Towards Auditable and Intelligent Privacy-Preserving Record Linkage

Thiago Nóbrega[1], **Carlos Eduardo S. Pires**[1], **Dimas Cassimiro Nascimento**[1]

[1]Programa de Pós-Graduação em Ciência da Computação
da Universidade Federal de Campina Grande (PPGCC/UFCG)

thiagonobrega@gmail.com, cesp@dsc.ufcg.edu.br, dimas.cassimiro@ufape.edu.br

Abstract. *Privacy-Preserving Record Linkage (PPRL) intends to integrate private/sensitive data from several data sources held by different parties. It aims to identify records (e.g., persons or objects) representing the same real-world entity over private data sources held by different custodians. Due to recent laws and regulations (e.g., General Data Protection Regulation), PPRL approaches are increasingly demanded in real-world application areas such as health care, credit analysis, public policy evaluation, and national security. As a result, the PPRL process needs to deal with efficacy (linkage quality), and privacy problems. For instance, the PPRL process needs to be executed over data sources (e.g., a database containing personal information of governmental income distribution and assistance programs), with an accurate linkage of the entities, and, at the same time, protect the privacy of the information. Thus, this work intends to simplify the PPRL process by facilitating real-world applications (such as medical, epidemiologic, and populational studies) to reduce legal and bureaucratic efforts to access and process the data, making these applications' execution more straightforward for companies and governments. In this context, this work presents two major contributions to PPRL: i) an improvement to the linkage quality and simplify the process by employing Machine Learning techniques to decide whether two records represent the same entity, or not; and ii) we enable the auditability the computations performed during PPRL.*

1. Introduction

In recent times the companies and government significantly increased the amount of the collected data. Much of this data is about personal information, such as shopping transactions, browsing history, telecommunication records, financial information, or electronic health records. This data has been employed in data mining and analytic techniques that can provide relevant information for several areas of knowledge. For instance, personal data can i) be employed to perform crime and fraud detection [6], ii) lead to better patient outcomes or to detect a disease outbreak in the health sector [1], iii) be of vital importance to national security [8] or be a competitive edge to a commercial enterprise[2].

Data mining and analysis often require information from multiple data sources to be integrated in order to enable precise and useful analysis [1]. However, to execute data integration, first, we have to identify and aggregate records that relate to the same entity (e.g., people, restaurants, publications, products, among others) from one or more data sources [2]. This process is known as Record Linkage (RL), Data Matching (DM),

or Entity Resolution (ER) [2, 1]. Although the process receives several names in the literature, in this work, we will adopt RL.

The RL process is composed of four major steps. The first one is data preprocessing, which ensures the data from several data sources are in the same format. The second step, indexing, intends to reduce the number of comparisons performed by selecting entity pairs to be matched (compared) in the subsequent step. In the third step, the actual entity pair comparison occurs. In the comparison step, each entity pairs receives a similarity value. These pairs are compared using various attributes (for a person, it can include name, sex, and age) and comparison functions. Finally, in the last step (classification), the record pairs are classified into matches, non-matches, and potential matches, depending on the decision model used [1].

A recurring problem that Record Linkage faces is the absence of attributes capable of uniquely identifying entities, which refer to the same entity, in the different data sources. The absence of a unique identifier, such as an ID, makes the use of simple comparison operations (e.g. SQL joins) impossible, making the linkage to be carried out with sophisticated comparisons involving a set of common attributes to all entities in the different data sources. Such a set of attributes is called quasi-identifiers (QIDs) [2].

Currenttly, Record Linkage not only faces computational and operational challenges intrinsic to the comparison and classification methods, but it also has to address privacy preservation challenges due to recent laws and regulations such as European General Data Protection Regulation (GDPR), Brazilian General Data Protection Law (LGPD) and the US HIPAA Privacy Rule. In this context, Privacy-Preserving Record Linkage (PPRL) emerges, aiming to identify matching entities across private data sources, ensuring that the data's privacy and confidentiality are preserved throughout the linkage process.

In order to address privacy-related issues the basic idea of Privacy-Preserving Record Linkage (PPRL) is to execute the linkage process in anonymized data (by perturbating the original data with the use of encryption, hash functions, and noise additions), ensuring that the privacy and confidentiality of the data are preserved during the linkage process. PPRL reveals only a limited amount of information. For instance, a party only knows which of its own records exist in the other party's data source or the number of duplicated entities presented in the datasets used as input to the PPRL process [8].

A PPRL solution needs to address two issue (or characteristics): privacy, and linkage quality. In the following, we outline the PPRL characteristics.

1. **Privacy**: in order to fulfill privacy-preserving requirements, PPRL solutions employ sophisticated anonymization techniques (e.g., homomorphic encryption and Bloom Filter) to preserve the privacy of the entities at a linkage quality level and an extra computational cost. However, the use of the anonymization techniques do not guarantee information privacy, several privacy attacks are able to break the privacy of anonymized data. Therefore, the use of privacy-preserving protocols along with anonymization techniques is required to ensure privacy during the PPRL process;
2. **Linkage Quality**: in general, real-world data sources are 'dirty' [6], which means they contain errors, typos, variations and values that could be missing. For in-

stance, the name 'Anna Estella' could be entered as 'Ane Stela' by a hospital employee, making it hard to link patient data across different data sources. Therefore, the exact comparison of QID values is not sufficient to achieve accurate linkage results. Thus, to improve the linkage quality, the use of approximate comparison techniques[1], as well as accurate classification techniques, are needed to achieve accurate linkage quality in record linkage applications. These quality problems are exacerbated due to the privacy requirements, i.e., anonymized QIDS. Thus, every PPRL process needs to address the linkage quality issues.

For a PPRL solution to be used in real-world applications, it should address these two characteristics. Furthermore, the PPRL solution needs to provide a comprised among privacy, and quality according to the needs of the PPRL parties' requirements. There have been many different approaches proposed for PPRL [5, 7, 8]. However, some approaches attempt to address the problem of PPRL fall short in providing a reliable solution, either because they do not provide sufficient privacy capabilities or because they cannot provide high linkage quality.

2. Limitations of Privacy-Preserving Record Linkage

As previously introduced, PPRL needs to address two issues. However, it is worthwhile to mention that Efficiency, Quality, and Privacy are conflicting. In other words, if a PPRL solution prioritizes one of these three characteristics, the other two will suffer. For instance, if we employ a complex anonymization, such as Homomorphic Encryption [3], technique we add an extra computational cost in every comparison. Furthermore, we force the linkage process to be carried based only on exact comparisons due to encryption limitations [6]. Therefore, the exact comparisons have an impact on the linkage quality because the QID's values need to be the same for a pair of entities to be considered a match; for example, the entities' ana' and 'Ana' are classified as "no match" by exact comparisons techniques.

While PPRL techniques help overcome the privacy-preserving linkage of sensitivity data, they present their own problems. Recent surveys [2, 5, 7, 8, 6] indicate that the main challenges for the extensive use of PPRL are related to the linkage quality and privacy issues. In the following, we outline some of the high-level challenges of the PPRL that are marked as open issues by the literature:

- **New adversarial models**: the parties PPRL need to make assumptions about the behavior of the other parties, and this assumption is named as adversarial models. The currently used adversarial models require that the PPRL parties fully trust other parties [2]. However, this adversarial model is not realistic for real-world applications [6], mainly because it is hard to find PPRL parties that will not try to learn from the exchange information. Therefore, the need for a more realistic adversarial model is an open issue to the PPRL community;
- **Anonymization techniques**: many of the anonymization techniques used in the PPRL process currently lack evidence that verifies whether these techniques cannot be attacked by an adversary, such as phonetic encoding and generalization

[1] Approximate comparison techniques return the degree of similarity among two entities, a number between 0 and 1, where 0 means dissimilarity and one total similarity. For instance, if we employ an approximate comparison technique over the 'Anna' and 'Ane' example, it will return a value of .75, indicating that the names are 75% similar, while the exact comparison will indicate that 'Anna' and 'Ane' are not similar.

techniques [6]. On the other hand, those techniques based on secure multiparty computation and encryption, while probably secure, are currently less scalable to link large data sources. Thus, in order to improve the linkage, novel anonymization techniques are required that are more secure than current approaches while still efficient and accurate, allowing the approximate comparisons of the QIDs values [2];

- **PPRL classification**: most PPRL solutions employ a simple classifier. In order to classify the entity pairs, the PPRL parties define a threshold and compare it against the value that represents the similarity calculated for an entity pair. However, the threshold value definition is a complex task that requires expert operators to "guess" the appropriate value. For instance, if the threshold value is too high (e.g., 0.9 or 1), PPRL will miss true match entities. On the other hand, if this value is too low, PPRL will likely classify false positive matches. Therefore, novel classification techniques are required in order to help the PPRL operators to classify the entities correctly.

Unless progress is made along with these issues mentioned above, it will not be easy to employ PPRL in real-world data. Next, we present the aims of our research.

3. Aim of Research

This work intends to address the PPRL process's bottlenecks that represent limitations to extensive use of PPRL in a real-world application. Therefore, this work's main goal is **to improve the PPRL in order to make it more suitable to be used in real-world applications**. The work's central hypothesis is to use our contributions to improve the process and tackle some of the existing PPRL bottlenecks.

3.1. Specific goals

Considering the proposed main goal and the fact that the privacy, efficiency and quality of linkage issues are the most limiting PPRL characteristics to widespread use of real-world applications, this work has the following specific goals:

1. Improve the privacy-preserving capabilities of the Bloom Filter (BF) anonymization technique;
2. Propose a novel adversary model that reduces the need of trust by PPRL parties;
3. Propose a machine learning-based classifier to PPRL;
4. Investigate the usage of PPRL in real-world applications.

4. Research Contributions

In order to illustrate our contributions to the PPRL process, we plotted Figure 1. It depicts the PPRL steps, highlighting the steps directly impacted by our contributions. Notice that the figure illustrates a general workflow for two dataset owners. We would like to state that our contribution can be applied in different scenarios, including multi-party PPRL.

Notice that we propose a contribution to the Anonymization step. This step is critical to the entire PPRL process, impacting the privacy, quality, and efficiency of the PPRL. The majority of the PPRL process considers the Bloom Filter (BF) anonymization technique. The BF is able to produce an accurate similarity distance between two entities'.

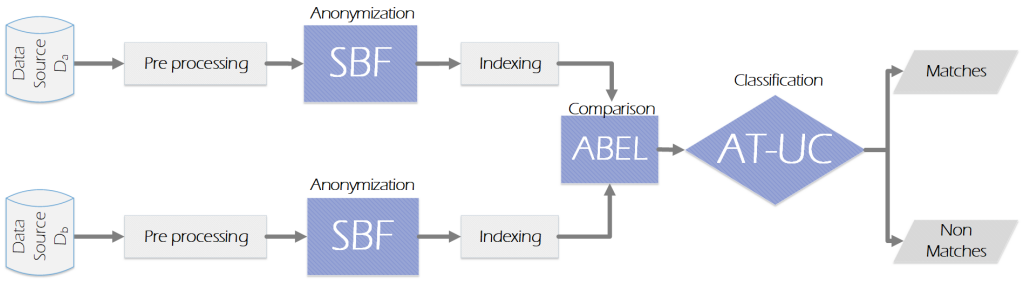

Figure 1. Our contributions within the PPRL process.

However, recent studies [4, 9] demonstrate that if an attacker has access to a complete database anonymized with this technique, he/she can re-identify the entities, breaking the privacy of the information.

In this context, we propose the Splitting Bloom Filter (SBF). The SBF aims to enable an iterative comparison of the entities' similarity by breaking the entities' anonymized representation in splits regarding the BF privacy enhance technique. In other words, the SBF modifies the anonymization step's output to enable the auditability in the PPRL comparison step, our second contribution.

In the Comparison step lays our second contribution. A major deficiency in the PPRL context is that the PPRL party needs to consider an unrealistic adversary model. The majority of the PPRL solutions assume an Honest-But-Curious (HBC) adversary model. This adversary model assumes that all PPRL parties will follow a pre-agreed protocol and will not try to re-identify the anonymized information exchanged during the PPRL. Therefore, having such trust in the PPRL context is unrealistic [2, 5].

To address the issue mentioned above, we propose the Auditable Blockchain-Based PPRL (ABEL) to provide auditability during the comparison step, eliminating the need to trust the other PPRL parties fully. Moreover, ABEL enables the auditability of the entity's similarity computation using Blockchain technology, with on-chain and off-chain computations. It is worthwhile to mention that the Blockchain stores all processed data on-chain to provide a transparent and temper evident computation. However, this Blockchain characteristic, in a PPRL context, poses a threat to entities' privacy. The usage of off-chain computation by the parties is a fundamental aspect to preserve the privacy of entities during the PPRL execution.

Our third contribution is placed in the Classification step of the PPRL. Due to privacy limitations, the classification step i) can not be performed or assisted by humans (oracle), and ii) there is no available label data, making it hard to train Machine Learning (ML) classifiers. The majority of the PPRL process utilizes a simple threshold (guessed by a specialist) to define whether an entity pair is a match, or not. It is worthwhile to remark that PPRL is used in law enforcement and medical applications, and an erroneous classification of the PPRL could have a serious outcome to a person. For instance, an innocent man could be flagged as a criminal, or a physician could prescribe the wrong treatment to a patient.

In this context, we propose the Auto-Tuned Unsupervised Classification approach (AT-UC) to provide the PPRL with better classifiers; eliminating the need for a specialist to guess a threshold and improve the linkage quality. The AT-UC utilizes a Transfer

Learning technique to employ non-private datasets for training and modify a classifier to be executed in a private dataset to tackle the absence of labeled data. Moreover, the AT-UC also has to define a proper feature space, select a related dataset, and modify the classifier.

In summary, our research intends to improve the PPRL process to make it more suitable for real-world applications. Until this point of the research, we focus mainly on linkage-quality and privacy bottlenecks of the PPRL.

5. Partial Results

In order to summarize our partial results, we arrange our contribution into two groups, contributions to privacy and quality of the PPRL process. Regarding the contribution to the privacy of PPRL, the SBF and ABEL could provide an auditability to the comparison step (covert adversary model). Furthermore, it could improve PPRL privacy by reducing the amount of information shared during linkage execution. It is worth mentioning that our contribution could reduce the linkage quality, as demonstrated by the F1 in Figure 1.

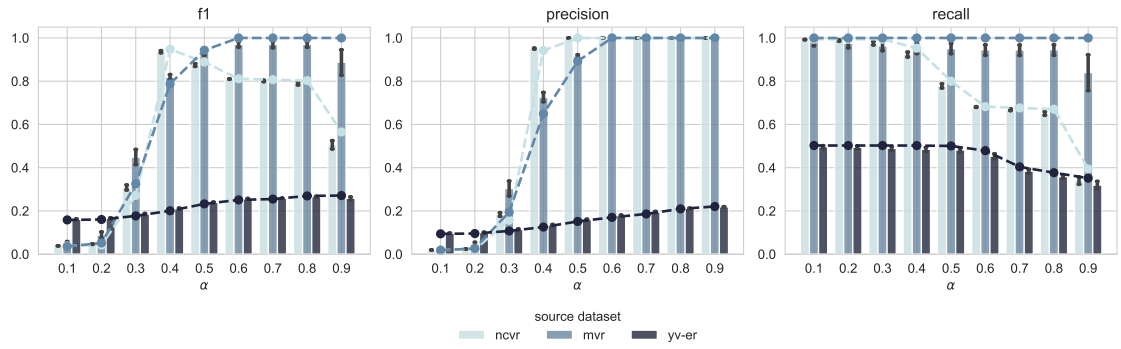

Figure 2. Effectiveness results of ABEL for different α values.

To improve the PPRL quality, we propose the AT-UC, which enables the adoption of an ML-based classifier to the classification step of PPRL. In other words, our contribution could improve the quality and simplify the PPRL process. Moreover, our contributions overcome the quality result of baseline (threshold-based) and two competitors, which do not consider the entities' privacy. Figure 2 outlines our results.

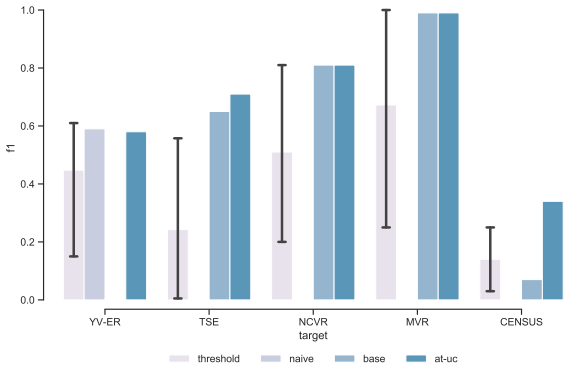

Figure 3. AT-UC quality results.

In summary, we provide contributions that: i) enable the usage of a novel adversary model and ii) improve the linkage quality by proposing an automatic classification

approach to the PPRL process. Moreover, besides privacy and quality improvements, our work impacts the adoption (usability) of PPRL by companies and governments by reducing the level of thurst to execute the PPRL and eliminating the need for an expert to define the classification threshold.

References

[1] Carlo Batini and Monica Scannapieco. *Data and Information Quality*. Data-Centric Systems and Applications. Springer International Publishing, 1 edition, 2016.

[2] Peter Christen, Thilina Ranbaduge, and Rainer Schnell. *Linking Sensitive Data*. Springer International Publishing, Cham, 2020.

[3] Thiago P. Nóbrega, Carlos E. S. Pires, and Tiago Brasileiro Araujo. Avaliação Empírica de Comparações Privada Aplicadas na Resolução de Entidades. In *SBBD*, 2016.

[4] Thilina Ranbaduge and Peter Christen. Privacy-Preserving Temporal Record Linkage. *IEEE International Conference on Data Mining*, pages 377–386, 2018.

[5] Dinusha Vatsalan, Dimitrios Karapiperis B, and Aris Gkoulalas-divanis. *An Overview of Big Data Issues in PPRL*, volume 2. Springer International Publishing, 2019.

[6] Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969, 2013.

[7] Dinusha Vatsalan, Dimitrios Karapiperis, and Vassilios S Verykios. Privacy-Preserving Record Linkage. (January), 2018.

[8] Dinusha Vatsalan, Ziad Sehili, Peter Christen, and Erhard Rahm. Privacy-Preserving Record Linkage for Big Data : Current Approaches and Research Challenges. In *Big Data Handbook*. Springer, 2016.

[9] Anushka Vidanage, Thilina Ranbaduge, Peter Christen, and Rainer Schnell. Efficient Pattern Mining based Cryptanalysis for Privacy-Preserving Record Linkage. *Proceedings - International Conference on Data Engineering*, pages 1698–1701, 2019.