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@uepb.edu.br, 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. In this context, our work presents contributions to improve the privacy and quality capabalities of the PPRL. Moreover, we propose 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 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 [Vatsalan et al. 2013a], ii) lead to better patient outcomes or to detect a disease outbreak in the health sector [Batini and Scannapieco 2016], iii) be of vital importance to national security [Vatsalan et al. 2016] or be a competitive edge to a commercial enterprise [Christen et al. 2020].

Data mining and analysis often require information from multiple data sources to be integrated in order to enable precise and useful analysis [Batini and Scannapieco 2016]. 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

[Christen et al. 2020]. This process is known as Record Linkage (RL), Data Matching (DM), or Entity Resolution (ER) [Christen et al. 2020, Batini and Scannapieco 2016]. 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 [Batini and Scannapieco 2016].

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) [Christen et al. 2020].

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 [Vatsalan et al. 2016].

A PPRL solution needs to address two issues (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'

[Vatsalan et al. 2013a], which means they contain errors, typos, variations and values that could be missing. For instance, 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 [Vatsalan et al. 2019, Vatsalan et al. 2018, Vatsalan et al. 2016]. 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 [Nóbrega et al. 2016], 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 [Vatsalan et al. 2013a]. 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 [Christen et al. 2020, Vatsalan et al. 2019, Vatsalan et al. 2018, Vatsalan et al. 2016, Vatsalan et al. 2013a] 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 [Christen et al. 2020]. However, this adversarial model is not realistic for real-world applications [Vatsalan et al. 2013a], mainly because it is hard to

[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.

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 techniques [Vatsalan et al. 2013a]. 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 [Christen et al. 2020];
- **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. Research Objectives and Contributions

Based on the challenges summarized in Section 2, our work intends to address the PPRL process's bottlenecks that represent limitations to its extensive use of PPRL. Given the current demands for improvement to the PPRL process, this **work's main goal covers improving privacy and the linkage quality of PPRL**. The privacy improvements will be concentrated on the anonymization and comparison steps using a novel anonymization technique and auditable data comparison protocols, respectively. The linkage quality improvements are focused on automatic (Machine Learning-based) classifiers to PPRL. Each contribution will be employed to tackle a different bottleneck, detailed in the specific goals section

3.1. Specific goals

Considering the proposed main goal and the fact that the privacy 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 anonymization technique;
2. Propose a novel adversary model that reduces the need of thrust by PPRL parties;
3. Propose a machine learning-based classifier to mitigate the threshold selection during the PPRL Classification step;
4. Propose a novel encoded/anonymized record pair representation that enables the use of novel ML-based classifiers (e.g., deep learning-based classifiers) to improve the linkage of the PPRL process;

4. Research Contributions

In order to illustrate our contributions to the PPRL process, we plotted Figure 1. It depicts the PPRL steps, further detailed in Section 1, highlighting the steps directly impacted by our contributions.

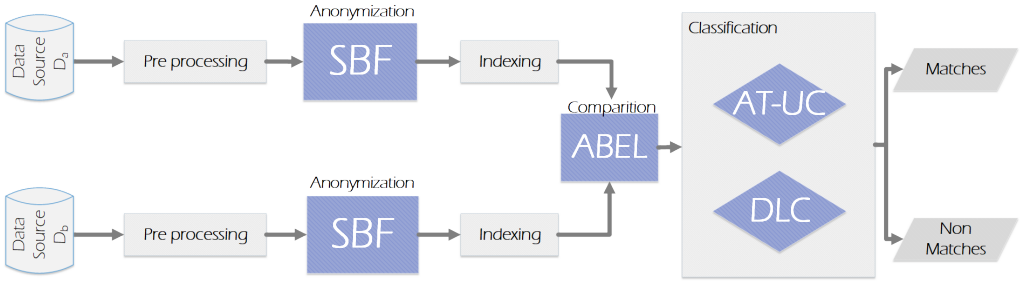

Figure 1. Our contributions within the PPRL process.

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 processes consider the Bloom Filter (BF) anonymization technique, further detailed in our thesis. BF is able to produce an accurate similarity distance between two entities'. However, recent studies [Vidanage et al. 2020, Christen et al. 2017, Ranbaduge and Christen 2018, Vidanage et al. 2019, Christen et al. 2019, Vidanage et al. 2022] 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.

Thus, in this context, we propose the Splitting Bloom Filter (SBF). 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, SBF modifies the anonymization step's output to enable the auditability in the PPRL comparison step, our second contribution. The SBF address our first specific goal.

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.

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. A detailed explanation of the ABEL is presented in our thesis. The ABEL addresses our second specific goal.

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 processes utilize 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 PPRL with better classifiers; eliminating the need for a specialist to guess a threshold and improve the linkage quality. AT-UC utilizes a Transfer Learning technique to employ non-private datasets for training and modifying a classifier to be executed in a private dataset to tackle the absence of labeled data. Moreover, AT-UC also has to define a proper feature space, select a related dataset, and modify the classifier. AT-UC is presented in detail in our thesis. The AT-UC addresses our third specific goal.

In the PPRL context, most of the automatic classifiers employ statistical learning techniques (e.g., Support Vector Machine and Logistic Classifiers) [Christen and Vatsalan 2012, Dong and Rekatsinas 2018, Christen et al. 2020]. Moreover, these classifiers often employ the similarity measures of the records as input (features). Furthermore, standard similarity measures often do not manage well the heterogeneity of the underlying input data. This requires experts to design and configure such measures manually [Loster et al. 2021]. Therefore, due to the limitation of the similarity measures employed in PPRL [Christen et al. 2020, Loster et al. 2021], the classifier task of delineating a suitable separation region (e.g., hyperplane or line) between matching and non-matching records gets more challenging.

Our fourth contribution seeks to mitigate the problem of the similarity measures influence over the classifiers employed in PPRL. Our contribution, the Deep Learning Classifiers (DLC) to identify patterns that indicate whether an anonymized record pair is a match or not. We also propose a novel representation of the encoded record pair based on a dynamical system representation of the data (Recurrence Plot, also detailed in our thesis). It is worth noting that the DLC could improve the linkage quality, mitigating the problems of miss classification presented in PPRL. The DLC address our last specific goal.

In summary, this thesis intends to improve the PPRL process in terms of privacy and linkage quality. Moreover, the contributions introduced in our thesis can be employed to solve problems beyond the PPRL scope. For example, the contribution can be employed to: i) create a Federated Data Linkage solution to integrate multiple sensitive databases (e.g., patient records), providing a tool for epidemiological studies in a country, and ii) adequate data integration tasks to privacy laws (e.g., Brazilian LGPD).

5. Research Relevance

Data privacy or information privacy has recently gained relevance for individuals, governmental institutions, and corporations. The relevance of data privacy is reflected by the number of laws and regulations presented by different countries worldwide [Christen et al. 2020]. Due to this regulation, organizations cannot share their data without addressing the privacy of the individuals [Vatsalan et al. 2013a, Vatsalan et al. 2018]. In this context, the PPRL process aims to improve the input quality to data applications, such as data mining and analytics.

Identifying duplicated entities across private data sources has an important outcome for any data application. For instance, in the decision-making context, the low quality of the data negatively influences the analyses' interpretation based on these data and, consequently, compromises the decisions. For example, a production chain planning process involving the purchase and stock of raw materials, production, and storage of products. It will most likely be hampered if decisions are made based on reports that do not include duplicated materials stored in a different warehouse.

The previous example was made considering non-private data. Now imagine the consequence of duplicated records in health application. For instance, the existence of duplicated patient records may lead to the wrong conclusion in an investigation if a specific medication is efficient against a disease. In summary, PPRL is an important step to analyze, mine, and process private data sources.

PPRL can be employed in different scenarios besides medical and health applications. As an example, we have anti-terrorist, organized crime, and national security applications. For instance, consider an investigation against the money laundry and corruption investigation. Such investigation needs to use and manage various national databases, from many different sources, including law enforcement agencies, financial institutions, travel history, phone records, and so on [Christen 2012, Christen et al. 2020]. It is obvious that this database is highly sensitive and therefore need to be protected [Vatsalan et al. 2013b]. Thus, the PPRL may facilitate linking the information without all data being given to a criminal investigation unit. In other words, only linked information of suspicious individuals is available to the investigation, reducing privacy and confidentiality breaches [Christen et al. 2020].

The need for PPRL solutions is reflected by the wide demand to link real-world private databases. Several countries with different privacy frameworks and legislation are linking some of their sensitive databases. In Brazil, sensitive databases are linking to investigate the consequences of cash welfare and housing programs for Brazil's most poor population concerning their health outcomes [Pita et al. 2015]. In Germany, information about newborns are linking sensitive to measure the quality of their medical systems. Switzerland, Canada, and Australia standardized their anonymization techniques and created a federal institution to coordinate the linkage of private and sensitive information [Christen et al. 2020].

Recently laws and regulations (such as the Brazilian LGPD[2], European Union GDPR[3], and USA HIPAA[4]) intend to enhance individuals' control and rights over their sensitive information, enforcing the institutions (governments or private companies) to protect and preserve the privacy of individuals and entities. Moreover, these laws make the right to privacy a basic human right [Bygrave 1998]. Therefore, privacy violations are beyond data protection laws and can be characterized as a violation of human rights. Consequently, the penalty for a violation of privacy is severe. In this context, the PPRL task is essential to perform data integration in light of existing privacy laws.

A field that can leverage the PPRL techniques is the Federated Data Linkage application. Federated Data Linkage is a data integration task that intends to link data of multiple institutions (e.g., a hospital) that integrates a large network/federation (e.g., all federal

[2]https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

[3]https://gdpr-info.eu/

[4]https://www.hhs.gov/hipaa/index.html

hospitals of a country) [Jarke and Quix 2022]. In the Federated Data Linkage context, laws (e.g., GDPR and LGPD) and regulations (e.g., HIPAA) require that the privacy of the individuals are preserved during the task execution [Boyd et al. 2012]. For example, during the 2019 pandemic, the German university hospitals [Prokosch et al. 2022] made available for researchers a database with information regarding the diagnostic and therapeutic approaches for COVID-19. This database was built over the information available only in isolated silos (comprising all 36 German university hospitals) and inaccessible to external researchers by using a Federated Data Linkage solution (the CODEX[5] project). Therefore, PPRL techniques could link records (e.g., patient records), preserving the privacy of the individual in the Federated Data Linkage applications.

The present work relevance is related to the proposal of methods that attempt to extend the use of PPRL techniques by eliminating bottlenecks. Specifically, this work intends to enhance the linkage quality and propose novel privacy assumptions. The contributions presented in our work can be incorporated into the existing PPRL process (i.e.,medical application) and/or can be employed in future researches.

6. Related Work

This section provides an overview of the contributions presented in this document compared to the state-of-the-art. Given the wide range of our contributions, this section is divided into two parts to enhance clarity: Privacy-preserving protocols and auditable computation in PPRL and Machine Learning applications in PPRL. In the following, we highlight the contributions and advancements in the related work field.

6.1. Privacy-preserving protocols and auditable computation

Most of the attack methods on the BF technique in the context of PPRL are based on the fact that plaintext values that frequently occur in the dataset will generate a BF with the same frequency as the plaintext values. This attack is named as frequent bit pattern attack [Christen et al. 2020, Christen et al. 2017, Kuzu et al. 2012, Kuzu et al. 2011, Vidanage et al. 2022]. In this context, the work of Christen et al. [Christen et al. 2017, Vidanage et al. 2022] shows that to reidentify the entities, an adversary needs to generate the bit pattern using the possible values of the attributes. For instance, to reidentify the encoded name of a patient thread for a specific pathology, an attacker could generate the bit pattern attack using a list containing the most common names in a social network.

Recent works [Ranbaduge and Christen 2018, Vidanage et al. 2019] propose a novel cryptoanalysis attack, a pattern mining-based cryptoanalysis attack. The major advantage of this attack method over the frequent bit pattern attacks is that it neither requires frequent BF nor frequent plaintext values. For instance, the work of Vidanage et al. [Vidanage et al. 2019] applies maximal frequent itemset mining [Cryan 2006] (using a language model) on a BF database to identify sets of frequently co-occurring bit positions that correspond to encoded frequent sub-string values. The Graph-based cryptoanalysis attack uses the q-grams encoded into one single BF or the similarities calculated between BF pairs [Christen et al. 2020] to build a graph and employs the generated graph to reidentify the anonymized data.

To mitigate the aforementioned attacks, **we proposed an extension of the BF (the SBF, detailed in [Nóbrega et al. 2021]) that intends to increase the privacy of the similarity computations of PPRL**.

[5]https://www.netzwerk-universitaetsmedizin.de/projekte/codex

In the PPRL context, Vatsalan and Christen's [Vatsalan 2014, Vatsalan and Christen 2016] presented Security Multiparty Computations (SMC)[6] considering an honest-but-curious adversary model. Vatsalan and Christen's proposes a two-party protocol that eliminates the need for a third party by iteratively revealing selected bits in the BFs between two parties. However, this work is unable to audit the comparison performed during PPRL.

Hybrid PPRL protocols that combine differential privacy techniques with SMC techniques have been proposed to reduce the computational cost of PPRL [He et al. 2017, Inan et al. 2010]. However, such protocols need to disclose all entities stored in their data sources amongst the parties compromising the privacy capabilities of the anonymization. To mitigate this issue, Rao et al. [Rao et al. 2019] propose a framework under a HBC security model that employs a trust-third party to coordinate the record matching between the parties. The parties send to the trust-third party synopses (with Differential Privacy guarantees) of the data. Based on the received synopses, the trusty-third party matches the entities with a distance beyond a threshold specified by the parties. Notice that the works [He et al. 2017, Inan et al. 2010, Rao et al. 2019] consider an HBC and trusty third party, which is a hindrance to the wide usage of PPRL in real-world applications [Christen et al. 2020, Vatsalan et al. 2013b].

In this context, we have identified a research opportunity and proposed the **A**uditable **B**lockchain-Based R**e**cord **L**inkage (**ABEL**), further explained in [Nóbrega 2022, Nóbrega et al. 2021]. Unlike previous works, **ABEL takes into account a new adversary model (the covert adversary) by considering a SMC protocol that is able to audit the computation performed during PPRL**.

6.2. Machine Learning usage in PPRL

Record Linkage (RL) researchers started to explore the usage of techniques originated in machine learning, data mining, artificial intelligence, information retrieval, and database research to improve the classification of linkage process [Christen and Vatsalan 2012, Dong and Rekatsinas 2018]. Many of these approaches are based on supervised learning techniques [Christen 2008, Loster et al. 2021] and assume that training data is available - i.e., record pairs with labels indicating whether they are a match or not. However, such datasets with labels are often unavailable in real-world applications or must be prepared manually (considering a traditional RL context). Furthermore, in the PPRL context, it is not possible to manually label these datasets due to the privacy constraints imposed by the PPRL.

The use of ML in the PPRL context is closely related to Privacy-Preserving Machine Learning (PPML). PPML techniques intend to protect the privacy of the data (training and testing data), model, and prediction [Al-Rubaie and Chang 2019]. Thus, our work and PPML share one goal in common, the protection of data privacy. The works [Chaudhuri and Monteleoni 2009, Brickell and Shmatikov 2009, Tang et al. 2019] use encryption techniques, i.e., homomorphic encryption, to protect the privacy of data. The works of Rajkumar et al. [Rajkumar and Agarwal 2012] and Mivule et al. [Mivule et al. 2012] use the concept of differential privacy to provide privacy-preserving capabilities to their approaches. The work of Miyajima et al. [Miyajima et al. 2017] uses

[6]The SMC is a subfield of cryptography that aims to create methods for parties to jointly compute a function over their inputs while keeping those inputs private [Lindell 2017].

a SMC protocol, to train a classifier in a federated learning context. It is worthwhile to mention that all the presented PPML works [Tang et al. 2019, Miyajima et al. 2017] still need labeled data, which is unavailable within the PPRL context.

Transfer Learning (TL) is another technique that has been explored in recent years in the RL context. The work of Thirumuruganathan et al. [Thirumuruganathan et al. 2018] considers a traditional RL scenario to propose the usage of TL. The authors propose TL usage along with Distributed Representation for Words (called word embeddings). Distributed Representation for Words [Mikolov et al. 2013, Peters et al. 2018], recently introduced to deep learning, are learned from the data such that semantically related words have embeddings that are often close to each other. Typically, these approaches map each word in a dictionary into a high dimensional vector (e.g., 300 dimensions [Thirumuruganathan et al. 2018]) where the geometric relation between the vectors of two words – such as vector difference or cosine similarity – encodes a semantic relationship between them.

Kirielle et al. [Kirielle et al. 2022] propose a TL method for RL over structured data. The work assumes homogeneous domains with the same feature space (same attribute types and similarity functions). In other words, it employs the similarities of the entities' attributes (e.g., names and addresses) from one domain to train a classifier to be employed in another domain that shares the exact attributes. It is worthwhile to mention that the comparison in PPRL is usually performed over the complete record, harming the linkage quality in a PPRL context.

Several linkage processes employ automatic (ML-based) techniques to identify matching entities. However, few of them are compatible with PPRL. In this context, **we proposed the novel automatic classification step** (AT-UC, detailed in [Nóbrega et al. 2023]) and **deep learning-based classifier** (DLC, detailed in [Nóbrega 2022]) **for automatic classification in the PPRL context**.

Notice that the majority of works do not provide privacy-preserving guarantees. PPML, which provides privacy guarantees, requires labeled data. As mentioned early in this section, labeled data is usually unavailable in the PPRL context. The AT-UC and DLC can provide ML-based classifiers without labeled data to the PPRL process. Furthermore, DLC refers to a novel method to compare and classify entity pairs that do not rely on standard similarity measures, addressing the bias introduced by the standard similarity measures [Koudas et al. 2006, Loster et al. 2021].

7. Conclusions and Future Work

In this section, we summarize the contributions presented in this thesis. Moreover, this chapter reveals the perspectives of future research topics by commenting on weaknesses and topics not addressed by the contributions present in this document.

7.1. Contributions

Our work provides the 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 trust to execute PPRL and eliminating the need for an expert to define a classification threshold.

Our contributions are designed to alleviate the limitations imposed by data privacy laws such as the Brazilian *"Lei Geral de Proteção de Dados"* (LGPD) and the European *"General Data Protection Regulation"* (GDPR). These laws restrict the manipulation and operation of private data, and our efforts are aimed at finding ways to work within these limitations while still being able to provide valuable contributions. For example, considering a medical research context, our contributions can be employed to build Privacy-Preserving Federated Linkage System (such as the CODEX[5]) to link patients based on their medical records.

7.2. Results

The products derivate from our contributions are divided into three categories (publications, datasets, and source code), presented in the following.

Publications

Our research has resulted in publications that contribute to the advancement of this field. Among the publications, we have three journal articles and one conference paper. Moreover, we are currently in the final stages of preparing an article for publication in a high-impact journal, which focuses on the application of the DLC (Deep Learning Classifier) method.

Type	Publication
Jorunal	Blockchain-based privacy-preserving record linkage: enhancing data privacy in an untrusted environment - T Nóbrega, CES Pires, DC Nascimento. Information Systems 102, 101826 - 2021. 10.1016/j.is.2021.101826
Jorunal	Limitation of Blockchain-based Privacy-Preserving Record Linkage - T Nóbrega, CES Pires, DC Nascimento. Information Systems 108, 101935 - 2022. 10.1016/j.is.2021.101935
Jorunal	Nóbrega, Thiago, et al. "Towards automatic Privacy-Preserving Record Linkage: A Transfer Learning based classification step." Data & Knowledge Engineering 145 (2023): 102180. 10.1016/j.datak.2023.102180
Conference	Towards Auditable and Intelligent Privacy-Preserving Record Linkage - T Nóbrega, CES Pires, DC Nascimento. Anais Estendidos do XXXVI Simpósio Brasileiro de Bancos de Dados, 99-105S - 2020. DOI 10.5753/sbbd_estendido.2021.18170

Table 1. Publication list.

Datasets

We have curated a comprehensive dataset, named the "**Brazilian Politician dataset for Record Linkage**"[7], which is designed for testing our contribution using Brazilian real-world data. This dataset is a resource for researchers in various domains, such as RL, PPRL, and population studies. It contains structured and processed data and gold standards for evaluation.This dataset facilitates the development and evaluation of classifiers and enables other researchers to test and improve their methods.

Source code with the instruction to re-execute the experiments

Our contributions contain public repositories containing the source code and instructions for re-executing the experiments. These repositories provide convenient access to the implementation details and facilitate our work's replication and further exploration. The repositories for each contribution are as follows:

1. Splitting Bloom Filter[8]
2. Auditable Blockchain-Based PPRL[9]

[7]doi:/10.5281/zenodo.7957492

[8]`https://github.com/thiagonobrega/auditable_pprl`

[9]`https://github.com/thiagonobrega/bcpprl-simplified`

3. Auto-Tuned Unsupervised Classification [10]
4. Deep Learning Classifiers [11]

These repositories serve as valuable resources for researchers interested in exploring, adapting, or extending our work in privacy-preserving record linkage. They provide transparency, reproducibility, and a foundation for further advancements in the field.

7.3. Future Work

Future work based on our contributions includes exploring the integration of privacy-preserving techniques with blockchain technology, such as the Privacy-Preserving Blockchain approach. Researchers have proposed encrypting data on the blockchain and utilizing secure enclaves for executing blockchain nodes to enhance transactional privacy [Dwivedi et al. 2019, Russinovich et al. 2021, Weng et al. 2019].

Another promising direction for future research is the integration of differential privacy into PPRL. Differential privacy can also be integrated into PPRL for stronger privacy guarantees [Dwork 2008, Dwork and Roth 2014]. Techniques such as Bloom Filters and Adaptive Bloom Filters can be enhanced with differential privacy to provide enhanced privacy guarantees and reduce the privacy risk.

Additionally, privacy-preserving techniques can be applied to non-structured data domains like Natural Language Processing (NLP). Distributed Representation of Words (DR) has demonstrated successful utilization in NLP applications, particularly in deep learning models [James et al. 2021]. Leveraging privacy-preserving techniques to encode DR can enable the development of PPRL methods that link non-structured data while maintaining privacy. This opens up possibilities for privacy-preserving NLP solutions, such as linking medical records or identifying patients with specific conditions, with privacy guarantees intact [Dong and Rekatsinas 2018].

This section provides a brief list of future work based on our contributions, for further detail on future work possibilities see our full text [Nóbrega 2022].

References

Al-Rubaie, M. and Chang, J. M. (2019). Privacy-Preserving Machine Learning: Threats and Solutions. *IEEE Security & Privacy*, 17(2):49–58.

Batini, C. and Scannapieco, M. (2016). *Data and Information Quality*. Data-Centric Systems and Applications. Springer, 1 edition.

Boyd, J. H., Ferrante, A. M., O'Keefe, C. M., Bass, A. J., Randall, S. M., and Semmens, J. B. (2012). Data linkage infrastructure for cross-jurisdictional health-related research in australia. *BMC health services research*, 12(1):1–8.

Brickell, J. and Shmatikov, V. (2009). Privacy-preserving classifier learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5628 LNCS:128–147.

Bygrave, L. (1998). Data protection pursuant to the right to privacy in human rights treaties. *International Journal of Law and Information Technology*, 6(3):247–284.

[10] https://github.com/thiagonobrega/tl-pprl
[11] https://github.com/thiagonobrega/nn_pprl

Chaudhuri, K. and Monteleoni, C. (2009). Privacy-preserving logistic regression. *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference*, pages 289–296.

Christen, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. *ACM SIGKDD*, pages 151–159.

Christen, P. (2012). *Data Matching*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Christen, P., Ranbaduge, T., and Schnell, R. (2020). *Linking Sensitive Data*. Springer, Cham.

Christen, P., Ranbaduge, T., Vatsalan, D., and Schnell, R. (2019). Precise and Fast Cryptanalysis for Bloom Filter Based Privacy-Preserving Record Linkage. *IEEE Trans. Knowl. Data Eng.*, 31(11):2164–2177.

Christen, P., Schnell, R., Vatsalan, D., and Ranbaduge, T. (2017). *Efficient Cryptanalysis of Bloom Filters for Privacy-Preserving Record Linkage Peter*, volume 10235 of *Lecture Notes in Computer Science*. Springer, Cham.

Christen, P. and Vatsalan, D. (2012). A flexible data generator for privacy-preserving data mining and record linkage.

Cryan, M. (2006). *Probability and Computing Randomized Algorithms and Probabilistic Analysis*. JSTOR.

Dong, X. L. and Rekatsinas, T. (2018). Data Integration and Machine Learning. In *ICDM*, pages 1645–1650, New York, NY, USA. ACM.

Dwivedi, A. D., Srivastava, G., Dhar, S., and Singh, R. (2019). A decentralized privacy-preserving healthcare blockchain for iot. *Sensors*, 19(2):326.

Dwork, C. (2008). Theory and Applications of Models of Computation. 4978:1–19.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

He, X., Machanavajjhala, A., Flynn, C., and Srivastava, D. (2017). Composing Differential Privacy and Secure Computation. In *ACM SIGSAC*, number 1, pages 1389–1406, New York, New York, USA. ACM Press.

Inan, A., Kantarcioglu, M., Ghinita, G., and Bertino, E. (2010). Private record matching using differential privacy. In *EDBT*, page 123, New York, New York, USA. ACM Press.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer Texts in Statistics. Springer New York, New York, NY.

Jarke, M. and Quix, C. (2022). *Federated Data Integration in Data Spaces*, pages 181–194. Springer, Cham.

Kirielle, N., Christen, P., and Ranbaduge, T. (2022). Transer: Homogeneous transfer learning for entity resolution. In *EDBT*, pages 2:118–2:130. OpenProceedings.org.

Koudas, N., Sarawagi, S., and Srivastava, D. (2006). Record linkage: similarity measures and algorithms. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 802–803.

Kuzu, M., Kantarcioglu, M., Durham, E., and Malin, B. (2011). A Constraint Satisfaction Cryptanalysis of Bloom Filters in Private Record Linkage. *Privacy Enhancing Technologies*, 6794:226–245.

Kuzu, M., Kantarcioglu, M., Durham, E. a., Toth, C., and Malin, B. (2012). A practical approach to achieve private medical record linkage in light of public resources. *Journal of the American Medical Informatics Association*, pages 285–292.

Lindell, Y. (2017). *Tutorials on the Foundations of Cryptography*. Springer.

Loster, M., Koumarelas, I., and Naumann, F. (2021). Knowledge transfer for entity resolution with siamese neural networks. *Journal of Data and Information Quality (JDIQ)*, 13(1):1–25.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality.

Mivule, K., Turner, C., and Ji, S. Y. (2012). Towards a differential privacy and utility preserving machine learning classifier. *Procedia Computer Science*, 12:176–181.

Miyajima, H., Shigei, N., Makino, S., Miyajima, H., Miyanishi, Y., Kitagami, S., and Shiratori, N. (2017). A proposal of privacy preserving reinforcement learning for secure multiparty computation. *Artificial Intelligence Research*, 6(2):57.

Nóbrega, T., Pires, C. E. S., Nascimento, D. C., and Marinho, L. B. (2023). Towards automatic privacy-preserving record linkage: A transfer learning based classification step. *Data & Knowledge Engineering*, 145:102180.

Nóbrega, T. P. d., Pires, C. E. S., and Araujo, T. B. (2016). Avaliação Empírica de Técnicas de Comparação Privada Aplicadas na Resolução de Entidades. In *Proceedings of the 31 st of the Brazilian Symposium on Databases (SBBD16)*, pages 121–126.

Nóbrega, T. (2022). *Towards Auditable and Intelligent Privacy-Preserving Record Linkage*. PhD thesis, PPGCC/UFCG.

Nóbrega, T., Pires, C. E. S., and Nascimento, D. C. (2021). Blockchain-based privacy-preserving record linkage: enhancing data privacy in an untrusted environment. *Information Systems*, 102:101826.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations.

Pita, R., Pinto, C., Melo, P., Silva, M., Barreto, M., and Rasella, D. (2015). A Spark-based workflow for probabilistic record linkage of healthcare data. *CEUR Workshop Proceedings*, 1330:17–26.

Prokosch, H.-U., Bahls, T., Bialke, M., Eils, J., Fegeler, C., Gruendner, J., Haarbrandt, B., Hampf, C., Hoffmann, W., Hund, H., et al. (2022). The covid-19 data exchange platform of the german university medicine. In *Challenges of Trustable AI and Added-Value on Health*, pages 674–678. IOS Press.

Rajkumar, A. and Agarwal, S. (2012). A differentially private stochastic gradient descent algorithm for multiparty classification. *Journal of Machine Learning Research*, 22:933–941.

Ranbaduge, T. and Christen, P. (2018). Privacy-Preserving Temporal Record Linkage. *2018 IEEE International Conference on Data Mining (ICDM)*, pages 377–386.

Rao, F. Y., Cao, J., Bertino, E., and Kantarcioglu, M. (2019). Hybrid private record linkage: Separating differentially private synopses from matching records. *ACM Transactions on Privacy and Security*, 22(3).

Russinovich, M., Costa, M., Fournet, C., Chisnall, D., Delignat-Lavaud, A., Clebsch, S., Vaswani, K., and Bhatia, V. (2021). Toward confidential cloud computing. *Communications of the ACM*, 64(6):54–61.

Tang, F., Wu, W., Liu, J., Wang, H., and Xian, M. (2019). Privacy-preserving distributed deep learning via homomorphic re-encryption. *Electronics (Switzerland)*, 8(4).

Thirumuruganathan, S., Parambath, S. A. P., Ouzzani, M., Tang, N., and Joty, S. (2018). Reuse and Adaptation for Entity Resolution through Transfer Learning.

Vatsalan, D. (2014). *Scalable and Approximate Privacy-Preserving Record Linkage*. PhD thesis.

Vatsalan, D., B, D. K., and Gkoulalas-divanis, A. (2019). *An Overview of Big Data Issues in Privacy-Preserving Record Linkage*, volume 2. Springer.

Vatsalan, D. and Christen, P. (2016). Multi-Party Privacy-Preserving Record Linkage using Bloom Filters.

Vatsalan, D., Christen, P., and Verykios, V. S. (2013a). A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969.

Vatsalan, D., Christen, P., and Verykios, V. S. (2013b). Ef fi cient Two-Party Private Blocking based on Sorted Nearest Neighborhood Clustering. pages 1949–1958.

Vatsalan, D., Karapiperis, D., and Verykios, V. S. (2018). Privacy-Preserving Record Linkage. (January).

Vatsalan, D., Sehili, Z., Christen, P., and Rahm, E. (2016). Privacy-Preserving Record Linkage for Big Data : Current Approaches and Research Challenges. In *Big Data Handbook*. Springer.

Vidanage, A., Christen, P., Ranbaduge, T., and Schnell, R. (2020). A Graph Matching Attack on Privacy-Preserving Record Linkage. *Int. Conf. Inf. Knowl. Manag. Proc.*, pages 1485–1494.

Vidanage, A., Ranbaduge, T., Christen, P., and Schnell, R. (2019). Efficient Pattern Mining based Cryptanalysis for Privacy-Preserving Record Linkage. *Proceedings - International Conference on Data Engineering*, pages 1698–1701.

Vidanage, A., Ranbaduge, T., Christen, P., and Schnell, R. (2022). A taxonomy of attacks on privacy-preserving record linkage. *Journal of Privacy and Confidentiality*, 12(1).

Weng, J., Weng, J., Zhang, J., Li, M., Zhang, Y., and Luo, W. (2019). Deepchain: Auditable and privacy-preserving deep learning with blockchain-based incentive. *IEEE Transactions on Dependable and Secure Computing*, 18(5):2438–2455.