SecUREmatch: Integrating Clerical Review in Privacy-Preserving Record Linkage

Florens Rohde
ScaDS.AI Dresden/Leipzig, Leipzig University
Leipzig, Germany
rohde@informatik.uni-leipzig.de

Victor Christen
ScaDS.AI Dresden/Leipzig, Leipzig University
Leipzig, Germany
christen@informatik.uni-leipzig.de

Erhard Rahm
ScaDS.AI Dresden/Leipzig, Leipzig University
Leipzig, Germany
rahm@informatik.uni-leipzig.de

Abstract

Data integration processes that combine sensitive personal information from different institutions are usually subject to strict privacy regulations. Privacy-preserving record linkage (PPRL) methods can be used in such projects to conceal the identities. However, the quality of such linkages may be low as the parametrization mostly must be done blindly or based on estimations from previous supposedly similar linkage problems. Our framework, SecUREmatch, integrates a multi-layer clerical review system to adapt the matching algorithm to the actual linked data in order to achieve high linkage quality while following the privacy-by-design principle.

CCS Concepts

• **Information systems** → **Entity resolution**; • **Security and privacy** → **Privacy-preserving protocols**.

Keywords

Record Linkage, Privacy, Clerical Review, Active Learning

ACM Reference Format:
Florens Rohde, Victor Christen, and Erhard Rahm. 2025. SecUREmatch: Integrating Clerical Review in Privacy-Preserving Record Linkage. In *Companion of the 2025 International Conference on Management of Data (SIGMOD-Companion '25), June 22–27, 2025, Berlin, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3722212.3725131

1 Introduction

Record linkage enables recognizing different representations of the same real-world entity, such as an individual. This process is essential for many data integration tasks, facilitating the combination of multiple data sources to enhance data analysis. Since unique record identifiers are often unavailable for straightforward join operations, records are typically compared in pairs based on identifying attributes like first name, last name, and date of birth, and are then classified as either a match or a non-match.

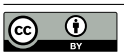
Privacy-preserving record linkage (PPRL) methods facilitate the comparison of records without the need to exchange the sensitive plaintext records between data owners or with third parties [1]. A typical use case for PPRL is the secure combination of microdata from multiple institutions, e.g., hospitals, for secondary research. The process involves separating the microdata from identifying data, performing entity resolution on the latter to link records across the sources, and assigning a globally unique pseudonym to each entity. The microdata is merged into a pseudonymized dataset which ensures privacy while retaining utility for research purposes.

To protect the identities, data owners encode the identifying attributes prior to transmitting them to an independent linkage unit, which conducts the matching process solely on the encoded data. While various perturbation-based encoding techniques have been proposed, the most widely used and quasi-standard approach is based on Bloom filters [7]. However, these methods generally do not allow for clerical review to assess quality. Research on privacy-preserving clerical review (PPCR) systems for record linkage is limited, where attribute values are gradually revealed and displayed using visual masks [9, 10, 14]. Nonetheless, these masks are only for display purposes, and the reviewing institution still receives the complete plaintext data. Furthermore, this approach does not focus on minimizing labeling efforts or enhancing an automatic classification model based on labeled samples.

Contributions. We present SecUREmatch, an open-source privacy-preserving record linkage framework which integrates clerical review with minimal information disclosure. It is based on a recently published linkage protocol that incorporates multi-layer active learning to improve linkage quality while ensuring privacy [15]. This demonstration allows end-to-end execution of such a protocol on datasets without ground truth and thus without pretrained classification models by providing a graphical user interface for control and clerical review. In particular, this demonstration comprises the following aspects:

- Introduction to Bloom filter based privacy-preserving record linkage and multi-layer comparison approaches
- Interactive protocol execution with masked clerical review and active learning for configuration adaption
- Analysis of privacy effects, shared data between the institutions and evolving linkage quality (in evaluation mode)
- Inspection of modular service API documentation

Related frameworks. While a variety of record linkage frameworks exist, none provide the functionality required for our multi-layer clerical review system. On the one hand, there are general record linkage frameworks [3, 12] or specialized PPRL tools like PRIMAT [5] that provide implementations of record linkage and perturbation methods. However, they are either programming libraries or provide only command line interfaces and thus, are not designed as stand-alone services. JedAI [13] has a GUI but lacks support for PPRL and clerical review. On the other hand, frameworks like SOEMPI [18] employ a service-oriented architecture which



SIGMOD-Companion '25, Berlin, Germany

ACM ISBN 979-8-4007-1564-8/2025/06
https://doi.org/10.1145/3722212.3725131

is useful for realizing iterative linkage protocols that require data flow orchestration. None of those tools integrate active learning processes to reduce reviewing effort. dedupe.io used active learning methods only in its discontinued non-free version. Some master patient index tools like Mainzelliste [16] or E-PIX [8] comprise Bloom filter based record linkage but have no native clerical review functionality at all or do not support an incremental improvement of previous labels based on active learning processes. To the best of our knowledge, none of these tools support pair-specific encoding secrets, which is required for secure usage of attribute level encodings in the multi-layer linkage scenario.

2 Method

Bloom filter based PPRL. Bloom filter encodings were initially proposed for PPRL by Schnell et al. [17] and became very popular in practical applications on large datasets [1]. The plaintext is tokenized and each element is mapped to multiple positions in the bit vector using cryptographic hash functions. Similar inputs lead to similar encodings which allows for fuzzy matching to tolerate typical data errors like typos or changes of some attributes, e.g., due to marriage or moving to a new residence. However, due to the similarity preserving transformation, Bloom filter encodings are susceptible to frequency attacks [19]. In particular, attribute-level encodings that transform each attribute of a record separately should be avoided to prevent an alignment of frequent encoded attribute values to frequent plaintext values [2]. Various hardening techniques to hamper reidentification attacks have been proposed [6]. Most importantly, multiple or all attributes are combined in a single encoded representation. Thereby, tokens from different attributes lead to colliding hashes. Another very effective approach is the use of record-specific secrets in the hashing process so that equal values of different records are not encoded identically. Naturally, this is not applicable for primary linkage of datasets as it makes records in general incomparable. It can be used however if only certain subgroups or pairs that shall be compared, are encoded identically. This technique allows for attribute-level comparisons of uncertain pairs.

Multi-layer comparison. For above-mentioned privacy reasons, record-level encoding techniques are strongly recommended. Therefore, record pair comparisons should be mainly conducted based on these encodings. However, the information available for match classification is typically limited to a single similarity value, which can result in a significant number of incorrect or uncertain pairs. To improve the linkage quality, attribute similarities can be highly beneficial, as they provide additional context for resolving such ambiguities. For very similar or dissimilar records, the attribute similarities are not necessary and therefore no attribute-level encodings need to be provided. However, high-performance classification models leveraging these similarities to resolve uncertain cases often require training and may still fail to classify all pairs with high certainty. In such cases, human intervention is necessary to resolve the remaining uncertain pairs. Our linkage protocol is composed of these three layers of comparison.

Clerical review with minimal information disclosure. Fig. 1 illustrates different levels of disclosure for human decision making. For

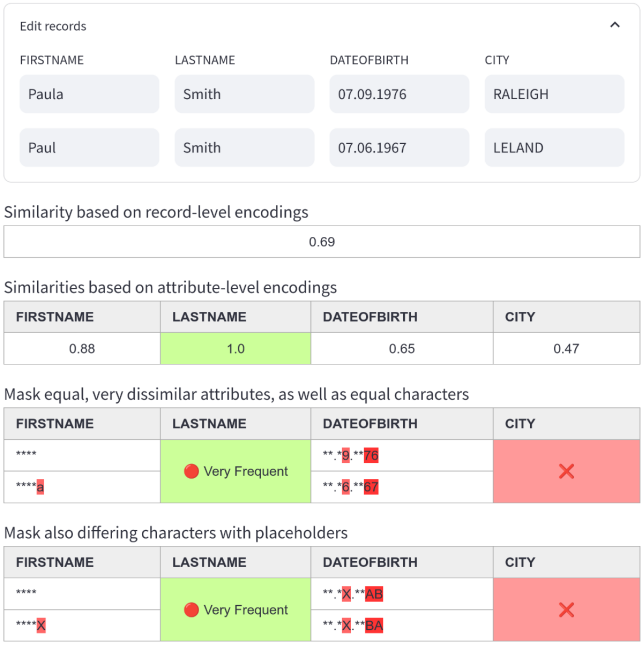


Figure 1: Illustration of different comparison methods for record pairs: While record-level encodings only provide very limited information for classification, attribute-level encodings and masked displays reveal much more features that can be used by classification models or human oracles.

very similar or dissimilar attributes, the actual values are hardly relevant for the match classification and thus, are not displayed. In contrast, the (categorical) frequency of agreeing attributes is relevant and therefore included in the visualization. For somewhat similar attributes, the agreeing characters may also be masked. The full masking even replaces disagreeing parts with placeholders. Previous studies indicate that a dynamic disclosure mechanism which incrementally reveals more details if required has little impact on the decision quality [10].

Integration of active learning. Although partially masked, the manual clerical review process inherently poses the risk of reidentification, e.g., in case of very unique attribute values. Therefore, the number of such reviews should be kept as low as possible. To prevent reviews of similar samples, the protocol comprises an active learning process where the classification models for attribute-level and record-level similarities are updated in multiple iterations based on the labeled samples from lower layers. Thus, pairs without (manual) review benefit from reclassification using refined models.

Privacy measures. The risk of attacks on Bloom filter based encodings is reduced when less frequent patterns are present which means that all bit positions have the same likelihood of being set to '1'. We report the Gini coefficient for measuring the dissimilarity of the bit frequency distribution with a uniform distribution. For masked clerical review, a k-anonymity-based privacy risk score (KAPR) has been proposed [11]. It measures how unique the records are based on the disclosed information.

3 System Overview

Architecture. Conventional Bloom-filter-based PPRL is based on two components, for the data owner and the linkage unit respectively. The data owner part comprises the preprocessing and in particular the encoding. The linkage unit is responsible for the actual matching process including blocking, comparison, classification and possibly clustering in a multi-source scenario. In the multi-layer protocol, another component, the protocol manager, is essential for orchestrating the data flow between those components. We develop all three components as Java Spring Boot Applications, that are controlled via RESTful interfaces. For this demonstration, we build a web-based frontend with streamlit[1] to visualize and control the state and outcome of the linkage protocol and enable masked clerical review. The clients for the backend services are automatically generated using the OpenAPI generator[2].

Project-based linkage. The PPRL services are designed for managing multiple linkage projects in parallel. Thereby, it is possible to execute multiple linkages independently on the same linkage unit instance. For each project, a secret has to be agreed on between the data owners. When requesting encoded records from the data owner service, multiple seeds for the cryptographic hash functions of each attribute are computed using a secret derivation algorithm. In the multi-layer linkage protocol, the layers are modeled as separate projects and therefore, can be managed by different service instances and data custodians. Lower comparison layers are given a reference project id that they use for reporting their predictions.

Protocol execution. The protocol manager service is responsible for triggering the operations of the other services such as providing attribute-level encodings with pair-specific secrets for the attribute-level comparison or selected plaintext attributes for the masked clerical review. Each step of the linkage protocol can be triggered manually so that the state can be inspected in between. For convenience, there is an autocontinue option to run the protocol until an optionally defined breakpoint is reached. The demonstrator displays the previous processing steps as well as a preview of the upcoming step.

Database. The PPRL services persist encoding, matching and protocol configurations as well as the datasets and record pairs independently from each other in MongoDB collections. For testing purposes, they can be provided by the same database management system, while in practical use cases, distinct instances under control of the respective institutions would be used. The multi-layer protocol comprises multiple units which makes it hard to assess who has access to which information. Therefore, the demonstrator contains a viewer so that users can inspect the data each layer holds. It shows aggregate privacy-related metrics such as the number and frequency distribution of the available data as well as a consumable subset of the raw data.

Machine learning and active learning. The PPRL linkage unit natively uses the Weka machine learning library [4] which provides a variety of common classification models. The record-level classification is limited to a single threshold. For attribute-level similarities, the demonstration uses an extension of the Random Forest implementation which replaces older decision trees in subsequent model updates to achieve a more robust evolution. The linkage unit also supports integrating external classification services via a RESTful API, for which we provide a reference implementation in Python to facilitate its rich machine learning ecosystem.

Masked clerical review. The visual masking of somewhat similar attributes is conducted in the streamlit frontend. For convenience, the review process can also be simulated with an adjustable error rate. Thereby, the audience can focus more on the overall linkage process without having to spend too much time on manual labeling.

Linkage quality development. The overall purpose of the multi-layer clerical review is an improvement of the linkage quality with limited labeling. Therefore, after each batch of reviewed labels and after each subsequent model update, the linkage quality is computed based on the ground truth of the experimental datasets. Linkage quality is measured using standard metrics for binary classification tasks with imbalanced classes, namely recall, precision and F1-score. The tool can also be used in a blind execution mode when no ground truth is available or shall not be displayed in order to simulate a real-world linkage scenario. The user interface illustrates the match prediction changes based on attribute-level comparison and masked clerical review which allows the users to assess their usefulness.

4 Demonstration

On the landing page of the web-based demonstration, the audience is introduced to Bloom filter based PPRL and the multi-layer clerical review approach using explanatory illustrations and the opportunity to encode and compare editable records based on the different encodings and clerical review masks, similar to Fig. 1.

Linkage problem selection. The demonstration includes small, medium and large sized subsets of the North Carolina Voter Registry (NCVR). The two sources are built from different snapshots in time and thus, contain real-world changes like new addresses but also data errors like name abbreviations. The small dataset can be linked in reasonable time on mobile computers while the medium and large sized version illustrates the benefits of the active learning process for linkage improvements with limited manual review.

Protocol execution. Initially, the PPRL process is conducted based solely on record-level encodings. The result serves as a baseline that the user tries to exceed by making use of the more detailed comparison layers. The first iteration of the protocol is executed step-by-step to familiarize the audience with the process and data flow. Afterwards, the automatic execution mode is used to fast-forward to the steps where manual intervention is possible. Users can conduct the clerical review with different masks and optionally get feedback on their performance. They can also adapt the record-level threshold manually based on the similarity histogram, either for the full dataset or for the reviewed pairs only.

Analysis. To illustrate the privacy benefits of the multi-layer approach, the privacy measures are compared to a linkage result solely based on attribute-level encodings as well as a linkage process where uncertain pairs from record-level comparison are manually reviewed without the intermediate attribute-level comparison.

[1] https://streamlit.io
[2] https://openapi-generator.tech/docs/generators/python

Figure 2: Linkage execution interface of SecUREmatch with columnar views for comparison layers. From top to bottom, users can inspect the database contents, the protocol configuration, privacy and quality measures and an overview of the previous processing steps. The overlaid box allows comparison of the current linkage quality with reference results.

Technical details. Users who are interested in the protocol execution implementation may inspect and test the backend services using the REST API documentation tool Swagger UI[3].

[3]https://swagger.io/tools/swagger-ui/

Availability. The source code and a demonstration video are available at https://github.com/floroh/pprl-goodall.

Acknowledgments

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany and by the Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus for ScaDS.AI. The masked clerical review implementation is partially based on work by Max Schrodt.

References

[1] Peter Christen, Thilina Ranbaduge, and Rainer Schnell. 2020. *Linking Sensitive Data.* Springer, Cham. doi:10.1007/978-3-030-59706-1
[2] Peter Christen, Thilina Ranbaduge, Dinusha Vatsalan, and Rainer Schnell. 2019. Precise and Fast Cryptanalysis for Bloom Filter Based Privacy-Preserving Record Linkage. *TKDE* 31, 11 (2019), 2164–2177. doi:10.1109/TKDE.2018.2874004
[3] Jonathan De Bruin. 2019. *Python Record Linkage Toolkit: A toolkit for record linkage and duplicate detection in Python.* doi:10.5281/zenodo.3559042
[4] Eibe Frank, Mark A Hall, and Ian H Witten. 2016. *The WEKA workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"* (4th ed.). Morgan Kaufmann, San Francisco, California.
[5] Martin Franke, Ziad Sehili, and Erhard Rahm. 2019. Primat: a toolbox for fast privacy-preserving matching. *Proceedings of the VLDB* 12, 12 (2019), 1826–1829.
[6] Martin Franke, Ziad Sehili, Florens Rohde, and Erhard Rahm. 2021. Evaluation of Hardening Techniques for Privacy-Preserving Record Linkage. In *Proceedings of the 24th EDBT.* 289–300. doi:10.5441/002/edbt.2021.26
[7] Aris Gkoulalas-Divanis, Dinusha Vatsalan, Dimitrios Karapiperis, and Murat Kantarcioglu. 2021. Modern Privacy-Preserving Record Linkage Techniques: An Overview. *IEEE Transactions on Information Forensics and Security 16* 16 (2021), 4966–4987. doi:10.1109/TIFS.2021.3114026
[8] Christopher Hampf, Lars Geidel, Norman Zerbe, Martin Bialke, Dana Stahl, Arne Blumentritt, Thomas Bahls, Peter Hufnagl, and Wolfgang Hoffmann. 2020. Assessment of scalability and performance of the record linkage tool E-PIX® in managing multi-million patients in research projects at a large university hospital in Germany. *Journal of Translational Medicine* 18 (2020), 1–11.
[9] Hye-Chung Kum, Ashok Krishnamurthy, Ashwin Machanavajjhala, Michael K Reiter, and Stanley Ahalt. 2014. Privacy preserving interactive record linkage (PPIRL). *JAMIA* 21, 2 (2014), 212–220. doi:10.1136/amiajnl-2013-002165
[10] Hye-Chung Kum, Eric D Ragan, Gurudev Ilangovan, Mahin Ramezani, Qinbo Li, and Cason Schmit. 2019. Enhancing Privacy through an Interactive On-demand Incremental Information Disclosure Interface: Applying Privacy-by-Design to Record Linkage. *Proceedings of the 15th SOUPS* (2019), 175–189.
[11] Qinbo Li, Adam G. D'Souza, Cason Schmit, and Hye-Chung Kum. 2019. Increasing Transparent and Accountable Use of Data by Quantifying the Actual Privacy Risk in Interactive Record Linkage. arXiv:1906.03345.
[12] Robin Linacre, Sam Lindsay, Theodore Manassis, Zoe Slade, Tom Hepworth, Ross Kennedy, and Andrew Bond. 2022. Splink: Free software for probabilistic record linkage at scale. *IJPDS* 7, 3 (Aug. 2022). doi:10.23889/ijpds.v7i3.1794
[13] George Papadakis, Leonidas Tsekouras, Emmanouil Thanos, George Giannakopoulos, Themis Palpanas, and Manolis Koubarakis. 2018. The return of jedai: End-to-end entity resolution for structured and semi-structured data. *Proceedings of the VLDB* 11, 12 (2018), 1950–1953.
[14] Eric D. Ragan, Gurudev Ilangovan, Hye Chung Kum, and Han Wang. 2018. Balancing privacy and information disclosure in interactive record linkage with visual masking. *Proceedings of the 2018 CHI conference on human factors in computing systems.* doi:10.1145/3173574.3173900
[15] Florens Rohde, Victor Christen, Martin Franke, and Erhard Rahm. 2025. Multi-Layer Privacy-Preserving Record Linkage with Clerical Review based on gradual information disclosure. In *BTW 2025.* Gesellschaft für Informatik, Bonn, 393–415. doi:10.18420/BTW2025-18
[16] Florens Rohde, Martin Franke, Ziad Sehili, Martin Lablans, and Erhard Rahm. 2021. Optimization of the Mainzelliste software for fast privacy-preserving record linkage. *Journal of Translational Medicine* 19, 33 (2021).
[17] Rainer Schnell, T. Bachteler, and J. Reiher. 2009. Privacy-preserving record linkage using Bloom filters. *BMC Med. Inf. & Decision Making* 9 (2009), 41.
[18] Csaba Toth, Elizabeth Durham, Murat Kantarcioglu, Yuan Xue, and Bradley Malin. 2014. SOEMPI: A Secure Open Enterprise Master Patient Index Software Toolkit for Private Record Linkage.. In *AMIA Annual Symposium Proceedings*, Vol. 2014. 1105–14.
[19] Anushka Vidanage, Thilina Ranbaduge, Peter Christen, and Rainer Schnell. 2022. A Taxonomy of Attacks on Privacy-Preserving Record Linkage. *Journal of Privacy and Confidentiality* 12, 1 (2022). doi:10.29012/jpc.764