*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 3, Ser. I (May. – June. 2026), PP 98-105*
*www.iosrjournals.org*

# A Study to Enhance Multi-Party Record Linkage Through Secure Approximate Matching

Biplab Biswas
*Research Scholar, Department of Computer Science, RKDF University, Ranchi, India*

***Abstract***
*The multi-party Privacy-Preserving Record Linkage (PPRL) seeks to recognise and correlate same entities across various data sources from multiple parties, while safeguarding all confidential information from exposure, save for the ultimate matching outcomes exchanged among the parties. This research introduces a scalable Multi-Party Approximate Matching (MPAM) technique that facilitates secure record linking while safeguarding sensitive personal data. The suggested approach utilizes Bloom filter-based encoding and decentralised similarity calculations to provide effective approximation matching while maintaining anonymity. Empirical findings indicate that the suggested methodology attains nearly linear scalability as dataset sizes expand, markedly diminishes computing burden relative to current exact matching techniques, and preserves elevated linkage precision despite data corruption. Moreover, the protocol demonstrates improved privacy safeguards when the chance of exposure diminishes with a greater number of involved parties. The suggested framework adeptly harmonises efficiency, precision, and confidentiality, rendering it highly appropriate for safe data amalgamation in healthcare, government, finance, and other sectors necessitating collaborative scrutiny of sensitive data.*
***Keywords:*** *Multi-Party Record Linkage, Approximate Matching, Record, Scalability, Disclosure Risk*

## I. INTRODUCTION

With the ever-changing nature of the Internet, big data has emerged as a vital strategic asset that has far-reaching consequences in sectors as diverse as healthcare, business, and government. Data integration and connecting not only benefits many businesses, but also propels humanity forward in an always improving social fabric. Integrating and exchanging data is made easier by record linkage, a fundamental technique that reliably matches things from many data sources to the same real-world object. On the other hand, privacy and security issues are among the biggest obstacles to record linking. Due to the sensitive nature of the data (e.g., personal information, medical records, etc.), privacy protection concerns must be thoroughly considered and appropriately addressed throughout the data linking process to guarantee the data's security. A new technique called PPRL has arisen to deal with this problem.

To ensure security, PPRL strives for reliable record linking across different data sources without disclosing any further information about the data sources. The medical field provides this example. It is critical that patients' medical records be able to be shared and identified when they receive treatment at multiple healthcare facilities. Finding a way to provide just the user's medical condition information while protecting the patient's privacy is the main challenge of this approach. Sharing this data securely using PPRL technology improves medical quality and efficiency, increases accuracy in analysing health conditions, and safeguards patient privacy. Healthcare isn't the only industry that stands to gain from this technology's use; other sectors also have practical options for achieving data exchange while preserving privacy.

In theory, PPRL has a lot of potential, like protecting privacy in medical data, applying it to a national statistical institute, achieving efficient record linkage in large datasets, and protecting deidentified records within a public health surveillance system. However, in practice, it hasn't been without its limitations and challenges. The difficulty in finding a happy medium between security and computing performance is the primary cause of this. Data encoding and encryption are two regularly used data protection techniques in PPRL technology. While PPRL approaches based on encryption can offer high levels of security, they are notoriously computationally expensive, which makes them impractical for use with massive datasets. But Bloom Filter encoding (BFS) is commonly used in PPRL approaches that rely on encoding.

While BFS is great at increasing computing efficiency, it is vulnerable to assaults when it links many data sources. The likelihood of privacy leakage is increased due to these attacks, which include brute-force collisions on hash methods, frequency attacks, and password attacks on BFS. An increased risk of privacy leakage can occur in PPRL if attackers take use of the features of technologies like BFS to disclose personal identifying information through methods such as precise password analysis. It is now critically important to discover solutions to the pressing challenge of how to make PPRL technology more secure without sacrificing computing performance. This calls for technical optimisation and innovation as well as an increased awareness of the

fundamental needs of data privacy protection, which in turn calls for the creation of more robust and effective PPRL techniques.

## II. REVIEW OF LITERATURE

Han, Shumin et al., (2024) To ensure that parties communicate matching records for the same entity without releasing other sensitive data, privacy-preserving record linkage (PPRL) links records from different data sources. Unfortunately, the security of the linkage is not guaranteed by most existing PPRL techniques since they depend on third parties for connecting, which can lead to malicious manipulation and privacy breaches. So, to tackle the problem of semi-trusted third-party validation, we provide a consortium blockchain-based parallel multi-party PPRL technique that checks each participant in the PPRL procedure for signs of malicious interference. We offer a realistic Byzantine fault tolerance consensus method that is based on matching efficiency to enhance the efficiency and security of consensus inside a consortium blockchain. To further improve the security of Bloom filter encoding, we have used homomorphic encryption. Our goal is to maximise computing efficiency by implementing a binary storage tree for similarity calculation and using the MapReduce architecture for parallel encryption. Our solution successfully guarantees data security, exhibits relatively good linkage quality, and is scalable, according to the testing results.

Randall, Sean et al., (2022) Bloom filter-based privacy-preserving record linkage (PPRL) approaches have demonstrated potential for application in operational linkage contexts. Nevertheless, to verify their practicality, assessments in the actual world are necessary. Bloom filters were used to encrypt and connect records from the Western Australian (WA) Hospital Morbidity Data Collection 2011–2015 and the WA Death Registrations 2011–2015, while privacy-preserving procedures were employed. We were able to directly investigate the comparison step by comparing the results to a standard, un-encoded linking of the identical datasets using the same blocking criteria. In a blindfolded context, neither the unencoded data nor a "truth set" were available throughout the execution of the encoded linkage. Since 99.3% of the "groupings" between privacy-preserving and clear-text linkage were identical, the PPRL approach utilising Bloom filters produced linkage quality that was comparable to the standard un-encoded linkage. It seems that when clear-text IDs are not available for linking, the Bloom filter approach is a good fit.

Xue, Wanli et al., (2020) An effective masking method for privacy-preserving matching functions, bloom filter encoding has seen widespread application. Unfortunately, the current matching methods can only handle text, category, and signal integer values, which are quite simple kinds. Here, we provide a novel approach that vastly broadens the scope of matching primitives grounded in the privacy-preserving Bloom filter mechanism. Popular distance-based ML algorithms like KNN and SVM, as well as sequence data matching, are examples of such primitives. In order to efficiently assess the similarity of the data points with low utility loss, our technique adds a timestamp (bit) for each element in the data represented with its neighbouring values and hash-maps a sequence data vector into the Bloom filter space. To compensate for the limitations of encoding-based deterministic probability, it employs a Laplace-like perturbation approach on the built Bloom filters. This means that private data records are protected against discrimination in the event of collisions and differential privacy thanks to the planned effort. By directly introducing Laplace noise to the data, our technique achieves a far better utility-privacy trade-off than the state-of-the-art differential privacy-based method, according to experimental findings on three real-scenario based datasets.

Stammler, Sebastian et al., (2020) Record Linkage is useful in many real-world data analysis scenarios when there isn't a unique identifier for each record yet many datasets need to be connected at the record level. Patient medical records stored in different institutions' systems that must be connected based on uniquely identifying information such as name, DOB, or ZIP code. It may not be possible to outsource the record linking process to a reliable third party due to privacy regulations that forbid the transfer of this personally identifiable information (PII) across different institutions. In order to link relevant records while preventing the leaking of personally identifiable information (PII), we suggest using privacy-preserving record linkage (PPRL) approaches. Our results show that by utilising secure multi-party computing and the medical record keeping program Mainzelliste as our data source, we can build a framework for fault-tolerant PPRL. Even under the most basic assumptions of cryptographic security, our approach is certain that no personally identifiable information (PII) will ever fall into the wrong hands. Our solution has been tested in real-world networking environments and was shown to be feasible: linking a patient record to a database of 10,000 records takes 48s over a severely delayed (100ms) network connection or 3.9s with a low-latency connection.

Christen, Peter et al., (2019) The capability to recognise records that pertain to the same item across many databases is crucial in data analytics. Privacy-preserving record linkage (PPRL) refers to methodologies that connect records in a manner that, other from the pairs identified as matches, no confidential information on the entities in the associated databases is disclosed. A widely utilised method in Privacy-Preserving Record Linkage (PPRL) involves encoding confidential information into Bloom filters, facilitating approximate matching using character q-grams. PPRL with Bloom filter encoding has demonstrated both precision and scalability for

extensive datasets. Nonetheless, research indicates that Bloom filters employed for Privacy-Preserving Record Linkage (PPRL) are susceptible to assaults that might uncover sensitive information. Prior attack techniques were sluggish and necessitated an understanding of encoding settings; we provide an innovative assault that capitalises on the encoding of data within Bloom filters. Our assault does not need awareness of the encoding mechanism or its parameter configurations employed. It can accurately re-identify q-grams that were not hashable to specific Bloom filter bit locations; utilising these re-identified q-grams, it can then re-identify attribute values with great precision. Our approach surpasses previous approaches in speed and can effectively re-identify attribute values from extensive datasets within minutes.

Han, Shumin et al., (2017) The procedure of aligning and consolidating records pertaining to the same entity from many datasets is referred to as record linkage, which has gained significance across various domains such as commerce, public administration, and healthcare. The material from these regions frequently includes confidential data. To avert privacy violations, documents ought to be interconnected privately, ensuring that just the matching outcome is disclosed, a method referred to as privacy-preserving record linkage (PPRL). As data volume escalates, scalability emerges as the primary obstacle for PPRL, prompting the development of many private blocking methodologies for PPRL. Their objective is to decrease the quantity of record pairs subjected to comparison in the matching procedure by eliminating evident non-matching pairings while safeguarding privacy. Nonetheless, the majority of them are tailored for dual databases and exhibit significant variation in their capacity to reconcile the conflicting objectives of precision, efficacy, and security. In this manuscript, we introduce an innovative private blocking method for PPRL utilising dynamic k-anonymous blocking and the Paillier cryptosystem, applicable to two or more databases. In dynamic k-anonymous blocking, our methodology actively creates blocks that fulfil k-anonymity while providing more precise data to characterise the blocks with differing k. We furthermore provide an innovative similarity measurement technique that operates on numerical features and integrates with the Paillier cryptosystem to assess the similarity of several blocks securely, ensuring robust privacy protections that prevent any information disclosure, even in cases of collusion. Investigations performed on a publicly available dataset of voter registration records confirm that our methodology is adaptable to extensive datasets while maintaining superior blocking quality. We evaluate our approach against alternative methods and illustrate the enhancements in security and precision.

Niedermeyer, Frank et al., (2014) Bloom filter encoded IDs are progressively utilised in privacy-preserving record linkage applications, as they accommodate inaccuracies in encrypted identifiers. Nonetheless, scant investigation on the security of Bloom filters has been released to far. This work delineates an effective assault on Bloom filters constructed using bigrams. Prior scholarship has posited that an assailant have knowledge of the overarching data collection from which a sample is extracted. Conversely, we assume that an assailant is unaware of this comprehensive data collection. Conversely, we presume that the antagonist is aware of a publicly accessible compilation of the most prevalent traits. The assault relies on nuanced filtering and fundamental statistical examination of encrypted bigrams. The assault detailed in this document can facilitate the decryption of a whole database rather than only a limited selection of the most prevalent names, as shown in other studies. We demonstrate our suggested approach by an assault on a database containing encrypted last names. Ultimately, we delineate alterations to the Bloom filters aimed at thwarting analogous assaults.

Schnell, Rainer et al., (2009) The amalgamation of many databases containing disparate or supplementary information about the same individual is becoming more prevalent in research. In the absence of unique identifying numbers for these persons, probabilistic record linking is employed to identify corresponding record pairings. In several applications, IDs must be encrypted to address privacy issues. A novel approach for privacy-preserving record linking utilising encrypted identifiers that accommodates mistakes in identifiers has been established. The protocol relies on Bloom filters utilising q-grams of identifiers. Experiments conducted on both simulated and real databases produce linking outcomes that are equivalent to those obtained with non-encrypted IDs and exceed the results derived from phonetic encodings. We suggested a system for privacy-preserving record linking utilising encrypted identifiers that accommodate mistakes in those identifiers. Given that the protocol is readily augmentable and possesses minimal processing demands, it may be advantageous for several applications necessitating privacy-preserving data linking.

## III. EXPERIMENTAL SETUP

This section evaluates the proposed Multi-Party Approximate Matching (MPAM) protocol based on three key PPRL criteria: scalability, linkage quality, and privacy preservation. The proposed approach is compared with the exact matching method of Lai et al., which serves as the baseline. All experiments were conducted on a server with four 6-core 64-bit Intel Xeon 2.4 GHz processors, 128 GB RAM, and Ubuntu 14.04. Both methods were implemented in Python 2.7.3. The experimental parameters were set as l = 500, k = 20, q = 2, s<sub>t</sub> = 0.8, and P = {3, 5, 7}. A Soundex-based phonetic blocking technique was used to generate candidate blocks, which were then processed independently using the proposed approximate matching protocol for secure record comparison and classification.

**Datasets**

To evaluate the proposed approach under realistic conditions, experiments were conducted using the North Carolina Voter Registration (NCVR) database, which contains records of over 8 million registered voters. Following previous PPRL studies, the attributes given name, surname, suburb, and postcode were selected as quasi-identifiers (QIDs). Since no publicly available real-world dataset exists for multi-party record linkage, multiple party-specific datasets were generated by partitioning the NCVR data into subsets representing different organizations.

Dataset subsets containing 5,000, 10,000, 50,000, 100,000, 500,000, and 1,000,000 records were created for 3, 5, and 7 parties, with 50% overlapping records shared across all parties. To simulate real-world data quality issues, a data corruption tool was used to introduce one to three random modifications, including character insertions, deletions, substitutions, transpositions, optical character recognition (OCR) errors, and phonetic variations. The identities of all modified records were tracked to accurately distinguish true matches from false matches during evaluation.

For each dataset, additional versions were generated with 0%, 20%, and 40% corrupted overlapping records to assess the robustness of the proposed approximate matching protocol. This setup reflects practical scenarios in which some organizations maintain accurate records while others contain incomplete or inconsistent information, enabling a comprehensive evaluation of scalability, linkage accuracy, and privacy preservation under varying data quality conditions.

**Evaluation Measures**

The proposed multi-party PPRL protocol was evaluated using three key performance criteria: scalability, linkage quality, and privacy preservation. Scalability was assessed by measuring the runtime and memory consumption required for the linkage process. In addition, the effectiveness of the proposed filtering strategy (MPAM-F) was evaluated using the Filtering Reduction Ratio (RRf), which measures the reduction in candidate record comparisons.

Linkage quality was measured using the F-measure (F1-score), the harmonic mean of precision and recall, providing a balanced assessment of matching accuracy. Privacy preservation was evaluated using the Disclosure Risk (DR) metric based on the probability of suspicion (ps), which estimates the likelihood of correctly linking a masked record to records in a publicly available database. To assess resistance against privacy attacks, a frequency linkage attack was simulated under the worst-case scenario, where the global database was assumed to be identical to the linkage database. The attack attempted to match exposed Bloom filter bit patterns, enabling an evaluation of the protocol's resilience against re-identification while maintaining accurate record linkage.

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

Figures 1 and 2 present the scalability performance of the proposed MPAM protocol in terms of runtime, memory consumption, and the number of candidate record sets processed during linkage. The results show that the execution time increases gradually as the number of participating parties increases and exhibits an approximately linear relationship with dataset size, demonstrating good scalability. In contrast, the memory requirement per party decreases with an increasing number of parties because each participant processes only a smaller Bloom filter segment of size (l/P), and the similarity computations are distributed across all parties. However, the overall memory consumption increases for larger datasets due to the growth in candidate record sets, despite the use of Soundex-based phonetic blocking and the proposed MPAM-F filtering strategy, both of which effectively reduce unnecessary record comparisons.

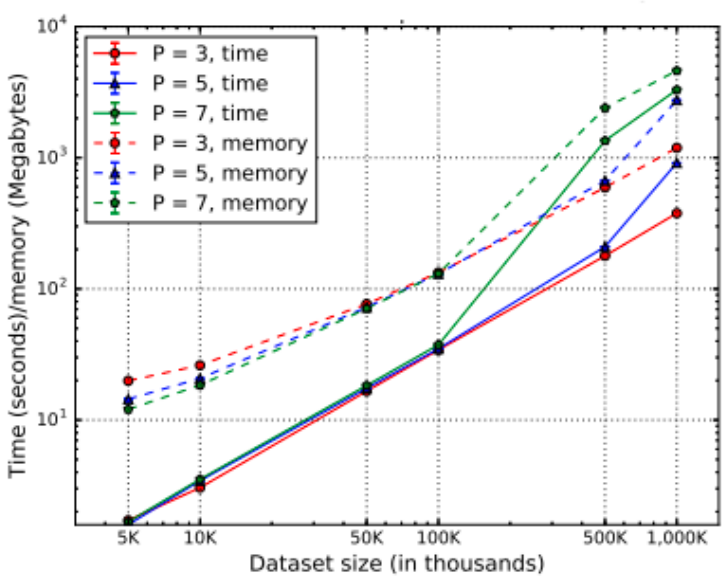


**Figure 1: Performance Analysis of Record Linkage Based on Time and Memory Utilization**

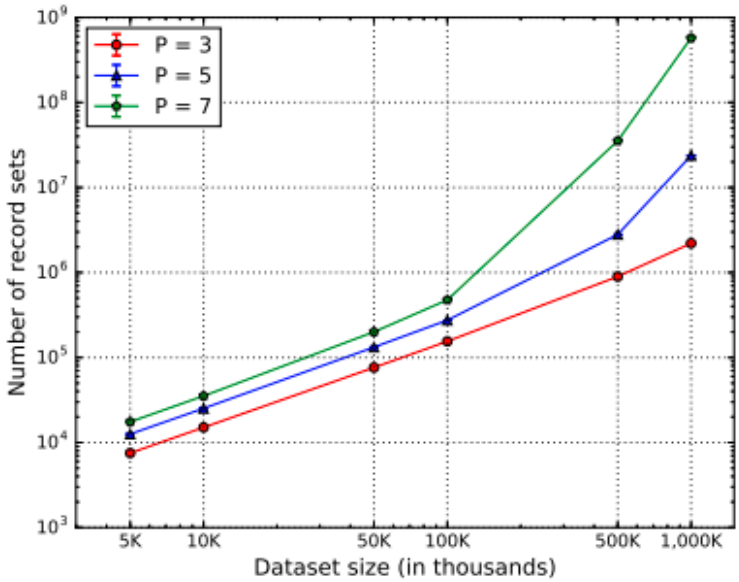


**Figure 2: Comparison of Candidate Record Sets for Varying Dataset Sizes**

Figure 3 illustrates the Filtering Reduction Ratio (RRf) achieved by the proposed MPAM-F strategy. For smaller datasets, the reduction in candidate record comparisons is relatively limited due to the lower number of records. However, as the dataset size increases, the filtering strategy becomes significantly more effective, resulting in a higher RRf. The results also show that RRf improves with an increasing number of participating parties (P), demonstrating the scalability and efficiency of the proposed filtering approach. These findings highlight the potential for further improvements through optimized communication strategies, advanced blocking and indexing techniques, and enhanced multi-party PPRL methods.

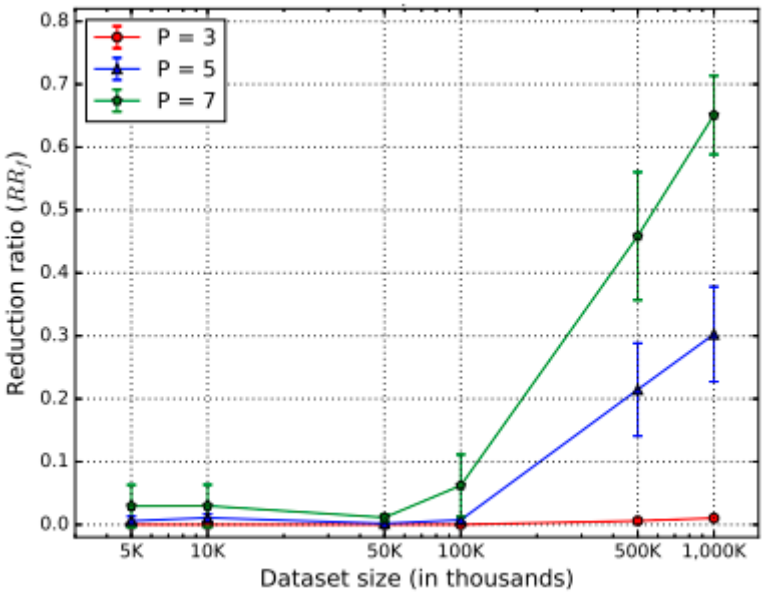


**Figure 3: Comparison of Filtered Record Set Reduction Ratios for Different Dataset Sizes**

Figure 4 compares the proposed MPAM protocol with the baseline method of Lai et al. in terms of runtime and memory consumption. The results demonstrate that the proposed approach provides superior scalability and computational efficiency, particularly as the number of participating parties increases. Unlike the baseline method, which requires each party to perform additional membership tests on its Bloom filters after identifying common bit patterns, the proposed protocol computes record similarities centrally using the aggregated counts of Bloom filter 1-bits and shared 1-bits, and then distributes the results to all parties. This design significantly reduces computational overhead and communication costs. Consequently, both the execution time and memory requirements of the proposed method increase only marginally with the number of participating parties, whereas the baseline approach exhibits a much steeper growth in computational and memory demands.

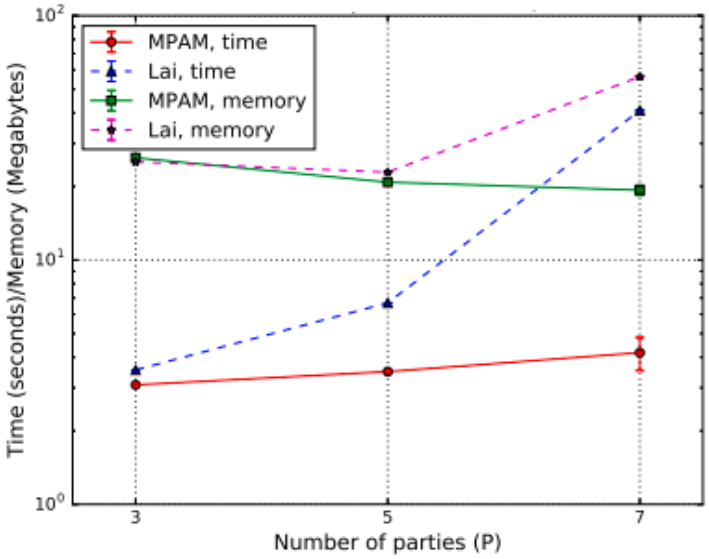


**Figure 4: Comparison of Runtime and Memory Requirements for Record Linkage across Different Numbers of Parties**

Figure 5 compares the linkage performance of the proposed MPAM and MPAM-F methods with the baseline approach of Lai et al. using the F-measure (F1) on the NCVR-10,000 datasets. The results indicate that all methods achieve high F1-scores on the original datasets without data corruption. However, as the number of participating parties and the level of data corruption (20% and 40%) increase, the F1-score decreases due to the growing number of modified records, which reduces the likelihood of identifying true matches. The proposed MPAM-F method achieves linkage quality comparable to MPAM, demonstrating that the filtering strategy has

only a minimal impact on matching accuracy while effectively reducing unnecessary comparisons. On the original datasets, MPAM-F slightly improves the F1-score by eliminating false candidate matches.

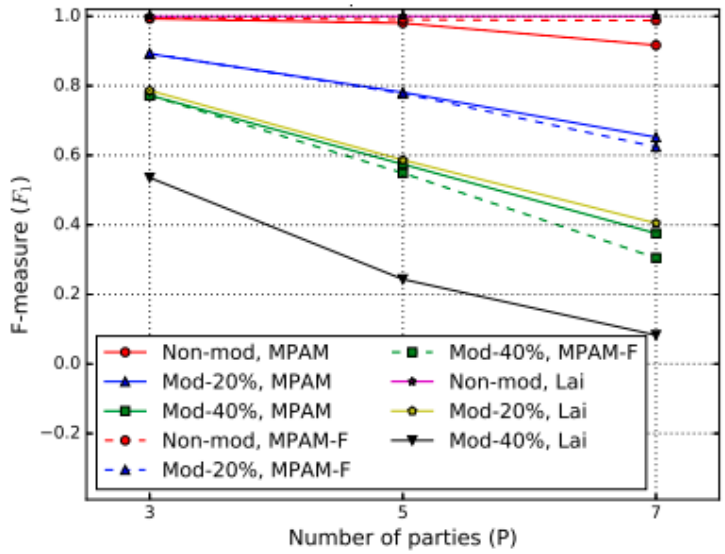


**Figure 5: Comparison of F Measure for Record Linkage across Different Numbers of Parties**

Figure 6 presents the privacy performance of the proposed protocol using Disclosure Risk (DR) metrics, including mean disclosure risk and marketer disclosure risk, under a frequency linkage attack on the NCVR-10,000 datasets. The evaluation assumes the worst-case scenario, where the global database is identical to the linkage database (G ≡ NCVR-10,000). The results show that disclosure risk decreases as the number of participating parties increases, indicating stronger privacy protection. This improvement is achieved because the Bloom filter segments become shorter (l/P), making each segment correspond to a larger number of possible records in the global database. Consequently, the probability of suspicion (ps = 1/ng) decreases, resulting in lower disclosure risk and enhanced resistance against re-identification attacks.

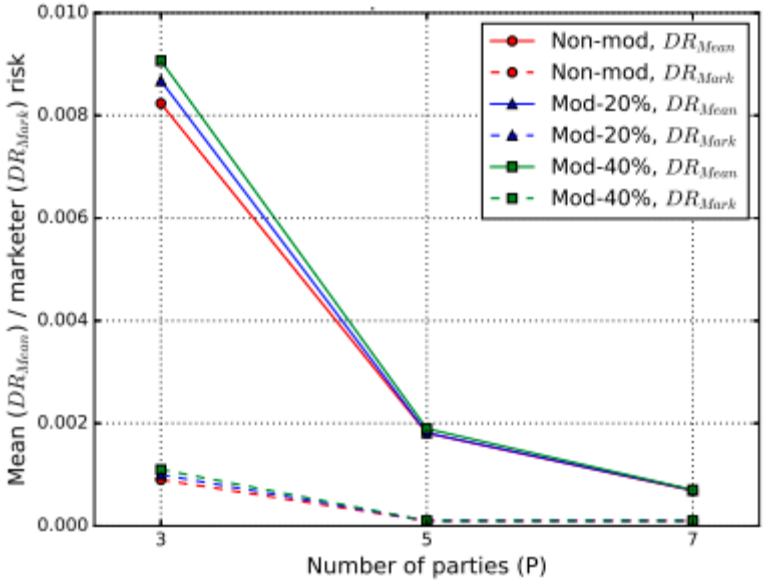


**Figure 6: Comparative Evaluation of Disclosure Risk Measures in Multi Party Record Linkage**

## V. CONCLUSION

By striking a good balance between scalability, linkage accuracy, and privacy protection, the proposed Multi-Party Approximate Matching (MPAM) protocol offers a safe and fast solution for Privacy-Preserving Record Linkage. The experimental findings show that the framework may reduce memory needs and computational overhead while still achieving great matching performance, especially on large-scale datasets. Secure data integration across different organisations is possible with the help of the Bloom filter-based

approximation matching method, which effectively deals with data discrepancies and reduces disclosure risk. Medical, governmental, and financial sectors are only a few examples of data-intensive industries that can benefit from the suggested protocol's dependable and practical foundation for privacy-preserving record linking.

## REFERENCES:


[1]. S. Han, Z. Wang, D. Shen, and C. Wang, "A parallel multi-party privacy-preserving record linkage method based on a consortium blockchain," *Mathematics*, vol. 12, no. 12, pp. 1–26, 2024.

[2]. B. Ünal, N. Pfeifer, and M. Akgün, "A privacy-preserving approach for cloud-based protein fold recognition," *Patterns*, vol. 5, no. 9, pp. 101–123, 2024.

[3]. S. Randall, H. Wichmann, A. Brown, J. Boyd, T. Eitelhuber, A. Merchant, and A. Ferrante, "A blinded evaluation of privacy preserving record linkage with Bloom filters," *BMC Medical Research Methodology*, vol. 22, no. 1, pp. 1–7, 2022.

[4]. S. Stammler, T. Kussel, P. Schoppmann, F. Stampe, G. Tremper, S. Katzenbeisser, K. Hamacher, and M. Lablans, "Mainzelliste SecureEpiLinker (MainSEL): Privacy-preserving record linkage using secure multi-party computation," *Bioinformatics*, vol. 38, no. 6, pp. 1657–1668, 2022.

[5]. W. Xue, D. Vatsalan, W. Hu, and A. Seneviratne, "Sequence data matching and beyond: New privacy-preserving primitives based on Bloom filters," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2973–2987, 2020.

[6]. P. Christen, T. Ranbaduge, D. Vatsalan, and R. Schnell, "Precise and fast cryptanalysis for Bloom filter based privacy-preserving record linkage," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 11, pp. 2164–2177, 2019.

[7]. D. Karapiperis, A. Gkoulalas-Divanis, and V. S. Verykios, "FEDERAL: A framework for distance-aware privacy-preserving record linkage," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 2, pp. 292–304, 2018.

[8]. S. Han, D. Shen, T. Nie, Y. Kou, and G. Yu, "Private blocking technique for multi-party privacy-preserving record linkage," *Data Science and Engineering*, vol. 2, no. 3, pp. 187–197, 2017.

[9]. E. A. Durham, M. Kantarcioglu, Y. Xue, C. Toth, M. Kuzu, and B. Malin, "Composite Bloom filters for secure record linkage," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2956–2968, 2014.

[10]. F. Niedermeyer, S. Steinmetzer, M. Kroll, and R. Schnell, "Cryptanalysis of basic Bloom filters used for privacy preserving record linkage," *Journal of Privacy and Confidentiality*, vol. 6, no. 2, pp. 59–79, 2014.

[11]. D. Vatsalan, P. Christen, and V. S. Verykios, "A taxonomy of privacy-preserving record linkage techniques," *Information Systems*, vol. 38, no. 6, pp. 946–969, 2013.

[12]. R. Schnell, T. Bachteler, and J. Reiher, "Privacy-preserving record linkage using Bloom filters," *BMC Medical Informatics and Decision Making*, vol. 9, no. 1, pp. 1–11, 2009.

[13]. S. Trepetin, "Privacy-preserving string comparisons in record linkage systems: A review," *Information Security Journal*, vol. 17, nos. 5–6, pp. 253–266, 2008.

[14]. M. Eichelberg, T. Aden, and W. Thoben, "A distributed patient identification protocol based on control numbers with semantic annotation," *International Journal on Semantic Web and Information Systems*, vol. 1, no. 4, pp. 24–43, 2005.

[15]. J. J. Berman, "Zero-check: A zero-knowledge protocol for reconciling patient identities across institutions," *Archives of Pathology & Laboratory Medicine*, vol. 128, no. 3, pp. 344–346, 2004.

[16]. R. Schnell, T. Bachteler, and S. Bender, "A toolbox for record linkage," *Austrian Journal of Statistics*, vol. 33, nos. 1–2, pp. 125–133, 2004.

[17]. A. Broder and M. Mitzenmacher, "Network applications of Bloom filters: A survey," *Internet Mathematics*, vol. 1, no. 4, pp. 485–509, 2003.

[18]. G. Gatta, R. Capocaccia, M. Sant, C. M. J. Bell, J. W. W. Coebergh, R. A. M. Damhuis, J. Faivre, C. Martinez-Garcia, J. Pawlega, M. Ponz de Leon, D. Pottier, N. Raverdy, E. M. I. Williams, and F. Berrino, "Understanding variations in survival for colorectal cancer in Europe: A EUROCARE high-resolution study," *Gut*, vol. 47, no. 4, pp. 533–538, 2000.