A Framework for Consensual and Online Privacy Preserving Record Linkage in Real-Time

Daniel Müller, Stefan Mau, Irena Pletikosa Cvijikj
Department of Management, Technology and Economics
ETH Zurich
Zurich, Switzerland
danielmueller@ethz.ch, smau@ethz.ch, ipletikosa@ethz.ch

***Abstract*—In the face of heterogeneity, privacy laws and the scale of various data sources, Privacy Preserving Record Linkage is an increasingly relevant topic for organizations that intent to collaborate on a data level. In addition, new collaboration scenarios require an exchange that would take place online and in real-time. To address these needs, in this paper we present a framework for consensual and online privacy preserving record linkage in real-time. The envisioned framework builds upon the proposition that an individual should agree a priori to reveal his identity to an external organization. We discuss the security features necessary for avoiding potential attacks and misuse of the transmitted data, and evaluate our approach through a usage scenario and performance measurements to demonstrate its applicability in real-life business scenarios.**

Keywords – privacy preserving record linkage; consensual; online data integration; real-time

I. INTRODUCTION

Record linkage (RL), also known as entity resolution, or duplicate detection, is a deterministic or heuristic process in which tuples that represent the same real-world entity are being identified [8]. Despite the maturity of this method, due to the high degree of heterogeneity in data structures, the absence of unique identifiers used for the matching process, and the growing privacy concerns, the cross-organizational collaboration on a data level still represents a challenging problem [1]. While there has been extensive research on data integration architectures, RL methods and schema matching techniques (see [1] and the references therein for a detailed overview), privacy considerations have received much less attention.

Previous studies, such as those presented in [3], [4], [5], and [6], showed different possibilities to match entities from separate databases, without compromising the confidentiality of individuals' data. These methods, commonly referred to as privacy preserving record linkage (PPRL), were shown to work efficiently with large datasets in an offline context and were predominantly used in the field of statistical research and medical applications [7]. Moreover, none of the existing approaches has been adapted for commercial applications in online context or has asked the user if he approves the exchange of his personal data. Thus, in the face of heterogeneity, data privacy laws and the scale of various data sources, traditional RL techniques do not provide a solution for emerging problems [7], [8].

In the modern business environment, the data owned by companies is becoming one of their most valuable assets [27]. Better data quality was shown to hold potential for reduction of operational inefficiencies [33]. Moreover, to obtain regulatory compliance and improve their customer relationsip strategies, many industries require accurate and complete customer information. However, aquiring such data is a task that is sometimes not feasible due to the limitations imposed by the design of existing systems and respectively the level of details of the stored customer data, which is usually focused only on attributes relevat for the specific business context [1]. Some of these challenges might be addressed by merging data from external sources, thus gaining new insights and thereby enhancing organization's record base. The obtained external data could further be beneficial for imporvement of marketing and sales campaigns. This approach would be of great value for many industries, and especially for financial services, such as insurance [34], where current customers were found to be more valuable than new customers [11].

While general understanding exists that external data sources could bring benefits to the companies, obtaining external information is not always an easy task. Organizations interested in collaboration on a data level are usually legally restrained to share personally identifiable information (PPI) without the consent of their customers. Societal concerns, reputational risks and regulatory constrains further restrict the possibility for data exchange and thus also for the use of RL approaches for matching of individual data records [31]. The existence of legal wording revealed in organization's terms and conditions is not sufficient to avoid reputational damage and draw criticism from consumer activists [36]. Thus overcoming these challenges is of great concern for many companies.

To contribute in the direction of cross-organizational data exchange in an online context, while taking in consideration the privacy regulations, in this paper we present a framework for consensual PPRL in real-time. The proposed framework builds upon the requirement that an individual should agree a priori to reveal his identity to a collaborating company after the successful outcome of the RL process. As such, this framework is generic and can be applied to any business case

where organizations plan to collaborate on a data level, while asking the individual for a permission in real-time.

The continuation of this paper is structured as follows. In section II we provide an overview of the previous work from the relevant research streams. Section III describes the proposed framework for consensual PPRL in real-time. Section IV describes the data and the methodology used to evaluate the proposed framework, while Section V presents the preliminary results. We discuss the proposed framework from the perspective of overcoming the known challenges as well as our findings from the preliminary evaluation in Section VI. Finally, we conclude our work with a summary and an overview of future research possibilities.

II. Related Work

A. Record Linkage and Data Integration

Record linkage describes a process of identification of near identical records from one or more datasets, which belong to the same real-world entity [8]. This method allows finding of information, which would otherwise not be possible to obtain [9]. The foundation for RL was set in 1969 [13], and has been extended since then. RL approaches originated in epidemiology and census, but gained further importance when organizations merged or collaborated, hence had to join their databases [1]. Applications in health care and census generally have database schemata (column names) which are identical. However, merging organizations often do not follow any standards when naming entity attributes. The same applies for organizations that like to collaborate on a data level in order to increase their revenues, reduce operational inefficiencies or improve regulatory compliance procedures [1].

Existing approaches to identify near identical records can be classified into *unsupervised* or *supervised* methods [11]. Recently, techniques from Data Mining and Artificial Intelligence were applied to improve the performance of unsupervised RL processes. One popular unsupervised approach is the edit-distance measure, which calculates the number of operations that should be performed in order to create identical records (e.g., [12], [13], [14]). In turn, supervised techniques often require the availability of a suitable training data set, which represents a challenge for real world settings. These methodologies can further be classified into learning based approaches (e.g., [10], [11], [12], [13]), or database and graph-based methods (e.g., [14], [15], [16], [17]).

It should be noted that majority of the previously proposed RL methods were evaluated by using small datasets [15], which may not allow a generalization of the results in terms of their applicability to industrial business settings with millions of customer records. Moreover, existing data integration techniques are mostly optimized for offline datasets that should be compared to each other. This approach assumes that both data sets can be compared in full text and exchanged freely between data owner and the organization that performs the matching. Thus, an adaption of existing methods is needed to address the privacy constraints and the demand for online data exchange that are typical for new collaboration scenarios.

B. Privacy Preserving Record Linkage

Due to legal and ethical considerations, performing RL across organizations while maintaining the confidentiality of the transmitted data has gained importance in use cases where records containing personal information are linked [25]. An approach on how to link micro data without access to unencrypted identifiers was first published by [3]. However, it was considered inefficient, hence never gained popularity [1]. A further approach for RL where the exact values must remain unknown to the linking party was introduced by [23]. Similarly, members of the German Record Linkage Center applied privacy-preserving methodologies and developed a protocol to accomplish approximate string comparison on encrypted values [24]. In the study by [30], the authors have sampled 20'000 records from North Carolina voters, which were encrypted and evaluated using a Bloom filter. They were able to launch a frequency attack, which revealed the vulnerability of Bloom filters when applied to datasets which have non-homogeneous frequencies of hashes, known to represent bi-grams. However, they concluded that their results were low in precision and had high computational costs. Therefore, optimization of the privacy preserving RL methods is still a pending task for scholars and practitioners.

C. Online Data Integration

Recently, cross-organizational RL approach in an online setting started attracting the attention of scholars. The first publication in the field discussed query-time of entity resolution applications [10]. The authors of this paper constructed an approach to resolve one query record, instead of reconciling complete databases. They achieved a response time of 31 seconds per query. Further, the need for RL in a networked and online world was recognized by [25]. They provided a matching procedure in (near) real-time by applying inverted indexing techniques, commonly used in Web search engines. In another study, a duplicate detection and fusion technique that works in an online setting was introduced [26]. The goal was to provide a solution, which returns feedback if a newly arriving record (obtained as a response to a query), already exists in an organization's database. The proposed approach addressed the performance issue, but required exact matches and was applied to a single database.

Improvements in terms of efficiency over the traditional RL in the online world were introduced for a *pay-as-you-go* application [10]. This approach was based on a family of techniques for constructing hints, which efficiently maximize the number of matches, given a limited amount of time. In their work, the authors applied the concept to match products in e-commerce stores in real-time. However, they did not take in consideration the privacy constraints.

To contribute to the field of online data integration in real-time, we propose a framework for online RL, while maintaining confidentiality of the exchanged data. We emphasize the importance of the opt-in process to obtain a consensual and real-time data exchange. We discuss the security features necessary for avoiding potential attacks and misbehavior through the involvement of a trusted third party (TTP). We further experimentally evaluate our approach and demonstrate the usability of the proposed approach in real-life

business situations. To the best of our knowledge, our work represents a first attempt for consensual data integration approach over the web in a privacy preserving manner. Thus, it provides an overarching framework which integrates several distinct approaches, each of them addressing only one of the components required in the modern context of online and privacy preserving organizational collaboration on a data level.

III. A Framework for Consensual and Online Privacy Preserving Record Linkage in Real-Time

In a typical data exchange scenario between two collaborating organizations, we assume there is a seller (S) and a buyer (B) of the information. The buyer is interested in acquiring some of the seller's data to improve his own customer records, cross- or upsell his customers, or fulfill regulatory compliance by gaining knowledge about his customers' life circumstances. The customer of the seller needs to be asked individually, if his identity may be revealed to the buyer. In addition, not all customers are interesting for the buyer and they need to be evaluated based on a certain criteria, e.g. the existence of a previous business relationship with the buyer. Finally, we assume that individuals, who know more about how their data records are used for marketing purposes, are more willing to share their personal data in exchange for an incentive, hence they are willing to reveal their identity to an inquiring organization [37].

Building upon these points, the envisioned framework for consensual PPRL in real-time should accommodate the following requirements:

- To address the privacy concerns, customers should agree in real-time to share their information with another party.
- Further, none of the collaborating organizations can have access to the dataset of the other organization, revealing the identity of some or all of the other organizations customers, unless explicitly given permission from the customer.
- To avoid potential data misuse, no external organization involved in the linkage process should be able to resolve the data provided to the real-world entity.
- To allow real-time application, there should be no human intervention.
- To avoid unnecessary data exchange and costs related to it, as well as the potential negative effect for the customers (e.g. ill targeted marketing campaign or inappropriate cross/up-selling), only customers that have a case-specific predetermined features should be asked to opt-in in the sharing process (i.e. common customers of two organizations).

Assuming that two parties have agreed on common identifiers, they include a TTP as a mediator, which performs the RL for them. This way, the two parties do not have to reveal their data sets directly to each other. Instead, they can define a data-specific threshold for similarity, which will be used for the matching classification. If a match is detected by TTP, the customer will be asked to opt-in.

The exact protocol between the three parties involved in the process, i.e. buyer, seller and TTP, contains an initiation phase with two general steps (1-2) that are conducted initially (or on a regular interval), and an iteration phase with four repetitive steps (3-6) that occur each time a data exchange is taking place:

Initiation Phase:

1. Organizations agree which attributes will be used in the linkage process (e.g. surname, date of birth, address, etc.).
2. Organizations agree on a pre-processing method and a hashing algorithm, used to encrypt transmitted data to the TTP.

Iteration Phase:

3. Organizations upload the encoded data to the TTP.
4. Matching of the records takes place at the TTP.
5. In case a match is found, opt-in process for the customer is triggered at the seller.
6. If the customer approves, an exchange process of the records identified as matches occurs (potentially followed by a compensation for the customer).

Figure 1 illustrates the previously described data exchange protocol.

In order to prevent data misuse by any of the involved parties, certain security features should be integrated in the envisioned PPRL framework. First, the seller should not receive any information about the customers of the buyer, and the buyer should not receive customer's information from the seller unless the customer at the seller side has approved the exchange.

Further, to prevent misuse of the data at the TTP, the data has to be protected in a way that will allow the buyer and the seller to exchange the information, but would prevent the TTP from resolving the received data. Instead, TTP should only able to perform RL and resolve if both records represent the same entity.

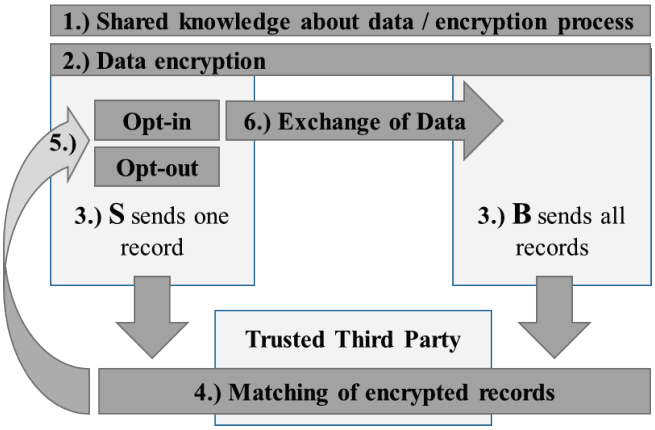

Figure 1. Framework illustrating the six steps for consensual PPRL.

Finally, no human intervention should be involved, allowing a real-time application of the framework, which is particularly interesting for e-commerce companies searching for collaboration with more traditional businesses.

In the continuation, we provide the details of the main concepts integrated in the envisioned framework.

A. Encoding and Matching Methodology

There are several options to encode the data in order to prevent access to it by unauthorized third parties, including the TTD. One possibility would be to encode the data by using a one-way hash function. In addition, adding a shared secret (i.e. a password) to the hash would ensure that only the seller and the buyer would be able to validate the received input. A commonly used approach is the usage of encrypted hashes based on a Cipher, such as the Advanced Encryption Standard (AES), defined by the U.S. National Institute of Standards and Technology (NIST) in 2001 [36]. Alternatively, the secret can be included in the hash itself, accomplished through an encryption algorithm, such as the Keyed-Hash Message Authentication Code (HMAC), or the widely used Secure Hash Algorithms (SHA), ideally SHA-3. These secure hash algorithms are used to encrypt passwords and serve our purpose well. To illustrate the results of the process and the difficulty it would introduce for reconstruction of an initial string, we provide the following example of a hashing outcome:

'Anna' -> 97a9 d330 … 894a 5438, and
'Ann' -> 822e 335f … be66 2110.

The provided example shows that conversion of slightly different strings into a hash-code creates a completely different outcome, even though only one character differs. Thus, to maintain the comparability of two slightly deviating strings, which is the basis for RL, we cannot encode full strings, but need to convert each data record into an array of *n*-grams. For example, when using bi-grams (2-grams), the name 'Dieter' would be converted to the following array:

'Dieter' -> ['_D', 'Di', 'ie', 'et', 'te', 'er', 'r_'].

Both parties, the buyer and the seller, should pre-process all records, and store them as *n*-grams and then apply a password-dependent hashing algorithm. This allows a structured comparison using a Bloom filter [7]. Moreover, to improve the security, both parties should further concatenate a random string (known as "salt") to the pre-hashing value [31].

Once both parties have sent the hashed data, the TTP performs the linking and reports the matching results. To determine if a record pair is a match or not, the Dice coefficient approach can be used [1]. The Dice coefficient is a similarity measure of two *n*-gram quantities and is composed of the share of common *n*-grams and the total *n*-grams:

$$D_{ab} = 2h / (|a| + |b|') \quad (1)$$

where h is the number of shared *n*-grams, and a and b are the number of *n*-grams in strings a, b. High values of the Dice coefficient indicate a good match.

The threshold for classifying the records as a match should be agreed between the buyer and the seller based on the specific data exchange scenario. One option to determine the optimal threshold value would be to consider the offered incentive and compare it to the cost of a false positive. If the record is classified as a match, the TTP sends back a request to the seller to trigger an opt-in action for the customer at the responsive website of the seller.

B. Security Features

1) Dictionary Attack of the Seller and the Buyer

The main reason for inclusion of a TTP is based on the argument that if the seller and the buyer would collaborate without the TTP, they might technically turn evil and store all of the requests. Due to the knowledge of the hashing algorithm, they could resolve the *n*-grams used in a form of a dictionary attack and learn about each other's datasets without paying any incentive to the end-user or without asking for permission. Therefore, the TTP plays a crucial role to avoid such misbehavior.

2) Dictionary Attack of the Trusted Third Party

To avoid a dictionary attack by the TTP, knowledge about the secret hashing algorithm remains with the seller and the buyer. Consequently, the TTP cannot find out which *n*-gram represents which character combination as it does not have the necessary secret, which determines how the hashing works. In addition, a security parameter strength (similar to key length) can be used, which would make the private inputs of both decrypting parties a computationally intractable problem, thus preventing an external party from resolving the hash.

Additional security features to avoid a dictionary attack by the TTP (or any other party) could be achieved through the salting approach [31], i.e. adding a secret component to each *n*-gram before it is hashed. Moreover, the seller and the buyer could further change the hashing algorithm frequently or introduce randomness in the hashing at an agreed scheme. Alternatively, a second hash function over the initial hash would increase the complexity for the TTP regarding a dictionary attack.

3) Frequency Attack of the Trusted Third Party

By transcribing all queries sent from the seller to TTP, or by evaluating the data provided by the buyer, the TTP could eventually learn about the frequencies of the hashed values. In order to avoid a frequency attack of this kind by the TTP, the series of queries from the seller and the records from the buyer should include dummy records specifically designed to make frequencies at which *n*-grams appear to look unilateral. In addition, the seller and the buyer should design some of the dummy records to appear as matches, thus prohibiting the TTP to evaluate the matching records in a frequency attack. For example, the frequency of all *n*-grams could be increased to reach a certain frequency threshold, which could be defined as the frequency of the highest naturally appearing *n*-gram.

The possibility for a frequency attack is further reduced by the fact that the TTP does not know the length of the used *n*-grams, as well as which customer attributes are used for the

matching. Hence, TTP would not know against what to compare the frequency of received *n*-grams.

C. Usage Scenario

To illustrate a scenario of collaboration on a data level that would benefit from the proposed framework, we provide an example in which an e-commerce company, i.e. an online platform allowing individuals to trade with cars, would represent the seller. This company has established a collaboration with an insurance company, which has the role of a buyer. The insurance company would like to know if an existing customer is considering selling an insured car or buying a new one over the e-commerce platform. The insurance company could use this knowledge for marketing and retention campaigns, i.e. to target such customers timely and prevent them from switching to another insurance company. Customers placing classified ads to sell their cars need to pay a fee for the creation of the listing. If an external party would offer a monetary compensation for the private information of such customers, then the e-commerce company, i.e. the seller, could provide these customers a discount.

Both parties have APIs, which allow them to communicate with each other via a TTP. In the initiation phase, both parties have exchanged the knowledge about the hashing algorithm and the structure of their respective database schemes (Step 1). In addition, both parties have applied encryption to their data records (Step 2).

As the customer has finished entering his details into the e-commerce platform, this platform sends a request, which includes the encrypted identifiers to the TTP. The TTP has all of the insurer's data stored (Step 3), but limited to the encrypted identifiers. The insurer may further exclude some records, e.g. low value customers.

When the TTP receives the information from the e-commerce platform (Step 4), it compares it against the predetermined encrypted identifiers provided by the insurer. Once a match has been identified, the e-commerce platform displays the opt-in window to the customer (Step 5), which asks the customer if he is willing to share his personal data with the insurance company. It would be beneficial to postpone this step to a point where the customer has reached the check-out page, in order to avoid the risk of compromising the purchase by imposing an additional decision to be made to the customer [39], which would eventually affect the conversion performance.

If the customer agrees to share his data, an incentive could be provided, e.g. a discount, voucher, or personal gift. Simultaneously, the insurance company would receive the relevant data record (Step 6).

IV. Evaluation

A. Dataset

To evaluate some of the above presented concepts, we used a dataset extracted directly from the data warehouse of a large Swiss insurer, i.e. the buyer. The company sells a broad range of life and non-life insurance products and is among the top three non-life insurers in the Swiss market. For the purpose of this study, only private customers were chosen. For each customer, the following characteristics were available: *First Name*, *Last Name*, *Street Name*, Street *Number* and *Zip Code*. The resulting dataset consisted of 3'689'796 entries from a period between 2010 and 2014.

In addition, a second dataset was obtained from a leading online car-selling platform, i.e. the seller. For consistency reasons, again only private customers were chosen. For each customer, the same characteristics were extracted from the data warehouse. The resulting dataset consisted of 20'104 entries, all from 2014.

To address the computational challenges, from the original sample, only a fraction consisting of 1'000 common customers was drawn offline. A consideration set of matching customers for clerical review were identified by using the Jarrow-Winkler distance approach [2], over the common data fields, i.e. the *First Name*, *Last Name*, *Street Name* and *Number*, and *Zip Code*. Moreover, only non-perfect matches on all fields were considered, based on the following rule:

```
UTL_MATCH.JARO_WINKLER
0.85 < (FNAME, FNAME) < 1
0.85 < (STREET, STREET) < 1
```

The goal of this approach was to create a ground truth for testing of the matching performance in the case when the data is noisy and incomplete, and where deterministic RL applications would fail.

To ensure the accuracy of the data, we carefully reviewed the records manually and removed 49 records, which did not belong to the same entity. The remaining 951 records were given unique IDs, which we used to evaluate the performance of the matching procedure. For this purpose, we were able to apply domain knowledge, as suggested by [38] (i.e. customer history with the insurance), to identify non-duplicates.

Further, to test the precision of our matching procedure, a random set of 1'098 and 2'098 records (without replacement), obtained from the buyer's dataset were added evenly to both datasets. To each of these records, a zero value was assigned as ID to indicate that they do not belong to the same entity, similar to the approach introduced by [21].

B. Methodology

To evaluate the performance of the proposed model, we translated all attributes from records belonging to both data sources into vectors of *n*-grams, and conducted several experiments, as described in the continuation.

First, to determine the role of the *n*-gram's length over the performance in terms of speed (measured in miliseconds), we repeated the matching process by using 2-grams, 3-grams and 4-grams. For simplicity reasons, we took in consideration only the *First Name*, *Street Name* and *Street Number*. Moreover, to test the effect of the number of parallel RL processes over the performance, we repeated the experiment by keeping the size of the seller's dataset to 2'000 records and changing the size of the buyer's dataset between 1, 2, 10 and 100 records.

Further, to determine the role of the *n*-gram's length over the RL performance, we repeated the process by using 2-

grams, 3-grams and 4-grams, and over the same attributes, i.e. *First Name*, *Street Name* and *Street Number*.

The experiments were based on datasets containing 951 records which belong to the same entity and evenly split dummy records of *n1* = 1'098 and *n2* = 2'098. In this case, the performance was estimated through the number of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) based on the previously assigned IDs, i.e. only the identical non-zero IDs were accurate matches (TP), while every other combination would result in a misclassification.

The performance measurement was based on the calculation of the accuracy, precision and recall measures, as shown in the following formulas [20]:

$$\text{accuracy} = (TP + TN) / (TP + FP + FN). \quad (2)$$

$$\text{precision} = TP / (TP + FP). \quad (3)$$

$$\text{recall} = TP / (TP + FN). \quad (4)$$

Finally, we used the F-score as a single measure of performance of the test for the positive class:

$$F = 2*\text{precision}*\text{recall} / (\text{precision} + \text{recall}) \quad (2)$$

It should be noted that depending on the use case, the cost of a false positive might be relatively low and such records could still be considered as leads from the perspective of the buyer. Therefore, the goal would be to achieve a high recall score.

Finally, to determine the *n*-gram frequency threshold needed for preventing the frequency attack, we conducted an analysis of the frequencies of occurrence of individual 2-grams in a dataset consisted of 14'500 customer names. The highest obtained value was chosen as a frequency threshold.

The RL method was implemented in the MergeToolBox 0.742 and Oracle 12c, Release 1. The experiments were performed on a notebook with an Intel Core i7-3520 2.9 Ghz, 8 GB RAM, and running a Windows 8.1 Enterprise operating system.

V. Preliminary Results

In general, resolving 1500 and 2000 record pairs took on average 8 and 20 minutes, respectively.

Moreover, the comparison of the results obtained from the performance testing for different combinations of dataset sizes and *n*-gram lengths, showed that no large differences exist between different combinations. The maximum time needed for the matching, i.e. 329ms, occurred when comparing 1 to 2'000 records, transformed in a form of 2-grams, while the minimal obtained time - 141 ms, which was almost the half of the maximal time, corresponded to the experiment comparing 2 to 2'000 2-grams. Table 1 provides the details of the obtained numbers.

It can be seen that no obvious pattern can be observed in terms of the influence of the number of records and the length of *n*-grams over the matching speed. When comparing one to 2'000 records, as opposed to comparing 100 to 2'000 records, the process took approximately the same time, regardless of the *n*-gram length. Thus, on average, for usage scenarios where less than 100 records are being compared, the throughput occurred in less than one third of a second.

The results of the performance testing in terms of the RL accuracy showed again no significant differences between the dataset when 2-grams, 3-grams and 4-grams were used. Details of the obtained values are provided on Figure 2 and Figure 3. It can be seen from the figures that the variations between the performance measures, i.e. accuracy, precision, recall, and F-score, across datasets based on different *n*-gram lengths do not exceed 3%.

TABLE I. Results of the Entity Resolution Speed Testing

# Records		Performance (milliseconds)			
DB1	***DB2***	***2-grams***	***3-grams***	***4-grams***	***AVG***
1	2'000	329	203	178	237
2	2'000	141	203	288	211
10	2'000	171	188	156	172
100	2'000	312	159	172	214
AVG		238	188	199	

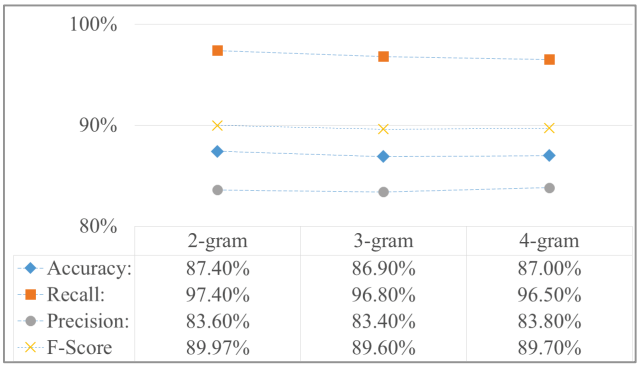

Figure 2. Results of the performance test (accuracy, precission, recall and F-score) with 1098 records

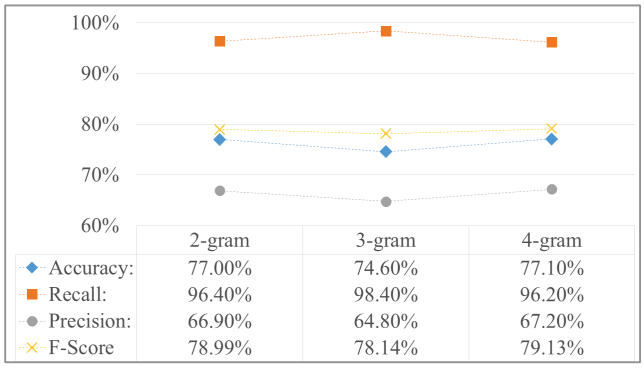

Figure 3. Results of the performance test (accuracy, precission, recall and F-score) with 2098 records

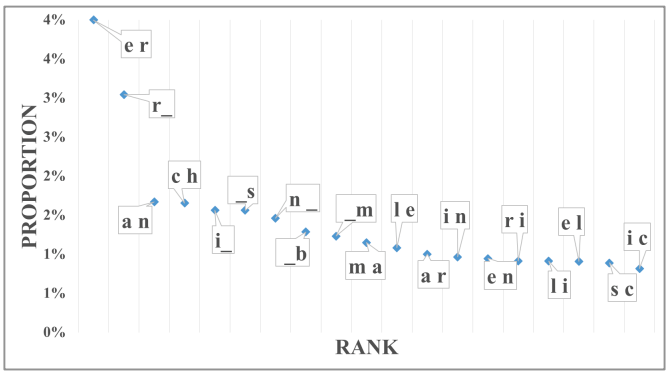

Figure 4. Frequency distribution of 2-grams in Swiss First and Last Names

While the accuracy and precision for the larger dataset (Figure 3) are smaller than in the case of a smaller dataset, the precision remains comparable. Thus, the initial goal of the RL procedure implemented within our framework for achieving high recall was successfully accomplished with values ranging between 96,20% and 98,40% across both datasets.

Finally, the results of the frequency attack simulation, representing the relative distribution of the 19 most frequent bi-grams of Swiss names are illustrated on Figure 4. It can be seen that the highest frequency in our dataset has a value of 4%. Thus in case of using bi-grams, increasing the frequency of the remaining bi-grams to this value would impede the frequency attack.

VI. Discussion and Conclusions

The PPRL framework presented in this paper has a potential to enable exchange and linkage of records between organizations online and in real-time. As such, it represents a solution towards allowing integration of components located in the non-crawlable web. In addition, the data exchange will occur only in those cases when the individual whose personal details are of interest to the collaborating organizations has approved sharing of his data records. This feature of the proposed framework represents an important distinction from the previous work, presented in [3], [9] and [20], where data records gathered within one organization are freely distributed to an external organization upon request, without the knowledge or consent of the affected individual. As such, our approach satisfies the *Privacy by Design* and *Privacy by Default* data processing requirements, which were introduced by government agencies, such as the European Commission, as a necessary component for RL applications [41].

Further, our framework includes advanced security measures, a critical component which was identified as missing by [7] within the original RL approaches presented in [6]. At the data sources belonging to collaborating organizations, personal identifiers are replaced by irreversible hashes and records are then subsequently sent to the registry. At the TTP, records are linked using the results of the secure hash, applied over the array of *n*-grams derived from the original data records. Our preliminary results have validated the assumption that the choice of *n*-gram length does not have an influence on the speed and accuracy performance of the proposed framework, i.e. the use of 2-grams, 3-grams or 4-grams resulted in comparable outcomes. Moreover, the speed of RL based on hashed *n*-grams was shown to be very small. Assuming that in real-life situations, the process will run on computers that are more powerful, these results would be acceptable for a real-time scenario.

Next, neither the TTP, nor individual data source owners are able to reverse-engineer the identifier values of both datasets because they (a) either miss the knowledge about the hash-function, or (b) only have access to their initial data sets. The TTP is also not able to mount a dictionary attack on the hashed *n*-grams because it is not aware of the cryptographic key or password. In addition, due to the insertion of dummy variables, a frequency attack from the TTP is not possible, as the frequency of occurring hashing can be altered to resemble a homogeneous distribution determined by the custom-defined frequency threshold.

Based on the above arguments, we argue that our approach can be used in scenarios that differ from the previously analyzed situations listed in [3], [4], [6], which are tailored to match the needs of RL for census or medical purposes in an offline world, where privacy considerations and data ownership play a subordinated role.

Moreover, as proposed by [47], our framework overcomes the standard approach of 1:1 matching of heterogeneous data sources, by providing an environment where unlimited number of data sources, such as data originating from e-commerce platforms, social networks, etc., can communicate and exchange records with a central registry.

The use of our framework could be generalized to a multi-party, or M:M data exchange. The TTP itself can act as a central hub, allowing the buyers and sellers to bid and offer individual records in real-time, while simultaneously asking individual users for their permission. Potentially, extending the scope of the TTP to represent a decentralized autonomous organization (DAO) would allow creation of a secure, global data exchange, where all participants would benefit by either buying or selling the data to the highest bidding organization. Due to the low marginal cost of the record exchange, we believe that the possibilities based on this framework are manifold.

VII. Summary and Future Work

In this paper, we proposed and evaluated a framework for consensual PPRL in real-time. We showed that by addressing privacy and security concerns, this framework has a potential to overcome the known challenges in the domain of online data exchange between independent organizations with heterogeneous data sources. Our framework allows organizations to collaborate, while leaving the decision about revealing the data to an external party to the affected individual. The results of our preliminary evaluation confirmed the applicability of the proposed approach to real-life scenarios in terms of speed and accuracy of the underlying RL process.

This paper represents an early attempt to provide proof-of-concept for the envisioned, fully functional framework, and as such, it faces few limitations. One limitation is the relatively small number of records and attributes used for the

experiments due to the limited computational resources. In order to overcome this limitation and be able to generalize our findings, we plan to repeat the analysis over the full dataset. In addition, we plan to test the potential effect of the proposed security measures over the performance. Moreover, we plan to evaluate the feasibility of the proposed approach for privacy-preserving blocking rule to reduce the number of RL operations, and thus further improve the performance.

Finally, we plan to implement and deploy the proposed framework in the previously described scenario for collaboration between an insurance company and an e-commerce platform, and evaluate the user acceptance, as well as the benefits for all of the involved parties, thus providing a seminal solution for organizations that plan to collaborate on a data level.

REFERENCES

[1] A. Soni and R. Duggal, "Reducing Risk in KYC (Know Your Customer) for large Indian banks using Big Data Analytics," *International Journal of Computer Applications,* vol. 97, no. 9, pp. 49-53, 2014.

[2] W. E. Winkler, "Overview of record linkage and current research directions," Tech. Rep. 2006-2, Statistical Research Division, US. Census Bureau, Washington.

[3] T. Churches and P. Christen, "Some methods for blindfolded record linkage," *BioMed Central Medical Informatics and Decision Making,* vol. 4, no. 9, 2004.

[4] E. Durham,, Y. Xue, M. Kantarcioglu and B. Malin, "Quantifying the correctness, computational complexity, and security of privacy preserving string comparators for record linkage," *Information Fusion,* vol. 13, no. 4, pp. 245-259, 2011.

[5] R. Hall and S. Fienberg, Privacy in Statistical Databases, Corfu: Springer, 2010.

[6] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *ACM SIGMOD Record,* vol. 3, no. 1, pp. 50-57, 2004.

[7] F. Niedermeyer, S. Steinmetzer, M. Kroll and R. Schnell, "Cryptanalysis of Basic Bloom Filters Used for," *Journal of Privacy and Confidentiality,* vol. 6, no. 2, pp. 59-79, 2014.

[8] Y. Altowim, D. V. Kalashnikov and S. Mehrotra, "Progressive approach to relational entity resolution," *Proceedings of the VLDB Endowment,* vol. 7, no. 11, pp. 999-1010, 2014.

[9] S. Whang, D. Marmaros and H. Garcia-Molina, "Pay-as-you-go entity resolution," *TKDE,* pp. 1111-11124, 2013.

[10] T. Papenbrock, A. Heise and F. Naumann, "Progressive Duplicate Detection," *IEEE Transactions on knowledge and data engineering,* vol. 27, no. 5, 2015.

[11] T. C. Redman, "The Impact of poor Data quality on the typical enterprise," *COMMUNICATIONS OF THE ACM,* vol. 41, no. 2, pp. 79-82, 1998.

[12] J. Y. Xiang, S. Lee and J. K. Kim, "Data quality and firm performance: empirical evidence from the Korean financial industry," *Information Technology and Management,* vol. 14, no. 1, pp. 59-65, 2013.

[13] P. C. Verhoef and B. Donkers, "Predicting customer potential value an application in the insurance industry," *Decision Support Systems,* vol. 32, no. 2, pp. 189-199, 2001.

[14] W. E. Winkler, "Machine Learning, Information Retrieval, and Record Linkage," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2000.

[15] K. Goise and P. Christen, "Towards Automated Record Linkage," in *Fifth Australasian Data Mining Conference*, Canberra, 2006.

[16] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical,* vol. 64, no. 328, 1969.

[17] R. Schnell, "Getting Big Data but avoiding Big Brother," *German Record Linkage Center Working Paper Series,* vol. 6, 2013.

[18] T. M. Mitchell, Machine Learning, Boston: McGrawHill, 1997.

[19] W. W. Cohen, P. Ravikumar and S. E. Fienberg, "A comparison of string distance metrics for name matching tasks," in *IJCAI'03 Workshop on Information Integration on the Web (IIWeb'03)*, Acapulco, 2003.

[20] M. Blienko and R. J. Mooney, "Adaptive d`uplicate detection using learning string similarity measures," in *Proceedings of ACM SIGKDD*, Washington DC, 2003.

[21] S. Chaudhuri, V. Ganti and R. Motawani, "Robust identification of fuzzy duplicates," in *Proceedings of the 21st international conference on data engineering (ICDE '05)*, Tokyo, 2005.

[22] I. Bhattacharya and L. Getoor, "Query-time Entity Resolution," *Journal of Artificial Intelligence Research,* vol. 30, pp. 621-657, 2007.

[23] W. W. Cohen and J. Richman, "Learning to match and cluster large high-dimensional data sets for data integration," in *ACM International Conference on Knowledge Discovery and Data Mining*, Edmonton,, 2002.

[24] M. G. Elfeky, V. S. Verykios and A. K. Elmagarmid, "TAILOR: A record linkage toolbox," in *International Conference on Data Engineering*, San Jose, 2002.

[25] X. Dong, A. Halevy and J. Madhavan, "Reference reconciliation in complex information spaces," in *ACM International Conference on Management of Data (SIGMOD'05)*, Baltimore, 2005.

[26] D. V. Kalashnikov and S. Mehrotra, "Domain independent data cleaning via analysis of entity relationship graphs," *ACM Transactions on Database Systems (TODS),* vol. 31, no. 2, pp. 716-767, 2006.

[27] M. Weis and F. Naumann, "Space and time scalability of duplicate detection in graph data," Technical report, University of Potsdam, Potsdam, 2007.

[28] X. Yin, J. Han and P. S. Yu, "LinkClus: Efficient clustering via heterogeneous semantic links," in *International Conference on Very Large Data Bases*, Seoul, 2006.

[29] P. Christen, "Probabilistic data generation for deduplication and data linkage," in *IDEAL '05*, Brisbane, Springer LNCS 3578, 2005, pp. 109-116.

[30] US General Accounting Office, "Record linkage and privacy: issues in creating new federal research and statistical information, Technical Report GAO-'1-126SP," US General Accounting Office, 2007.

[31] R. Hall and S. E. Fienberg, "Privacy-Preserving Record Linkage," in *Privacy in Statistical Databases*, Springer, 2010, pp. 269-283.

[32] T. Bachteler, J. Reiher and R. Schnell, "Similarity Filtering with Multibit Trees for Record Linkage," *Working Paper WP-GRLC-2013-02, German Record Linkage Center, Nuernberg,* 2013.

[33] M. Kuzu, M. Kantarcioglu, E. Durham and B. Malin, "A constraint satisfaction cryptoanalysis of bloom filters in private record linkage," *Privacy Enhancing Technologies,* vol. 6794, pp. 226-245, 2011.

[34] P. Christen and R. Gayler, "Towards Scalable Real-Time Entity Resolution using a Similarity-Aware Inverted Index Approach," in *Australasian Data Mining Conference (AusDM 2008), ed. John F Roddick, Jinyoung Li, Peter Christen, Paul Kennedy, Association for Computing Machinery Inc (ACM)*, 2008.

[35] E. K. Rezig, E. C. Dragut, M. Ouzzani and A. K. Elmagarmid, "Query-Time Record Linkage and Fusion over Web Databases," in

2015 IEEE 31st International Conference on Data Engineering (ICDE)2015, Seoul, 2015.

[36] J. Turow, M. Hennessy and N. Draper, "The tradeoff fallacy. How marketers are misrepresenting American Consumers And Opening Them Up to Exploitation," Annenberg School for Communication - University of Pennsylvania, 2015.

[37] U.S. National Institute of Standards and Technology (NIST) , "Announcing the Advanced Encryption standard (AES)," Processing Standards Publication 197 , Gaithersburg, 2001.

[38] E. Durham, Y. Xue, M. Kantarcioglu and B. Malin, "Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage," *Information Fusion,* vol. 13, pp. 245-259, 2012.

[39] V. Reding, "Data Protection Day 2014: Full Speed on EU Data," European Comission, Brussels, 2014.

[40] European Commission, "Regulation of the European Parlament and of the Counsil on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," European Commission , Brussels, 2012.

[41] B. Schneier, Applied Cyptogography: Ptrotocols, Algorithms and Source Code in C, Wiley, 2006.

[42] G. Lohse and P. Spiller, "Electronic shopping," *Communications of the ACM,* vol. 41, no. 7, pp. 81-87, 1998.

[43] N. Vesdapunt, K. Bellare and N. Dalvi, "Crowdsourcing algorithms for entity resolution," *Proceedings of the VLDB Endowment,* vol. 8, no. 5, pp. 641-641, 2015.

[44] D. G. Brizan and A. U. Tansel, "A Survey of Entity Resolution and Record Linkage Methodologies," *Communications of the IIMA,* vol. 6, no. 3, 2006.

[45] Accenture, "Data Privacy and Protection at the Tipping Point," 2009.

[46] J. Hipp, U. Guntzer and U. Grimmer, "Data Quality Mining: Making a Virtue of Necessity. In the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery," 2001.

[47] V. M. Shelake and V. S. Bhojane, "A Novel Approach for Multi-Source Heterogeneous Database Integration," in *International Conference on Machine Intelligence and Research Advancement*, 2013.

[48] W. A. Kamakura, M. Wedel, F. De Rosa and J. A. Mazzon, "Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction," *International Journal of Research in Marketing,* vol. 20, no. 1, pp. 445-465, 2003.

[49] E. Ngai, L. Xiu and D. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert Systems with Applications,* vol. 36, p. 2592–2602.

[50] P. S. Fader and B. G. Hardie, "Probability Models for Customer-Base Analysis," *Journal of Interactive Marketing,* vol. 23, pp. 61-69, 2009.

[51] P. Christen and A. Pudjijono, "Accurate synthetic generation of realistic personal information," in *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, Bankok, Thailand, 2009.

[52] M. N. Database. [Online]. Available: http://www.massmind.org/techref/ecommerce/nicknames.htm. [Accessed 19 08 2015].

[53] J. Brustein, Bloomberg, 24 03 2015. [Online]. Available: http://www.bloomberg.com/news/articles/2015-03-24/radioshack-s-bankruptcy-could-give-your-customer-data-to-the-highest-bidder. [Accessed 19 08 2015].

[54] Z. Yan, Q. Li, Y. Dong, L. Cao and P. Pan, "A Deep Web Data Integration Model for Pervasive Computing," *Third International Conference on Pervasive Computing and Applications (ICPCA 2008),* vol. 1, pp. 414 - 417, 2008.