggan, M., Gambs, S. and Kermarrec, A-M. BLIP: Non-interactive Differentially-private Similarity Computation on Bloom Filters. In Stabilization, Safety, and Security of Distributed Systems: 14th International Symposium, SSS 2012, Toronto, Canada, October 1–4, 2012. Proceedings, A. W. Richa and C. Scheideler, Eds. Berlin: Springer, (2012) 202–216.

[32] Schnell, R. and Borgs, C. Hardening Encrypted Patient Names Against Cryptographic Attacks Using Cellular Automata. 2018 IEEE International Conference on Data Mining Workshops (ICDMW), Singapore, Singapore (2018) 518-522.