International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Enhancing Fraud Detection with Privacy-Preserving Record Linkage and Digital Signature (EdDSA) Techniques: A Data Sharing Approach

MARY SHARMILA T[1], AMBATI GANESH[1], KURUVA MANJULA[1], CHINTAPALLI BHARANI KUMAR NAIDU[1*]

[1]*Department of Computer Science,*
Indian Institute of Industry Interaction Education and Research, Chennai, Tamil Nadu 600066

Abstract - Fraud as a threat to the telecom business is very real amid escalating and hopeless problem. Many telecoms also have different sets of measures that they employ against fraud. The prevention of fraud requires information and it is through this premise that detection comes in. There is also the lack of privacy rights for the exchange of information as people's privacy becomes restricted. There are a variety of techniques defined to ensure that data sharing can also include PPRL considerations. Current works are effective due to the reasons that many of the PPRL techniques adopt a similarity measure which could be Jacquard similarity on relevant datasets. As it has been mentioned earlier, the Bloom filter implementation is applied in a number of complex and slow algorithms; however, the given technique is critically sensitive to the cryptanalysis attacks. Based on current telco infrastructure and without having to go through a multistep protocol mechanism, this paper introduces a PPRL way which is indeed invulnerable to attacks. Starting with a fresh approach to matching non-homologous datasets is the application of the new attack-proof Digital signature system like the Edwards curve. Note that what is capability of this approach is only to estimate the Jaccard similarity here, which the datasets cannot be used. It's also important to note that basic request-response model is implemented for two-partner interactions. The performance, privacy, and matching accuracy of the approach have been evaluated by employing a scale of 1 to 5 using a large set of data set public. Concrete against attacks, this technique provides an enhanced pace and attains perfect match consistency.

Key Words: Telecom fraud, PPRL (Privacy-Preserving Record Linkage), Information exchange, Jaccard similarity, Bloom filter, Digital signature system (Edwards curve)

1.INTRODUCTION

Telecommunication organizations are the primary suppliers of technical connection around the world [1], as well as they become the objectives of a range of fraud attacks each day. This demand is steadily growing with the development of telecom goods and the increasing virtualization of our lives, in addition to their necessity to detect, prevent, and counteract these attacks. Therefore, telecommunications firms are the primary responsible for protecting internal and consumer data transmitted over their network. Telecom fraudulence is said to be in the billions of dollars and this has been voiced by the Communication Fraud Control Association. Nevertheless, many kinds of frauds can be singled out, and one of the most widely spread types is the subscription fraud. Telecom business can actually compromise their reputability and their customers in the process if a case of fraud is experienced. The businesses must also leverage their client traffic by accessing he other's networks. Thus, if one telecom business is attacked, it may complicate things for other related companies. Criminals are aware of and take advantage of the fact that verification and detection of traditional telecom fraud management system works in isolation. The moment they are acknowledged, they jump from one telco to a different one. In order to combat these scammers, telecoms can pool their efforts and information. Even this was impossible in the past due to state, federal, and international privacy laws such as Europe ' s GDPR and Computer Fraud and National Identity Act (CPNI). The process of passing of legislation or formulation of several similar laws is at present being considered. Telecommunications companies have to protect business information and to respect the customers' data privacy legislation in order to share the data. The three primary obstacles are as follows:The three purposes are (1) a safe, expedient, and easy protocol; (2) passing information without the inclusion of a third party; and (3) passing fraudulent information without sharing the data.This is made possible by the techniques of privacy-preserving record linkage (PPRL).

Another concern that researchers and practitioners may have with the existing PPRL approaches is that not all of them are security. The PPRL algorithms, and in fact, the majority of content-based retrieval models, make use of a similarity measure in order to retrieve a match. Recent work done by Vidanage, et al: [2] demonstrates that graphs can be utilized by present day cryptanalysis techniques to access similarity measurements, if these measures are identified on transcoded data. He proposes that all existing PPRL techniques are as a result unsafe. Anyway, the usage of Bloom filters is relevant to a multitude of PPRL implementations. According to Schnell, these implementations are insecure to frequency attacks in contrast to different cryptanalysis methods applied to the q-grams of a Bloom filter that contains the string.

Due to the desire for getting and selling safe qualities many individuals use a credible third party. Thus, the purpose of applying this technique is to avoid the two parties from sharing any information.

They are cumbersome, their implementation is cumbersome, they are slow, and they encompasses. Conclusively, there are many PPRL algorithms that use counting and randomization to adjust the Properties of the Bloom filter [3].

All existing PPRL implementations fail in the sense that both the sets of data are encrypted by using the same algorithm. However, even if the details are encrypted the records still share some measure of similarity. In that case, a conventional attacker may employ the graph attacks to decrypt the encrypted data by computing a similarity measure [2]. A method to compare data which have passed the entirely different routines and is processed only at the receiving side, is needed to prevent this attack. This is all that is needed; to encrypt the data and send it for comparison with the unencrypted data. This ensures that the two pieces of data will never hold the same value and the encrypted data is not easily decoded by the enemy when in transit. According to our findings, our research proposes a new PPRL technique that is as fast as using a TTP while being secure and capable of exchanging fraud

data between telcos using the simplest of request-reply formations.

This work proposes a new approach for computing the Jaccard index on nonhomologous datasets using (1) transmitted encrypted MinHash data through digital signature technique with no possibility of attack; and (2) non-transmitted conventional MinHash data. In summary, only the original input datasets of the type that has been processed using the present approach may the similarity measure be calculated.

- Contrary to the DSS, the sender is not required to pass on the original message while transferring the digital signature; only a request-response protocol is employed here. - As opposed to other implementations, this one employs the match in MinHash data, not a complicated q-grams or feature extraction pattern.

Instead, an applicant can choose the SuperMinHash method we proposed here, which we found to be faster than the EdDSA algorithm that is described as fast and secure; the only difference in the process is the transmission of the encrypted signature without the original message. Unlike the Dice coefficient that is computed using the EdDSA verification function, the Jaccard index (J) is one that satisfies the triangular inequality. To further ensure the speed and security of our PPRL approach, we will also give an indication of an optimum threshold value (JT) to compare our approach with other procedures and a length of signature (N).

2. IMPLEMENTATION

2.1 With Bloom filters

H and bit arrays L are the components of bloom filters and K is an independent hash function. If for instance we wish to map the set S = {a1, a2, …. . an} to the Bloom filter we might define Hi when and where we can get Hi that is an internship from 0 and K-1. Here, filter is set to 1 to every element ai's hash index. When the idea of validating the set membership is conceived, one can use same hash algorithms that hash the components that have to be matched. Collision and false positives are disadvantageously associated with the Bloom filters.

The IDs are decomposed into q-grams to be encrypted in the subsequent processing with the help of PPRL configuring Bloom filters. The term "privacy preserving" may be divided into the following words using q-grams splitting: These word graphs are labelled as "pr," "iv," "ac," "yp," "re," "se," "rv," "in," and "g. " To each of these word graphs, a total of K hash are applied before passing them to the Bloom filter. The sender and the receiver are able to evaluate if the set belongs to a given set and also calculate a similarity score, which is, for instance, the Dice coefficient, to look for a match. As a further defense against these assaults, they are employing balanced Bloom filters to strengthen their implementation. Specifically, Bloom filters of some length are joined with the negative of the same Bloom filters, because it is challenging to delete uncommon motifs with a constant accumulation distance. Very interesting mode based on Bloom filters and the Dice coefficient, whereby the transmitter and receiver each generate their own secret keys. While one party applies its secret key to encrypt not only the data of its owner, but data of another, the process is carried out merely with the following actions To encode data of the sender and recipient, and then send them back and forth. After the sender encrypts messages using the sender's key and sends the data, the receiver receives the data and the privacy algorithm decrypts the data.

2.2 Without Bloom filters

A suggested method that is not based on Bloom filters, but by applying the hashing-based encoding on qgrams and Longest Common Bit Sequence. In this implementation, a semi-trusted third-party linkage unit is employed to perform the matching and inform the appropriate party. Encrypting the information to be transmitted with a private-public key method and breaking them into q-grams. Matching is performed on that data that is encrypted. In cases where both parties are cryptographically vulnerable and are willing to trust each other, this implementation is done. Without or with q-grams or even Bloom filters regarding the investigated data. In order to make sure a match is found, a similarity metric such as the Dice coefficient is used after a pseudo randomisation method has replaced actual data by a safe pseudonym. An approach to a transaction where some independent third party is not directly involved in the transaction. Thus, the parties employ the Diffe-Hellman protocol for the transmission of information to preserve the privacy of the datasets involved. In any particular transaction, Secret Keys SA and SB would belong respectively to Parties A and B. To encrypt its message, each party employs its individual key: MA or MB. Both A and B switch between MASA and MBSB messages. At this point, both A and B encrypt the messages exchanged between them back to the other using their respective keys. B holds MBSAïsb and A holds MASAïsb. They will then be able to both find out if MA and MB are equivalent.

2.3 Issues

Cryptanalysis attacks can affect any of the current PPRL Distinctions since they are not directly guaranteed. As far as the blind search for a particular person who could become a match is concerned, the majority of the existing solutions utilize similarity measures. A dataset which has gone through the same processes of encryption as the actual data is utilised to obtain the similarity measure. Every PPRL uses vulnerability to assaults which was demonstrated by the similarity metrics to conduct [2] cryptanalyses. It is clear that the application of a Diffe-Hellman protocol or any other uncomplicated key exchange method exhibits unsuitable security parameters, and there was no presence of a similarity measure.

How Bloom filters are used: It also discourages cryptanalysts from attacking the Bloom filter as a specific part of PPRL because of the high success rate; [6]. In cryptanalysis, an attacker enjoys the areas of the encoded filter, Bloom filter to take advantage in some of the positions where a certain value say q-gram or frequency or hamming distance lies. It is possible to search for a match in a database with names if the applicable databases are accessible to the public; there are no requirements to understand all the parameters or the process of encoding them. They have also decoded all the known Bloom filter hardening strategies they identified through cryptanalysis.

In the implementations, the availability of using a semi-trusted third party who conducts matches was included. In some ways, the telecom business cannot employ the use of a third party because of privacy concerns. A mutually trusted third party is also something that was mentioned as something that telecoms may not settle on. This is so because several transactions need to occur among the involved third parties in an endeavor to get a match.

In fact, many systems deal with various transactions and high speeds where such data is processed to q-grams, the extraction of character frequencies together with hamming distance, and the

output is then further pseudo-random to have additional layers of protection. Because of this, problems with processing speed arise, as well as the higher implementation complexity for the telecom sector and mistakes made in matching at the level of accuracy.

3. PROPOSED PPRL METHOD

Telecommunication companies notice, monitor and indeed, stop fraud incidents on daily basis; however, for various reasons they don't talk about it. The following pieces of information may be stored by telcos: The elements to classify include the personal identification numbers such as name (first, middle, and last), postal/physical address, network address or location, social security number, account number, and/or signs of fraud. Without third-party verification, based on the PPRL approach described above, in Figure 1, the telcos may exchange fraud indicators with one another, while keeping individual data concealed through a request reply process. This is why, first, it is proposed to designate the code of a safe encryption scheme (Section 3. 1) and a successful MinHash algorithm (Section 3. Hence, a new secure matching procedure will be incorporated into the PPRL architecture as indicated in Section 3. 3.

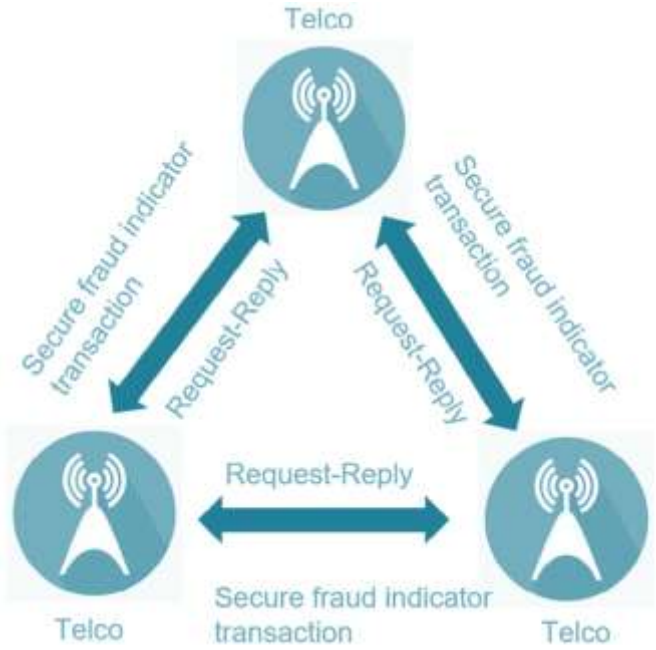

Fig 2: Fraud indicators across the Telecoms without a middle man

3.1 Encryption

When implementing PPRL without a third party NPDL recommends selecting a protocol for the secured transfer of data. The Diffie-Hellman key exchange is the key officer of a number of techniques of cryptology. Since the introduction of symmetric and asymmetric key exchange several security protocols have emerged namely: DSS, ElGamal and RSA. Thus, in a DSS, we incorporate a MinHash into the protocol to calculate the Jaccard index over the encrypted data. Due to the simplicity in using the basic request-response machineries, the telcos are able to transact securely: thus safeguarding each's information. It remains the common practice of sending the signed message together with the original one in case of a digital signature. It is also designed to make sure that the encrypted signature is the only thing sent out and not the content. Since the receiving telcos are going to have to search our protocol for a matching connection, this will remain anonymous.

3.2 MinHash method

In order to find a match between instances in datasets D1 and D2, we used the Jaccard index, where $J \in [0, 1]$ and $J = 1$, if D1 = D2. This was at first intended for application in ecological diversity indices and now adopted in many fields of study [7]. Concerning similarity measures, one learns that the Jaccard index is more accurate than the Dice coefficient because it satisfies the triangle inequality.
Thus, in addition to being more secure, MinHash is also faster than the Jaccard index due to its one-way operations. Even if the source of the hash could be found it is rather complex and may take a lot of time to compute. Different MinHash algorithms have been developed in order to improve the efficiency and accuracy of using the Jaccard index. Some of these are HyperMinHash, MinMaxHash, ProbMinHash and SuperMinHash. Because of it is efficiency and effectiveness, we used SuperMinHash in our implementation. The time complexity of the MinHash algorithm is on average O(NM), where N is the size of the signature and M is the number of elements in the data set. Therefore, the runtime complexity for the SuperMinHash algorithm was O(N + M log2M) which is far better.

4. RESULTS

As described in Section 4.1, after identifying the experimental setting. Thus in Section 4.2, we proceed further to explain three trials conducted using the procedure and their validity. Section 4. 3 focuses on the experimental analysis of the PPRL method, and Section 4.4 assess the privacy of the protocol. In section 4.5, we compare the performance of the developed PPRL with some other approaches..

4.1 Setup

For developing our PPRL protocol and conducting the trials, this research utilized the 16 GB RAM, the Apple M1 processor, and the Apple MacBook Air. We came up with a Python module that emulates the protocol that was discussed in the previous section. Many different libraries were incorporated into the work, including those related to EdDSA and SuperMinHash. Specifically, there were imports of the ED25519 libraries, which are used for EdDSA. Rather than generating data using simulation models, we decided to employ a real life voter database from North Carolina to generate a TB fraud database. Instead of the fraud indicator FiA for TA, we employed the voter database since names used in voting are familiar and selected only a handful of fake names. To facilitate the analysis, we selected a sample size of 10000 for the record size. The presence or absence of a match in FiA was determined by utilizing the identify, RONALD JOHN ADAMS as TB's fraud indication FiB. In this present work, we select a array of values for the Jaccard threshold value JT; starting from 0. 1 to 1, and a variable N to accept various signature sizes in the Python module, which feeds into SuperMinHash. We wrapped the Python time function around them in order to determine the runtimes of the different phases of the protocol.

4.2 Measurements and validity

Figure 3 depicts the data gained by our total run-time measurements of the PPRL method with signature lengths N between 2 to 44. The orange bars represent the total number of seconds required to do the following operations: Approximately the rates for generating EiA given FiA and FiB, using EiA to verify FiB, and calculating the Jaccard index. Since the protocol would go through all the data in FiA that is proposed to contain 5-signe length, the unidentified execution time increases by around 100s with every increase of 5-signe length. This corroborates the PPRL method's intended behaviour and means this experimental setting and data are justified.

Third experiment: when a match on FiB is present in FiA, comparing the success of the algorithm for the new, and the old FiB and FiA signature sizes. The overall spread of runtime for signature lengths N=3 to N=45 is pointed out with blue bars in Figure 4. With increasing mismatch degree, the matching run time increases and it grows by 0. 5 s for every MHz of the signature length when it increases by 5, due to all the iterations through FiA. The run time in experiment 3 is shorter by 1/10,00th of a second because Algorithm 1 shows a planned pause when there exists a match; results of experiment 3 and 2 have been depicted side by side in figure 3. The collected and used experimental data is correct and reliable as the measured values act in accordance with the basic concept of the PPRL approach.

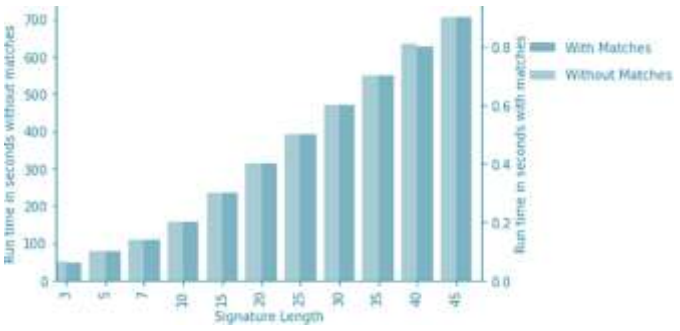

Fig 3: Comparison of four different performance indices for matched and unmatched M-sequences of equal length but different numbers of elements in each sequence.

4.3 PPRL method

The precision of the PPRL approach equation was established by computing four distinct basic signature lengths; N = 5, as shown in figure 4, N = 10, N = 15, and N = 20. To determine value accuracy, we employed a Jaccard threshold (JT) that was between 0 and 1. The patient's weight changed 1 to 1, with each increment of 0. 1. In search of an analogue of FiB in FiA, the same experiment was conducted as was outlined in Section 4. 2. If JT is below 0, then the values of the accuracy are lesser. 5, represented in figure 5, and it is able to keep the precision of 1 while having a minimum threshold of 0. 5. Moreover, in the case when the length of the signature increases, one can observe that our approach provides a satisfactory accuracy, based on comparatively small JT values. When it comes to longer signatures, it is understood that there will be more values that would have to be compared, and this considerably reduces FP and at low JT, the precision value comes out to be 1. This complicates the performance of the approach when the string length is expanded to high values for better accuracy at low JT values. It does, however, depict the right length of the signature and threshold length through which, when optimized, results in the best run-time performance of the algorithm when designed to be precise or with high accuracy. Thus, based on the given value of N = 7, we can conclude that the ideal values for the parameters of our experiment should meet the requirements of JT >= 0. 6.

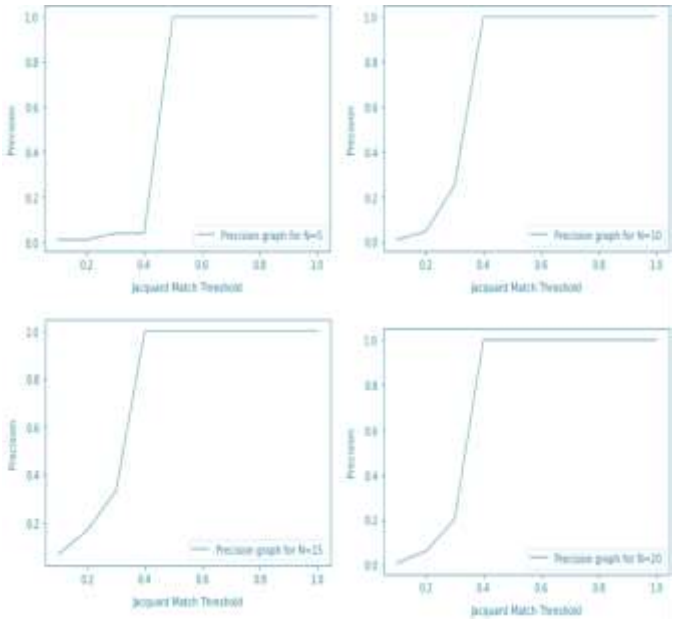

Fig 4: the length of the SuperMinHash signatures (N – number of bitmap features) determines the sensitivity of the PPRL approach to establish a match with the Jaccard threshold (JT).

4.4 Privacy analysis

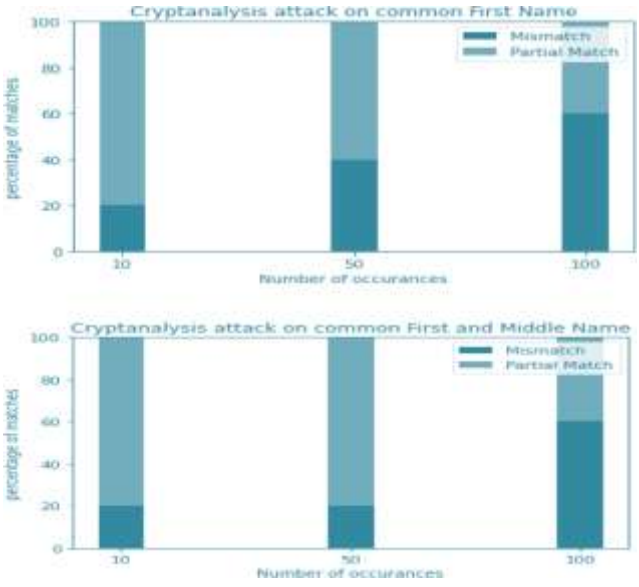

Fig 5: Even if a cryptoanalysis assault has been launched, it is evident that the effectiveness of the present protocol still meets acceptance criteria for real-world applications.

In [6], the authors and the present work adopted a frequency-based type of cryptanalysis method. The purpose of the study was in comprehending the extent to which the actual transfer of data between telecoms can be protected from similar onslaughts.

Security is again provided by the protocol that enhances secure transfer of data. Our experiment was actually conducted on the ‘‘North Carolina Voter record’’ database by sending them from our database to the Telcom B using the above protocol. Based on the assumption that Telecom B is a somewhat malicious but somewhat trustworthy enemy, we use a distinguishing shared first name and a distinguishing shared first and middle name to perform the cryptanalysis attack adopted by our experiment with 10, 50, and 100 occurrences. This cryptanalysis assault allowed the extraction of the data represented in the figure 5, where we can observe a zero frequency of RM matches and graphs illustrating the proportion of matches in part, and more specifically no match at all. Examining how we used first, middle, and last names that are commonly utilized in our protocol’s security was further evidenced. No matches were detected.

4.5 Comparison of our PPRL method against other implementations

To foster objectivity, four parameters have been proposed to compare are used to measure the performance of the proposed protocol to that of the other PPRL techniques.

The two articles [8, 9] were used for comparison while implementing with the help of Bloom filters. This made them more vulnerable to cryptanalysis attacks as debunked in [6].

Some more related problems include the amount of data processing overhead that arises due to noise in the differential privacy technique, which was mentioned in study [10], and the fact that it is difficult to implement several of these protocols. Propose to adopt Blm-DL, a deep learning [11], for Bloom filter data.

Linking: Using linkage units, entities that receive identification information from a solid source and return matching information. This will be displayed in our comparisons as Partition GC-RL.

The three criteria mentioned above will be evaluated under the ability to execute the protocol rapidly.

5. CONCLUSION

This problem is becoming rampant due to the inability of carriers to share information or have a proper method in handling cases of telecom fraud. Since many PPRL protocols base their matching on Bloom Filters, which are error-prone, or delegate their matching to third parties, we have outlined the primary weakness in current implementations which makes it relatively easy to induce data encoded with similar techniques. In view of this major concern affecting PPRL implementations, the study introduced EdDSA – a risk-free, efficient, and reliable DSS implementation of SuperMinHash – as a potent fraud detection tool for the telecommunication sector. We proved and demonstrated that it is only possible through our method to apply the Jaccard measure in order to compare two non-homologous sets in such a way that this comparison remains confidential. In a bid to know how fast PPRL is, we conducted some tests and evaluated our technique against existing approaches; we noted that our protocol performed better than all those approaches in terms of efficiency and security. Due to the fact that the protocol enables the real-time transfer of the fraud data without infringement of any privacy laws by the telecommunications companies, it constitutes a viable option. As we mentioned before, the present implementation only employs the indicators of fraud to determine if they are contained in the database; but the new proposed protocol should be designed in a way that it extracts further information without the invasion of the privacy of the user. In the long-run, the potential solution entails categorizing the fraud situations as low, medium, or high risk by down sampling a large number of MinHash methods and qualifying them with different fraud related terms. It is crucial for the fields related to healthcare to share data as well as maintain patients’ private information, and this can be applied to other areas.

REFERENCES

1. Babaei, Kasra, ZhiYuan Chen, and Tomas Maul. "A Study of Fraud Types, Challenges and Detection Approaches in Telecommunication." Journal of Information Systems and Telecommunication 7, no. 4 (2019): 248-261.
2. Vidanage, Anushka, Peter Christen, Thilina Ranbaduge, and Rainer Schnell. "A graph matching attack on privacy-preserving record linkage." In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 1485-1494. 2020.
3. Vatsalan, Dinusha, Peter Christen, and Erhard Rahm. "Scalable privacy-preserving linking of multiple databases using counting Bloom filters." In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 882-889. IEEE, 2016.
4. Kosub, Sven. "A note on the triangle inequality for the Jaccard distance." Pattern Recognition Letters 120 (2019): 36-38.
5. Meadows, Catherine. "A more efficient cryptographic matchmaking protocol for use in the absence of a continuously available third party." In 1986 IEEE Symposium on Security and Privacy, pp. 134-134. IEEE, 1986.
6. Christen, Peter, Thilina Ranbaduge, Dinusha Vatsalan, and Rainer Schnell. "Precise and fast cryptanalysis for Bloom filter based privacy-preserving record linkage." IEEE Transactions on Knowledge and Data Engineering 31, no. 11 (2018): 2164-2177.
7. Jaccard, Paul. "Lois de distribution florale dans la zone alpine." Bull Soc Vaudoise Sci Nat 38 (1902): 69-130.
8. Vatsalan, Dinusha, Ziad Sehili, Peter Christen, and Erhard Rahm. "Privacy-preserving record linkage for big data: Current approaches and research challenges." Handbook of big data technologies (2017): 851-895.
9. Randall, Sean M., Anna M. Ferrante, James H. Boyd, Jacqueline K. Bauer, and James B. Semmens. "Privacy-preserving record linkage on large real world datasets." Journal of biomedical informatics 50 (2014): 205-212.
10. Xue, Wanli, Dinusha Vatsalan, Wen Hu, and Aruna Seneviratne. "Sequence data matching and beyond: New privacy-preserving primitives based on bloom filters." IEEE Transactions on Information Forensics and Security 15 (2020): 2973-2987.
11. Ranbaduge, Thilina, Dinusha Vatsalan, and Ming Ding. "Privacy-preserving deep learning based record linkage." IEEE Transactions on Knowledge and Data Engineering (2023).