

A Privacy-Preserving Schema Matching Technique based on k -Anonymity*

Domenico Beneventano¹, Sonia Bergamaschi¹ and Carmelo La Corte¹

¹University of Modena and Reggio Emilia, Modena, Italy

Abstract

Most privacy-preserving schema matching methods are based on the assumption that direct access to plaintext data is not possible. Consequently, only metadata, such as attribute names and their associated descriptions, can be utilized. However, in many cases, even these metadata are not fully available, with attribute descriptions often missing.

Our claim is that privacy-preserving instance-based schema matching can still be achieved by **leveraging anonymized data**. Specifically, we propose to evaluate how k -anonymity influences the performance of instance-based schema matching. To this end, we investigate the effect of k -anonymity on the accuracy and efficiency of instance-based schema matching methods.

In this paper, we present preliminary results obtained by applying classical instance-based schema matching methods together with well-known k -anonymity techniques. To assess the impact of anonymization on schema matching performance, we go beyond a standard evaluation against a known gold standard by also comparing anonymized results to those obtained from cleartext data. This allows us to quantify the effect of k -anonymity on result quality, even in the absence of a reference alignment, and to determine whether results achieved on plaintext data remain reliable under privacy constraints.

Keywords

Privacy-Preserving Schema Matching, k -Anonymity, Instance-Based Schema Matching, Data Integration

1. Introduction

Data integration across organizations offers significant advantages, such as enhancing data quality and uncovering new insights that individual databases alone cannot provide [1]. Key steps in the data integration process include schema matching—identifying which attributes and tables across different datasets represent the same type of information—and record linkage, also known as entity resolution, which focuses on determining which records from one or more datasets correspond to the same real-world entity [2]. When integrating personal or sensitive data across organizations, ensuring privacy and confidentiality is essential to protect against unauthorized access.

As stated in [3], although there is extensive research on Privacy Preserving Record Linkage (PPRL) [4, 5], relatively few studies have addressed privacy concerns within schema matching. The goal in this area is to develop techniques that avoid exposing any sensitive information related to the source schemas or data; this is particularly important given that many PPRL approaches rely on having schemas already aligned.

In this paper, we propose a *Privacy-Preserving Schema Matching* technique that builds upon well-established results and existing frameworks from the literature in both privacy protection and schema matching, integrating them into a unified approach.

Regarding privacy protection, anonymization aims to prevent the re-identification of individuals by ensuring that data records can no longer be uniquely traced back to a specific person. Among various privacy paradigms, k -anonymity is the most widely adopted. It works by transforming quasi-identifiers (QIDs)—such as age, gender, or location—so that each record becomes indistinguishable from at least $k-1$ others in the dataset. However, even this relatively simple privacy model can introduce data distortion, leading to information loss and a potential negative impact on automated analysis—including techniques such as schema matching. The extent of this distortion directly affects data usability, making it essential

ITADATA2025: The 4th Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy

0000-0001-6616-1753 (D. Beneventano); 0000-0001-8087-6587 (S. Bergamaschi)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

to minimize information loss. For this reason, inspired by the findings presented in [6], our focus is primarily on *k-anonymity*, excluding other approaches such as ϵ -differential privacy, which—despite offering strong privacy guarantees—often compromises data utility for practical values of ϵ .

With respect to schema matching, existing techniques can be grouped according to the type of information they leverage. Specifically, *schema-level* approaches rely solely on schema-level metadata—such as attribute names, data types, and contextual information—to identify potential matches. In contrast, *instance-level* methods focus on the actual data values, using strategies such as comparing attribute value distributions or applying syntactic similarity metrics.

Instance-level schema matching often violates privacy constraints because it requires querying actual data on both sides. To address this, *SMAT* [7] introduces a schema-level approach that leverages only local table and attribute names, along with their descriptions. Developed for the privacy-sensitive healthcare domain, *SMAT* is evaluated on a benchmark dataset of real-world schema mappings, and serves as a baseline for aligning datasets to widely adopted standards such as the OMOP Common Data Model [8].

However, this method presents two main limitations. First, it requires the disclosure of local table and attribute names, along with their descriptions—an assumption that may not be acceptable in privacy-preserving scenarios. Second, in many real-world cases, such metadata are incomplete or unavailable, with attribute descriptions often missing. For these reasons, the *privacy-preserving schema matching* technique we propose is purely instance-based, meaning it does not rely on table or attribute names or descriptions, which can therefore be treated as meaningless symbols. Instead, it operates directly on *anonymized data*.

The main contribution of this paper is a *Privacy-Preserving Schema Matching* technique that leverages well-established results and builds on existing frameworks in the literature across both privacy protection and schema matching, combining them into a unified approach. More precisely, we present preliminary results obtained by applying classical instance-based schema matching methods in conjunction with well-known *k-anonymity* techniques. To assess the impact of anonymization on schema matching performance, we go beyond a standard evaluation against a known gold standard by also comparing the results obtained from *k-anonymized* data to those derived from cleartext datasets. This comparison allows us to quantify the effect of *k-anonymity* on result quality, even in the absence of a reference alignment.

The paper is organized as follows.

Section 2 introduces the foundational components of the proposed *Privacy-Preserving Schema Matching* approach. In particular, we describe the classical instance-based schema matching methods that serve as the core matching algorithms, and the well-known *k-anonymity* techniques employed to protect data privacy. These elements, drawn from well-established literature in both schema matching and privacy preservation, are combined into a unified framework designed to support schema alignment tasks while minimizing disclosure risks. In the Experimental Setup section 3, we present both the Dataset Preparation, i.e., the datasets used in our experiments, and the Experimental Workflow employed. Section 4 - Experimental Results presents some results. Section 5 is devoted to the related works. Section 6 contains conclusions, and directions for future work.

2. Building Blocks of the Proposed Framework

This section introduces the foundational components of the proposed *Privacy-Preserving Schema Matching* approach. We describe some classical instance-based schema matching methods implemented in the Valentine Framework [9] (section 2.1), and some well-known *k-anonymity* techniques (section 2.2).

2.1. Schema Matching with the Valentine Framework

In this section, we briefly describe the *Valentine* framework [9], focusing both on the schema matching methods we selected to evaluate the performance of our approach on *k-anonymized* data (see section 3), and on the evaluation techniques proposed within Valentine to assess their effectiveness.

Valentine is an extensible open-source experiment suite designed to execute and organize large-scale automated schema matching experiments on tabular data. It allows the application of multiple schema matching methods to pairs of denormalized tables, leveraging available metadata such as table and attribute names, data types, and the actual data values.

An important feature of Valentine is that it produces a ranked list of matching attribute pairs, ordered by the confidence score assigned by each method. This is particularly suitable for dataset discovery scenarios, where ranked results allow users and systems to explore candidate matches more efficiently and assess match quality based on position in the ranking.

To evaluate the effectiveness of the methods, Valentine introduces the *Recall@ground truth* metric, defined as follows:

Recall@ground truth [9]. Let $k = |\text{ground truth}|$. Then:

$$\text{Recall@ground truth} = \frac{\text{Number of top-}k \text{ relevant matches}}{k}$$

This metric reflects how many of the correct matches appear in the top- k results returned by a method, thus providing an intuitive measure of how helpful the ranked list is for a human evaluator or downstream process that considers only a limited number of top suggestions.

Below is a brief description of the instance-based matchers by Valentine utilized in our work.

Distribution-based Matcher. Distribution-based Matching [10] is an *instance-based* approach that identifies matches between columns by comparing the distributions of their values; more precisely, the similarity between columns is measured using the *Earth Mover's Distance (EMD)*, which quantifies the minimal effort required to transform one distribution into another. The algorithm first forms clusters of similar columns based on pairwise *EMDs*; then, columns that share many values, or that are both strongly associated with a third column, are considered to match.

Jaccard Distance Matcher. This is an instance-based matcher that uses Jaccard to calculate all pairwise column similarities; two values are considered identical if their Levenshtein distance (or other similarity measure) is less than a certain threshold.

2.2. K-anonymity and K-anonymization Tools

Access to data is fundamental for open science, supporting transparency, reproducibility, and research progress. However, privacy regulations like the GDPR [11] restrict the publication and sharing of datasets containing sensitive information.

A key challenge is the risk of re-identification, where individuals can be recognized even if direct identifiers (e.g., names) are removed. This risk increases when datasets include *quasi-identifiers*, such as age, postal code, and occupation, which combined can uniquely identify individuals.

To reduce this risk, *k-anonymity* requires that each combination of quasi-identifier values appears in at least k records, making each individual indistinguishable from $k - 1$ others. This reduces re-identification chances but causes some information loss.

K-anonymity has known limitations, including vulnerability to homogeneity and background knowledge attacks, which happen when all records share the same quasi-identifier values.

Optimal Lattice Anonymization. Optimal Lattice Anonymization (OLA) [12] is an algorithm for achieving *k-anonymity* by generalizing and suppressing data to protect sensitive information. It ensures each record is indistinguishable from at least $k - 1$ others, reducing re-identification risk. OLA searches a lattice of possible generalizations, where each node represents a different way of generalizing attributes. Using *predictive tagging*, it efficiently prunes the search by skipping nodes that are guaranteed to satisfy or violate *k-anonymity*, cutting down computations. After identifying all *k-anonymous* nodes,

OLA keeps only those with minimal generalization and selects the best one based on information loss measures like the *Discernibility Metric*, choosing the solution with the least loss of data utility.

Mondrian. The *Mondrian* algorithm [13] is a *greedy* top-down recursive partitioning method for achieving k -anonymity. It partitions the multidimensional *quasi-identifier* space into regions containing at least k indistinguishable records.

At each step, the attribute with the largest range is selected and the domain is split along the median, proceeding recursively until the k -anonymity constraint is satisfied or no further splits are possible. Mondrian supports both *global generalization*, which uniformizes the treatment of instances, and *local generalization*, which preserves specific details in some partitions, balancing privacy and utility. This approach, characterized by computational efficiency and simplicity, is widely adopted in practical data anonymization applications.

3. Experimental Setup

The goal of our experimental study is to assess how k -anonymity impacts the performance of instance-based schema matching algorithms. The evaluation metrics used are those defined in section 2.1.

The first type of evaluation adopted consists of classical comparison against a known *gold standard*. This mode, which we refer to as *Evaluation against the Gold Standard*, is applied to both the results obtained on the cleartext data and those derived from the k -anonymized data.

The second type of evaluation, which we will call *Evaluation against Cleartext Results*, compares the results obtained on k -anonymized data to those obtained on the original cleartext datasets, treating the latter as a gold standard proxy. This type of evaluation is strongly inspired by the methodology proposed in [6], which analyzes the effect of k -anonymization on the performance of machine learning models. In our context, we apply this idea to the domain of schema matching, focusing on the impact of privacy-preserving transformations on schema alignment effectiveness. Specifically, we evaluate and compare schema matching results obtained from k -anonymized datasets with those derived from original, non-anonymized (cleartext) datasets. This approach enables us to quantify how k -anonymity influences result quality, even in the absence of a reference alignment.

3.1. Dataset Preparation

The technique we adopted to construct the datasets used in our experiments is inspired by Valentine’s [9], where synthetic matching challenges are generated by systematically partitioning and perturbing existing tables. This includes horizontal and vertical partitioning of tables, as well as noise injection into schema information and instance values. We adapt and extend this dataset fabrication strategy to better align with the objectives of our study. Unlike the original method, which perturbs both schema and data, we focus exclusively on perturbing instance values. This design choice is driven by our goal of evaluating instance-based schema matching methods when applied to k -anonymized data.

The dataset preparation procedure is the following.

(1) Selection of the starting dataset. The starting point is a real dataset commonly used in studies on k -anonymization, namely the *Adult Dataset*, which contains 45,222 records derived from the 1994 United States Census. We consider the following attributes: sex, age, race, marital-status, education, native-country, workclass, and salary-class. This dataset is particularly suitable because it is already integrated into existing frameworks for the evaluation of k -anonymization, and it has known and available generalization hierarchies useful for applying anonymization techniques.

(2) Generation of datasets to be compared. Starting from the original dataset, to make the task more realistic, dummy columns were created for some attributes that are derived from the original ones but with different distributions. For example, two new columns, *native-country_1* and *native-country_2*, were added for the native-country attribute; these columns can be interpreted as representing two different country-related contexts: the country of current residence and the country of work, with

values initially extracted from the original column but modified to have different distributions. Similar operations have been performed on other columns, creating new variants with slightly perturbed distributions. Then we build pairs of datasets (S_i, S_j) through a controlled *horizontal overlap* (i.e., a certain percentage of records shared between the two datasets) and a controlled *vertical overlap* (i.e., a certain percentage of shared attributes).

(3) Perturbation of the datasets. To simulate realistic noisy data conditions, we apply a phase of *controlled perturbation* to the generated datasets. In particular, we use the *GeCo* tool [14] to introduce errors into the records, allowing a variable number of errors per record. Data corruption is performed based on the attribute type: for **textual** attributes, random typos are introduced based on keyboard key proximity; for **numeric** attributes, values are randomly modified following the statistical distribution of the original values. This phase allows simulating syntactic or semantic noise commonly present in real data, and evaluating the robustness of schema matching algorithms in the presence of such perturbations.

3.2. Experimental Workflow

The experimental workflow consists of the following steps:

1. **Matching on cleartext data:** For each dataset pair we apply a schema matching algorithm to the original data and compute evaluation metrics using a known gold standard. We evaluate the following two instance-based schema matching methods, both available within the Valentine framework and described in section 2: *DistributionBased* and *JaccardDistanceMatcher*.
2. **Data anonymization:** We apply well-known k -anonymity algorithms (such as Mondrian or OLA) to generate multiple anonymized versions of each dataset by varying the value of k .
3. **Matching on anonymized data:** Each instance-based schema matcher is applied to the anonymized datasets.
4. **Performance comparison:** We perform both *Evaluation against the Gold Standard* and *Evaluation against Cleartext Results*, in order to assess how k -anonymization affects the effectiveness of schema matching methods under different evaluation perspectives.

4. Experimental Results

In this section we present and discuss some preliminary results.

Cleartext Data - Evaluation against the Gold Standard To evaluate the performance of matchers (*JaccardMatcher* and *DistributionMatcher*, see section 2.1) on the plaintext data, we use the classic comparison of results with a Gold Standard and calculate the *Recall@GroundTruth* (see Table 1).

Pair	<i>JaccardMatcher</i>	<i>DistributionMatcher</i>	Best Matcher
(S_1, S_2)	0.57	0.71	Distribution
(S_3, S_4)	0.67	0.92	Distribution

Table 1

Comparison of matchers on cleartext data (*Recall@GroundTruth*)

Overall, the *DistributionMatcher* proves to be the most effective in all two cases, demonstrating better quality in matching tasks compared to *JaccardMatcher* on cleartext data. This is to be expected, since dummy columns were created with slightly modified distributions: while Jaccard is based on overlap of values, Distribution captures similarities at the distribution level.

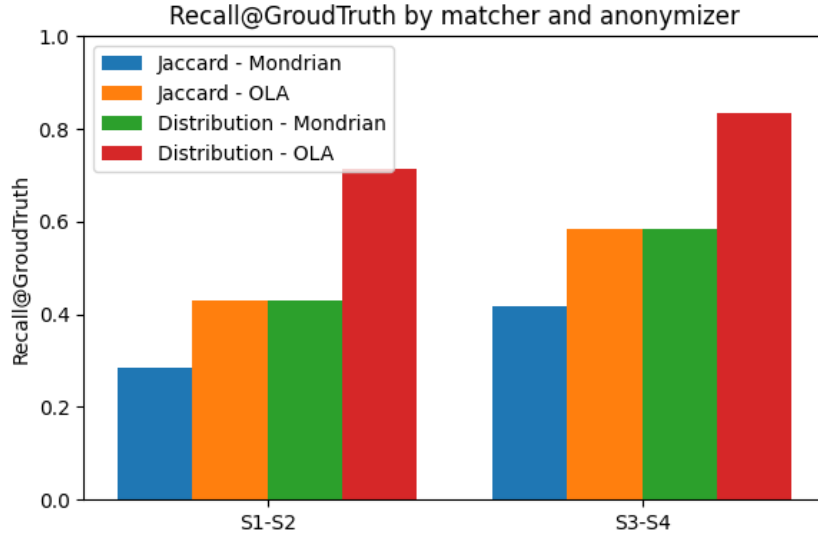


Figure 1: Chart of the Evaluation against the Gold Standard Results (*Recall@GroundTruth*)

Anonymized Data - Evaluation against the Gold Standard To assess which k-anonymizer between OLA and Mondrian allows for better quality schema matching tasks, we compared the values of *Recall@GroundTruth* obtained against the Gold Standard, for each dataset pair combination and for each matcher used (see Table 2).

Matcher	Pair	Mondrian	OLA
JaccardMatcher	(S1, S2)	0.29	0.43
JaccardMatcher	(S3, S4)	0.42	0.58
DistributionMatcher	(S1, S2)	0.43	0.71
DistributionMatcher	(S3, S4)	0.58	0.83

Table 2

Comparison of K-anonymizers (Mondrian vs OLA) on anonymized data (*Recall@GroundTruth*)

The analysis suggests that the quality of schema matching tasks with respect to the Gold Standard is better when done on anonymized data with OLA, for both the *JaccardMatcher* and *DistributionMatcher*. Regarding the effect of the matcher used, OLA improves more with the *DistributionMatcher* than with the *JaccardMatcher*.

Figure 1 summarizes the results for this evaluation.

Evaluation against Cleartext Results In this analysis, the *JaccardMatcher* and *DistributionMatcher* were considered as two semantic similarity tasks between pairs of datasets: the objective is to test which of the two k-anonymizers, OLA or Mondrian, is more effective in preserving the original similarity as measured by these tasks.

For each matcher *JaccardMatcher* and *DistributionMatcher*, for each pair of datasets, for each k-anonymizer, we compared the matches obtained between the k-anonymized datasets with those obtained between the plain-text datasets, calculating the *Recall@GroundTruth* (see Table 3).

We can see, the trend is very similar to the previous one: also here, OLA is always better than Mondrian regardless of the matcher adopted and of the dataset pairs used. This happens because OLA checks all possible generalizations and then selects the one with the best utility.

Figure 2 summarizes the results for this evaluation.

Finally, to evaluate the impact of k-Anonymity on matching quality, that is, to analyze how the degree of anonymization (parameter *k* in k-anonymity) influences instance-based schema matching quality, we compared the results of matching on k-anonymized data against the results obtained on cleartext data.

Matcher	Pair	Mondrian	OLA
JaccardMatcher	(S1, S2)	0.43	0.57
JaccardMatcher	(S3, S4)	0.50	0.67
DistributionMatcher	(S1, S2)	0.43	0.86
DistributionMatcher	(S3, S4)	0.58	1.00

Table 3

Evaluation against Cleartext Results ($Recall@GroundTruth$)

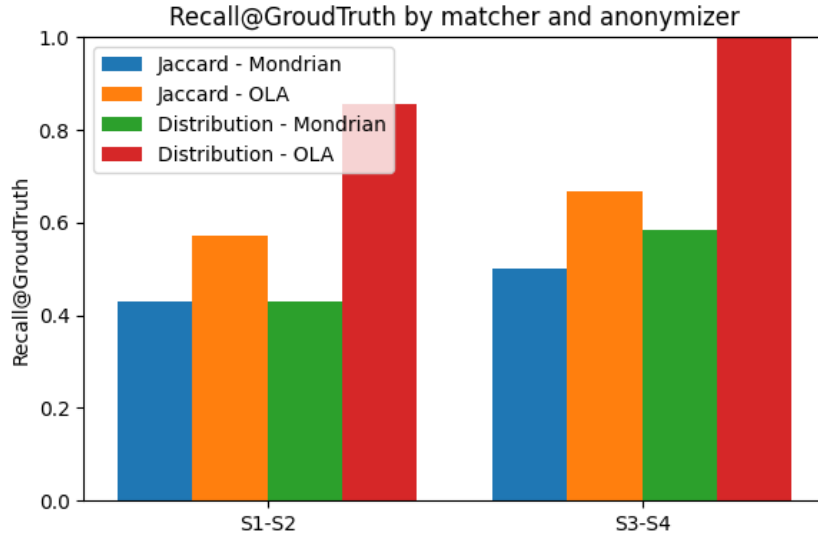


Figure 2: Chart of Evaluation against Cleartext Results ($Recall@GroundTruth$)

Table 4 presents the values of $Recall@GroundTruth$ for varying values of $k \in \{3, 5, 10\}$ using *OLA* as anonymization techniques (the ones for *Mondrian* are very similar and therefore are not reported).

When using *OLA* for anonymization, the matching results are good with $k = 3$ for both matchers. At $k = 5$, the results get a little worse, but remain acceptable. When $k = 10$, the results get much worse. This means that too high a level of anonymization makes it difficult to recognize similarities. Intermediate values such as $k = 5$ can be a good compromise between privacy protection and matching quality. As a last observation, across all our evaluations, we see the same trend: as k increases, the matching quality decreases. This suggests that the main driver is the level of privacy required, rather than the specific anonymization method used.

k	Matcher	Recall@GroundTruth (OLA)
3	DistributionMatcher	0.86
	JaccardMatcher	0.71
5	DistributionMatcher	0.71
	JaccardMatcher	0.43
10	DistributionMatcher	0.29
	JaccardMatcher	0.14

Table 4

$Recall@GroundTruth$ values for (S_1, S_2) using *OLA* with increasing k -anonymity.

5. Related works

The paper [15] presents a privacy-preserving approach to schema and data matching, where sensitive information is protected through embedding techniques that enable matching operations without

revealing original values. As discussed in the paper, the evaluation of the performance of matching techniques on anonymised data is carried out by comparing precision and recall metrics obtained in the embedded space (i.e., after anonymisation) with those obtained in the original space (i.e., without anonymisation).

In [7], the authors highlight that *instance-level schema matching* requires querying actual data, which can raise privacy concerns. To address this, they propose *SMAT* [7], a schema-level matching method that operates exclusively on local schemata, leveraging table and attribute names along with their descriptions. However, a key limitation of this approach is that it cannot be considered *truly privacy-preserving* unless metadata—such as table and attribute names or descriptions—are either inherently non-sensitive or properly anonymized before being shared with the data integration system. To overcome this, we propose a *Privacy-Preserving Schema Matching* technique that is purely instance-based and operates solely on anonymized data. This ensures that no information about schemas or data is disclosed to the integration system. A similar idea is presented in [16], where the authors introduce *Privacy-Preserving Quick Ontology Mapping* (P2QOM): each client’s ontology is transformed into a set of obfuscated features, which are then shared with the data integration framework.

PRISMA (Privacy-Preserving Schema Matcher) [17] is based on a privacy-preserving schema matching method that avoids access to raw data and relies only on metadata, such as functional dependencies and frequency distributions; it represents schemas as graphs and generates embeddings to compare attributes and identify one-to-one correspondences. Like our method, PRISMA also performs privacy-preserving schema matching without accessing the cleartext data. PRISMA, however, employs specific tools developed specifically for this purpose, while our method relies on classical schema matching techniques combined with well-established k -anonymity methods.

A recent trend in schema matching research explores the use of Large Language Models (LLMs) to improve both matching accuracy and scalability. For instance, the *Magneto* framework [18] proposes a two-phase approach that combines small and large language models to efficiently retrieve and rerank candidate matches. Its architecture relies heavily on both schema-level metadata (e.g., column names) and instance-level data (e.g., sampled values). In [19], the authors propose leveraging LLMs—such as GPT-3.5 and GPT-4—to generate semantic correspondences between attributes of heterogeneous schemas in the healthcare domain. The proposed method relies exclusively on textual metadata (e.g., attribute names and descriptions), deliberately avoiding access to instance-level data in order to reduce exposure to sensitive information.

The use of external LLM services, which entails transmitting data to third-party providers (e.g., OpenAI APIs), violates the requirements of a strictly privacy-preserving scenario, where no sensitive information—neither schema-related nor data-related—should be exposed outside the local environment. While this issue can be mitigated by deploying LLMs in a local environment, such a solution poses significant challenges in terms of computational resources, model maintenance, and the need for technical expertise. Moreover, locally deployed models may not always match the performance or up-to-dateness of cloud-based LLMs, potentially limiting their effectiveness in complex schema matching tasks.

6. Conclusions and Future Work

In this paper, we presented a preliminary study on privacy-preserving instance-based schema matching using k -anonymized data. Our results show that it is possible to achieve effective schema alignment while protecting sensitive information by operating exclusively on anonymized datasets, without relying on schema metadata that may be incomplete or unavailable.

We also demonstrated how comparing matching results on anonymized versus cleartext data provides valuable insights into the trade-off between privacy and data utility in schema matching scenarios.

As future work, we plan to extend our evaluation to real-world clinical datasets [8], where privacy concerns are paramount. Additionally, we aim to explore alternative anonymization techniques beyond k -anonymity, investigate more advanced instance-based matching algorithms, and develop strategies to

further minimize information loss while maintaining high matching accuracy.

Another possible future work involves the study of the trade-off between data utility and privacy protection. By *data utility* we mean how useful a dataset continues to be after applying protection techniques, such as *k*-anonymization. In general, the greater the protection applied, the greater is also the loss of useful information. Therefore, it is important to understand how to find a good trade-off between privacy and usefulness. In this work, this balance was evaluated indirectly by considering how the performance of schema matching algorithms changes when moving from the original data to anonymized data. Conduct a systematic assessment of the privacy-utility tradeoff, measuring both privacy gains (how much the risk of re-identification is reduced) and utility losses, not only in terms of the accuracy of the matching scheme but also with information loss metrics, so as to have more guidance on appropriate *k* values in practice.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to check grammar and spelling. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] S. Bergamaschi, D. Beneventano, F. Guerra, M. Orsini, Data integration, Handbook of conceptual modeling: theory, practice, and research challenges (2011) 441–476.
- [2] G. Simonini, L. Zecchini, S. Bergamaschi, F. Naumann, et al., Entity resolution on-demand, Proceedings of the VLDB Endowment 15 (2022) 1506–1518.
- [3] P. Christen, D. Vatsalan, V. S. Verykios, Challenges for privacy preservation in data integration, J. Data and Information Quality 5 (2014). URL: <https://doi.org/10.1145/2629604>. doi:10.1145/2629604.
- [4] P. Christen, T. Ranbaduge, R. Schnell, Linking Sensitive Data - Methods and Techniques for Practical Privacy-Preserving Information Sharing, Springer, 2020. URL: <https://doi.org/10.1007/978-3-030-59706-1>. doi:10.1007/978-3-030-59706-1.
- [5] L. Trigiante, D. Beneventano, S. Bergamaschi, Privacy-preserving data integration for digital justice, in: International Conference on Conceptual Modeling, Springer, 2023, pp. 172–177.
- [6] D. Slijepcevic, M. Henzl, L. D. Klausner, T. Dam, P. Kieseberg, M. Zeppelzauer, *k*-anonymity in practice: How generalisation and suppression affect machine learning classifiers, CoRR abs/2102.04763 (2021). URL: <https://arxiv.org/abs/2102.04763>. arXiv:2102.04763.
- [7] J. Zhang, B. Shin, J. D. Choi, J. C. Ho, SMAT: an attention-based deep learning solution to the automation of schema matching, in: L. Bellatreche, M. Dumas, P. Karras, R. Matulevicius (Eds.), Advances in Databases and Information Systems - 25th European Conference, ADBIS 2021, Tartu, Estonia, August 24-26, 2021, Proceedings, volume 12843 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 260–274. URL: https://doi.org/10.1007/978-3-030-82472-3_19. doi:10.1007/978-3-030-82472-3_19.
- [8] Observational Health Data Sciences and Informatics, The Book of OHDSI, Independently published, 2019. URL: <https://ohdsi.github.io/TheBookOfOhdsi/>.
- [9] C. Koutras, G. Siachamis, A. Ionescu, K. Psarakis, J. Brons, M. Fragkoulis, C. Lofi, A. Bonifati, A. Katsifodimos, Valentine: Evaluating matching techniques for dataset discovery, in: 37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021, IEEE, 2021, pp. 468–479. URL: <https://doi.org/10.1109/ICDE51399.2021.00047>. doi:10.1109/ICDE51399.2021.00047.
- [10] M. Zhang, M. Hadjieleftheriou, B. C. Ooi, C. M. Procopiuc, D. Srivastava, Automatic discovery of attributes in relational databases, in: T. K. Sellis, R. J. Miller, A. Kementsietsidis, Y. Velegrakis (Eds.), Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD

- 2011, Athens, Greece, June 12-16, 2011, ACM, 2011, pp. 109–120. URL: <https://doi.org/10.1145/1989323.1989336>. doi:10.1145/1989323.1989336.
- [11] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation), <https://eur-lex.europa.eu/eli/reg/2016/679/oj>, 2016. Accessed: 2025-06-30.
 - [12] K. El Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, J. Bottomley, A globally optimal k -anonymity method for the de-identification of health data, *Journal of the American Medical Informatics Association* 16 (2009) 670–682. URL: <https://www.sciencedirect.com/science/article/pii/S1067502709001236>. doi:<https://doi.org/10.1197/jamia.M3144>.
 - [13] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Mondrian multidimensional k -anonymity, in: *Proceedings of the 22nd International Conference on Data Engineering (ICDE)*, IEEE Computer Society, 2006, pp. 25–25. URL: <https://doi.org/10.1109/ICDE.2006.1>. doi:10.1109/ICDE.2006.1.
 - [14] K.-N. Tran, D. Vatsalan, P. Christen, GECO: An online personal data generator and corruptor, in: *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013)*, ACM, 2013, pp. 2473–2476.
 - [15] M. Scannapieco, I. Figotin, E. Bertino, A. K. Elmagarmid, Privacy preserving schema and data matching, in: C. Y. Chan, B. C. Ooi, A. Zhou (Eds.), *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Beijing, China, June 12-14, 2007, ACM, 2007, pp. 653–664. URL: <https://doi.org/10.1145/1247480.1247553>. doi:10.1145/1247480.1247553.
 - [16] T. Amagasa, F. Zhang, J. Sakuma, H. Kitagawa, A scheme for privacy-preserving ontology mapping, in: *Proceedings of the 18th International Database Engineering & Applications Symposium, IDEAS '14*, Association for Computing Machinery, New York, NY, USA, 2014, p. 87–95. URL: <https://doi.org/10.1145/2628194.2628232>. doi:10.1145/2628194.2628232.
 - [17] J. Hellenberg, F. Mahling, L. Laskowski, F. Naumann, M. Paganelli, F. Panse, PRISMA: A privacy-preserving schema matcher using functional dependencies, in: A. Simitsis, B. Kemme, A. Queralt, O. Romero, P. Jovanovic (Eds.), *Proceedings 28th International Conference on Extending Database Technology, EDBT 2025*, Barcelona, Spain, March 25-28, 2025, OpenProceedings.org, 2025, pp. 297–309. URL: <https://doi.org/10.48786/edbt.2025.24>. doi:10.48786/EDBT.2025.24.
 - [18] Y. Liu, E. Peña, A. S. R. Santos, E. Wu, J. Freire, Magneto: Combining small and large language models for schema matching, *CoRR abs/2412.08194* (2024). URL: <https://doi.org/10.48550/arXiv.2412.08194>. doi:10.48550/ARXIV.2412.08194. arXiv:2412.08194.
 - [19] M. Parciak, B. Vandevoort, F. Neven, L. M. Peeters, S. Vansummeren, Schema matching with large language models: an experimental study, in: *Proceedings of Workshops at the 50th International Conference on Very Large Data Bases, VLDB 2024*, Guangzhou, China, August 26-30, 2024, VLDB.org, 2024, pp. 1–12. URL: <https://vldb.org/workshops/2024/proceedings/TaDA/TaDA.8.pdf>.