IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 March 2025, accepted 17 April 2025, date of publication 22 April 2025, date of current version 6 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3563309

APPLIED RESEARCH

Scoring System for Quantifying the Privacy in Re-Identification of Tabular Datasets

JAKOB FOLZ[1,2], MANJITHA D. VIDANALAGE[2], ROBERT AUFSCHLÄGER[2], AMAR ALMAINI[2], MICHAEL HEIGL[2], DALIBOR FIALA[1], AND MARTIN SCHRAMM[2]

[1]Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, 301 00 Pilsen, Czech Republic
[2]Institute ProtectIT, Faculty of Computer Science, Deggendorf Institute of Technology, 94469 Deggendorf, Germany

Corresponding author: Jakob Folz (jakob.folz@th-deg.de)

This work was supported in part by the Research Project EAsyAnon "Verbundprojekt: Empfehlungs- und Auditsystem zur Anonymisierung" from German Federal Ministry of Education and Research (BMBF) under the Umbrella of the Funding Guideline "Forschungsnetzwerk Anonymisierung Für eine Sichere Datennutzung" under Grant 16KISA128K, and in part by the Package NextGenerationEU.

ABSTRACT This study introduces a System for Calculating Open Data Re-identification Risk (SCORR), a framework for quantifying privacy risks in tabular datasets. SCORR extends conventional metrics such as k-anonymity, l-diversity, and t-closeness with novel extended metrics, including uniqueness-only risk, uniformity-only risk, correlation-only risk, and Markov Model risk, to identify a broader range of re-identification threats. It efficiently analyses event-level and person-level datasets with categorical and numerical attributes. Experimental evaluations were conducted on three publicly available datasets: OULAD, HID, and Adult, across multiple anonymisation levels. The results indicate that higher anonymisation levels do not always proportionally enhance privacy. While stronger generalisation improves k-anonymity, l-diversity and t-closeness vary significantly across datasets. Uniqueness-only and uniformity-only risk decreased with anonymisation, whereas correlation-only risk remained high. Meanwhile, Markov Model risk consistently remained high, indicating little to no improvement regardless of the anonymisation level. Scalability analysis revealed that conventional metrics and Uniqueness-only risk incurred minimal computational overhead, remaining independent of dataset size. However, correlation-only and uniformity-only risk required significantly more processing time, while Markov Model risk incurred the highest computational cost. Despite this, all metrics remained unaffected by the number of quasi-identifiers, except t-closeness, which scaled linearly beyond a certain threshold. A usability evaluation comparing SCORR with the freely available ARX Tool showed that SCORR reduced the number of user interactions required for risk analysis by 59.38%, offering a more streamlined and efficient process. These results confirm SCORR's effectiveness in helping data custodians balance privacy protection and data utility, advancing privacy risk assessment beyond existing tools.

INDEX TERMS Anonymization, privacy, re-identification risk, GDPR, uniqueness, uniformity, correlation, open data.

I. INTRODUCTION

In the contemporary landscape, the scope of data is extensive, experiencing collection and processing at an unprecedented pace. From business and technology to healthcare and governance, data plays a pivotal role in shaping the world. Data empowers organisations to make well-informed decisions, optimize processes, and deliver personalised experiences to end-users. Moreover, data fuels advancements in artificial intelligence and machine learning, reshaping interactions with technology. The concept of Open Data is characterized by the idea of making data freely available and accessible to the public, with minimal restrictions on its use, distribution, and reuse [1], [2], [3]. However, the benefits of Open

The associate editor coordinating the review of this manuscript and approving it for publication was Sedat Akleylek.

Data are accompanied by significant challenges, particularly concerning data privacy and security [4], [5], [6], [7], [8]. The possible exposure of sensitive and personally identifiable information upon releasing datasets raises ethical and legal concerns regarding data usage and protection. Notable examples include the Cambridge Analytica scandal, where 50 million Facebook profiles were collected without consent [9], and Google's DeepMind accessing 1.6 million patient records from London hospitals to develop predictive tools [10]. Such cases highlight risks of privacy breaches, misuse of medical data, and diminished trust in institutions [11], [12].

Achieving a balance between utility and privacy becomes crucial to prevent potential misuse of data. Anonymisation addresses this by removing or replacing identifiable information, making it difficult for intruders to link data to specific individuals [13], [14], [15].

However, even with anonymisation in place, achieving a balance between utility and privacy remains challenging, as several incidents have demonstrated that individuals in anonymised data can still be re-identified. In 2014, the New York City Taxi & Limousine Commission released an anonymised dataset of 173 million taxi journeys, which was de-anonymised within an hour to re-identify vehicles and drivers [16]. Similarly, Netflix's 2006 release of 100 million anonymised movie ratings was de-anonymised within 16 days by correlating it with IMDb data [17]. That same year, AOL exposed 20 million web search queries from 650,000 users; despite removing usernames, individuals were still identifiable through URLs [18].

These incidents highlight the persistent risks associated with anonymised data, emphasising the need to verify anonymisation methods to protect individuals privacy. Despite these efforts, guaranteed privacy remains a challenge, as re-identification techniques can be used to link anonymised data to individuals, leading to significant breaches [19]. Addressing these issues requires robust mechanisms to assess the quality of anonymisation and mitigate re-identification risks. A key question driving this research is: *How can the quality and robustness of anonymisation in tabular datasets be systematically evaluated to minimise re-identification risks while preserving data utility?* To explore this, the work investigates privacy concerns associated with the re-identification of individuals in tabular datasets and proposes a scoring system to evaluate the degree of anonymisation and re-identifiability of such datasets. The primary objective is to develop a comprehensive scoring system that evaluates anonymised tabular datasets against multiple attack types using diverse re-identification metrics and risk analysis methodologies. The key contributions are outlined below.

1) **Comprehensive Scoring System:** SCORR introduces a comprehensive scoring system that integrates a diverse range of metrics to assess re-identification risk across multiple attack types, each assuming different intruder knowledge. Unlike previous approaches, SCORR extends beyond conventional metrics such as k-anonymity, l-diversity, and t-closeness by incorporating additional risks, including uniqueness, uniformity, and correlation attacks. This holistic approach provides a more technically robust and detailed assessment of vulnerabilities, enhancing the effectiveness of privacy risk analysis.
2) **Risk Analysis Framework:** SCORR extends beyond simple privacy score evaluation by providing a detailed re-identification risk analysis. Using conventional metrics, it assesses dataset compliance with legal regulations, while extended metrics categorise re-identification risk as low, medium, or high. By integrating both conventional and extended metrics, SCORR offers a comprehensive risk overview, assisting users in making informed decisions about public dataset release.
3) **Minimal User Interaction:** SCORR prioritises ease of use and accessibility, especially for users without specialised expertise in data privacy. It streamlines the decision-making process for dataset release by requiring only two user interactions: dataset upload and selecting relevant attributes. The system then automates the analysis, assessing worst-case scenarios to provide a comprehensive re-identification risk overview.

Based on the contributions outlined, the remainder of this article is structured as follows. Section II provides an overview of the key concepts and theories related to privacy and the anonymisation of tabular datasets. In Section III, we review related work and highlight the distinctions between our approach and existing studies. Section IV offers a detailed explanation of SCORR's functionality and operation, covering its initialisation, metrics, and risk analysis processes. Section V presents and discusses the results of our evaluation, focusing on the metrics used and the scalability of our SCORR framework. Section VI concludes the article by summarising the key findings. Finally, Section VII addresses the limitations of our work and outlines potential directions for future research.

II. BACKGROUND AND NOTATION

A. TYPES OF DATASETS

In this study, we classify datasets into two distinct types. The first type is the event-level dataset, which captures and records individual events as they occur. Each entry in this dataset corresponds to a specific event, with associated attributes such as event type, timestamp, and the relevant person or entity involved. This format allows for the possibility of multiple records for a single person or entity. Examples of such datasets include bank transaction records or patient logs from a hospital.

The second type is at the person-level, with each record (or row) representing a unique person. The columns capture various attributes or characteristics of the individual, including, but not limited to, their name, address, age, and other

demographic information. Importantly, this dataset includes at most one record per person. Examples of such datasets include employee records from a workplace or a voter registry. In this study, we classify datasets into two distinct types.

B. TYPES OF ATTRIBUTES

In tabular datasets, attributes or variables are categorised based on their inherent characteristics. Nominal variables, a type of categorical data, represent distinct categories that lack any intrinsic order. Examples include gender (male, female), colour (red, green, blue), and product type (electronics, clothing, food). These categorical values can be expressed in different formats, such as text or numerical codes. For instance, gender in a survey might be recorded as ''male'' and ''female'' (text) or as ''1'' and ''2'' (numerical codes), with both formats conveying equivalent information. Numeric attributes are variables with continuous and discrete values, such as age, height, weight, or income. Binary attributes are also part of categorical data but are limited to two possible values, such as yes/no, true/false, or 0/1. The type of attribute can influence the selection of anonymisation techniques and the metric to measure the risk of re-identification in tabular datasets. In the context of privacy, attributes in tabular datasets are commonly classified into Direct Identifiers (DIs), Quasi-Identifiers (QIs), and Sensitive Attributes (SAs) [20]. DIs refer to attributes that can be used to identify an individual directly, such as name, address, social security number, passport number, or email address [21]. A QI is a set of attributes that, when combined, have the potential to identify an individual, even if DIs are removed or anonymised. It includes several attributes such as age, gender, postcode, occupation, or education level. SAs are attributes in tabular datasets that, if exposed or disclosed, could raise privacy concerns for individuals or groups. SAs are attributes which are typically considered private or protected. Examples include health status, salary, or criminal record. These classifications are crucial for understanding and mitigating privacy risks associated with data sharing and use.

C. DATASET NOTATION

To develop the equations for each metric, we assume an event-level dataset with n records, m attributes (elements) in the QI, one SA, and data for p distinct persons. In a person-level dataset, the number of persons equals the number of records ($p = n$), as each record represents a distinct person. Conversely, in an event-level dataset, the number of persons can be less than or equal to the number of records ($p \leq n$) since multiple records can represent multiple events related to a single person. The following assumptions are made for the calculations.

To formulate the equations for each metric, we consider an event-level dataset with n records, m attributes in the QI, one SA, and data for p distinct persons. In a person-level dataset, each record corresponds to a unique person, meaning $p = n$. In contrast, an event-level dataset allows multiple records to represent events associated with the same person, resulting in $p \leq n$. The following assumptions are made for the calculations.

1) SCORR evaluates the worst-case risk scenario, assuming the intruder knows the QI values of the target data subject.
2) The intruder attempts to infer a single SA value associated with the target data subject.
3) In an event-level dataset, a specific attribute, Person ID, is used to reference the relevant individual.

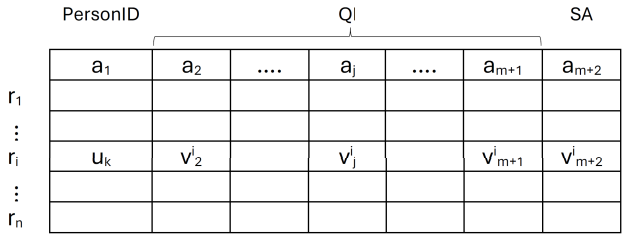


FIGURE 1. Notations of the dataset.

Figure 1 presents the notational conventions used in the equations. To streamline the discussion, we consider all attributes, except the Person ID and the SA, to be elements of QI, under the assumption that our data model includes exactly one QI and one SA. The indices i, j, and k are used to represent records (rows), attributes (columns), and distinct persons, respectively.

$$\begin{aligned}
\text{Attributes} &= \{a_1, a_2, \ldots, a_{m+2}\} \\
\text{Person ID Attribute} &= \{a_1\} \\
\text{QI} &= \{a_2, a_3, \ldots, a_{m+1}\} \\
\text{SA} &= \{a_{m+2}\} \\
\text{Records} &= \{r_1, r_2, \ldots, r_n\} \\
\text{Distinct Persons} &= \{u_1, u_2, \ldots, u_p\} \\
\text{Value of } a_j \text{ in } i^{\text{th}} \text{ record} &= v_j^i \\
\text{Person ID Value in } i^{\text{th}} \text{ record} &= v_1^i = u_k \\
\text{QI Value in } i^{\text{th}} \text{ record} &= QI^i = \{v_2^i, \ldots, v_{m+1}^i\} \\
\text{SA Value in } i^{\text{th}} \text{ record} &= SA^i = v_{m+2}^i
\end{aligned}$$

D. CONVENTIONAL METRICS

Emerging from the beginning of privacy-preserving data publishing, three pivotal privacy models have gained widespread recognition and extensive usage. Consequently, they are often referred to as conventional metrics in our paper. Each model addresses different aspects of privacy protection, and together they contribute to a holistic approach.

1) k-ANONYMITY

k-anonymity is a privacy concept that masks individual identities in a dataset by ensuring that each record cannot

be uniquely identified by its QI and therefore safeguards against linking attacks. A dataset is considered k-anonymous if each record cannot be distinguished from at least $k-1$ other records based on the QI. k-anonymity is enforced through transformations like generalisation and suppression, which reduce data granularity while retaining significant patterns and statistical information. However, it is crucial to acknowledge that any generalisation algorithm used to achieve k-anonymity inevitably results in some loss of information from the original dataset [22].

2) l-DIVERSITY

While k-anonymity is a useful privacy measure, it has limitations. It is identified that it cannot protect a dataset against homogeneity attacks and background knowledge attacks [23]. To address this concern, l-diversity builds upon the foundation of k-anonymity by introducing the idea of ensuring that each equivalence class (a group of records with similar values of QI) contains at least l distinct sensitive values [23]. This criterion enhances individuals' privacy by introducing diversity and uncertainty to the data, thereby making it harder for adversaries to deduce specific sensitive values associated with individuals. By enforcing l-diversity, the dataset becomes more resilient against attacks that exploit attribute disclosure [23].

3) t-CLOSENESS

t-closeness was introduced to address the limitations of l-diversity in preventing attribute disclosure, particularly its vulnerability to skewness and similarity attacks [24]. A dataset satisfies t-closeness when the distribution of an SA within each equivalence class is sufficiently similar to the distribution of the SA across the entire dataset. This similarity is quantified by a threshold t, which ensures that the risk of disclosing sensitive attributes within individual equivalence classes is minimised by limiting the distance between these distributions.

E. EXTENDED METRICS

In addition to these conventional metrics, there are several other metrics that quantify the risk of re-identification across various attack types and domains. As a result, we refer to these as extended metrics in our paper, highlighting their role in providing an additional layer of privacy protection beyond that offered by conventional metrics.

1) UNIQUENESS

Uniqueness is defined by the extent to which a value within a dataset can be distinguished from those in other entries [25]. For example, consider a dataset with the birth years of 10 individuals: [1991, 1991, 1991, 1993, 1993, 1993, 1993, 1993, 1993, 1999]. In this case, the value "1993" is less distinguishable due to its higher frequency, thereby offering greater privacy. Conversely, "1999" is notably unique and rare. When an individual's attribute values are distinct within a dataset, the risk of re-identification increases significantly. Unique individuals are more easily identified by matching their attribute values, making them more vulnerable to re-identification [11]. Uniqueness is commonly employed as a metric to assess re-identification risk in previous studies [26] and is closely related to the concept of k-anonymity. In SCORR, we specifically use the metric of *uniqueness-only risk* (R_{uq}) to assess the uniqueness of QI attributes. In this context, the probability P_d^i represents the relative occurrence of the QI value of the i^{th} record in the whole dataset. This probability, referred to as Duplication Probability, is calculated as shown in Equation 1.

$$P_d^i = \frac{f(QI^i)}{n} \in (0, 1], \tag{1}$$

where
P_d^i = Duplication Probability of i^{th} record
QI^i = QI value of i^{th} record
$f(QI^i)$ = Total occurrences of QI^i
n = Number of records

Subsequently, R_{uq} of the i^{th} record is calculated by Equation 2. This is done by taking the complement of the logarithmically transformed Duplication Probability associated with the i^{th} record P_d^i.

$$R_{uq}^i = 1 - \frac{\log_2[f(QI^i)]}{\log_2(n)} \in (0, 1], \tag{2}$$

Since R_{uq} is selected for SCORR and calculated for each record individually, the minimum, maximum, and mean are then determined to represent the overall risk of the dataset.

2) UNIFORMITY

Uniformity pertains to the likelihood of accurately identifying an individual based on consistent patterns in their behaviour or data. For instance, if only one student consistently achieves low scores on exams, it becomes easier to identify that student based on their performance. The more consistently an individual exhibits a particular behaviour across various situations, the more reliably that behaviour or data pattern can be associated with them. In other words, unique and consistent behavioural or data patterns increase the probability of precise identification [25]. For our system, we utilise the metric *uniformity-only risk* (R_{uf}) to assess uniformity characteristics. This metric is particularly applicable in scenarios involving event-level data, where multiple records may correspond to a single individual. When an individual demonstrates regularity in attribute values (high uniformity), the probability of their successful re-identification increases. This concept introduces an additional layer of privacy protection beyond mere uniqueness. Within SCORR, the R_{uf} metric assesses the association between an individual and a specific or a set of attribute values, focusing exclusively on the uniformity of QI values. The metric is computed for each record as described in Equation 3, considering both the entire

set of QI values and each QI element separately.

$$R^i_{uf} = P(\, u_k \mid QI^i \,) \in (0, 1], \tag{3}$$

where
R^i_{uf} = Uniformity-only Risk of i^{th} record
u_k = Person relevant to i^{th} record
QI^i = QI value of i^{th} record

3) CORRELATION

In addition to the uniqueness and uniformity of attribute values, the risk of re-identification is significantly influenced by the correlation between attributes when multiple attributes are analysed together. A high correlation between attributes increases the probability of inferring the value of one attribute based on the known values of other correlated attributes. For example, the level of education may be strongly correlated with salary, making it possible to estimate an individual's salary if their education is known. Attributes that exhibit strong correlations are therefore more vulnerable to re-identification when values from closely related attributes are available [25]. To specifically address the risk posed by correlated attributes, a specialised metric known as *Correlation-only Risk* (R_{co}) is used. This metric is designed to assess the correlation between the QI and the SA, providing a measure of protection against correlation attacks. The R_{co} is computed for each record using the equation specified in 4.

$$R^i_{co} = P(SA^i \mid QI^i) \in (0, 1], \tag{4}$$

where
R^i_{co} = Correlation-only Risk of i^{th} record
QI^i = QI Value of i^{th} record
SA^i = SA Value of i^{th} record

4) MARKOV MODEL RISK

The Markov Model Risk (R_{mm}), as introduced in [25], encapsulates key characteristics such as uniqueness, uniformity, and correlation. This model treats all attributes uniformly, without distinguishing between QI and SA. Utilising a Markov chain, the risk of re-identification is assessed by sequentially estimating attribute values based on their correlations, starting from a known attribute. In SCORR, where attributes are classified into QI and SA, we employ a modified version of the Markov Model to calculate the re-identification risk. This modification restricts the calculation to a single step, transitioning directly from the known QI to the SA. As a result, the process is simplified, avoiding the need for multiple sequential attribute-to-attribute calculations. This calculation is formally represented in Equation 5. Here, the term P^i_d denotes the uniqueness of the QI. The term $[1 - P(u_k|QI^i)]$ reflects the uniformity of the QI, while $[1 - P(SA^i|QI^i)]$ indicates the correlation between the QI and SA. Furthermore, $[1 - P(u_k|SA^i)]$ represents the uniformity of the SA.

$$R^i_{mm} = 1 - \Big[P^i_d \cdot [1 - P(u_k|QI^i)] \cdot [1 - P(SA^i|QI^i)] \cdot [1 - P(u_k|SA^i)]\Big] \in (0, 1], \tag{5}$$

where
R^i_{mm} = Markov Model risk of i^{th} record
P^i_d = Duplication Probability of i^{th} record
QI^i = QI Value of i^{th} record
SA^i = SA Value of i^{th} record
u_k = Person relevant to i^{th} record

F. DIFFERENTIAL PRIVACY AND SYNTHETIC DATA GENERATION

Differential privacy (DP), introduced by Dwork et al. [27], provides a mathematical framework to balance the trade-off between data utility and individual privacy in data analysis. It ensures that the inclusion or exclusion of any single individual in a dataset has a minimal impact on the outcome of an analysis, thereby offering a formal privacy guarantee. Typically, DP is implemented by introducing random noise to query outputs, making it difficult to infer specific details about any individual while preserving the overall statistical utility of the dataset.

A fundamental mechanism to achieve DP is the Laplace mechanism [28]. Given a query function f, this approach computes f on the dataset and adds noise sampled from a centred Laplace distribution. The scale of the noise is proportional to the inverse of the chosen privacy parameter ϵ, multiplied by the sensitivity of the function f. Sensitivity, in this context, quantifies the maximum possible change in the function's output due to the addition or removal of a single record. The selection of an appropriate noise scale is crucial in managing the privacy-utility trade-off and highly depends on the use cases of the dataset. While DP is effective for aggregate data publishing, its application to record-level Open Data publishing remains challenging [29]. The primary difficulty arises from the fact that future analyses of published data are unknown at the stage of noise addition, making it difficult to determine the appropriate sensitivity for noise generation. This concern is widely addressed in [30] as the Privacy-Flexibility-Accuracy trilemma.

An alternative approach to privacy-preserving data sharing is synthetic data generation (SDG). SDG techniques create artificial datasets that preserve the statistical properties of the original data, allowing analysts to perform computations using the same algorithms and pipelines as they would with real data while mitigating privacy risks. SDG has emerged as a more effective approach for integrating DP into record-level data publications [29]. Even the previous studies [31], [32] on using DP for record-level data publishing incorporate a synthetic data generation subprocess within the overall pipeline. Despite being a relatively new research

area, differentially private SDG has seen growing interest in the industry. The US National Institute for Standards and Technology (NIST) recently launched a challenge to develop DP-based SDG models for public-use datasets [33]. Similarly, the International Organisation for Migration and Israel's Ministry of Health have released differentially private synthetic datasets [34], [35].

It is important to distinguish synthetic data and perturbed data with noise addition from anonymised data generated using traditional anonymisation techniques such as generalisation and suppression. Unlike anonymisation, which reduces the granularity of the original data, synthetic and noised data aim to maintain granularity while ensuring privacy. Ideally, such data should be resilient against re-identification attacks, including linking attacks, homogeneity attacks, background knowledge attacks, similarity attacks, skewness attacks, and uniformity attacks, as they do not directly represent real individuals. However, correlation attacks remain a concern, as relationships between attributes are often preserved to maintain data utility. In practical scenarios, residual privacy risks persist in synthetic and noised data, necessitating empirical evaluation of their vulnerability to re-identification attempts [36]. Since real personal data is only present in the original dataset, privacy risks must be assessed by comparing synthetic or perturbed data with their original data [37]. This differs from anonymisation risk assessment methods used in SCORR, which evaluate risk solely based on anonymised datasets in a context where it does not have access to the original dataset. Consequently, risk metrics for synthetic and perturbed data differ from those used in SCORR, representing a distinct and emerging research area. Several ongoing studies such as [35], [37] focus on addressing these challenges, while further research efforts are discussed in detail in [38].

G. DATA UTILITY MEASUREMENT

Assessing the utility of an anonymised dataset involves evaluating how well it preserves the analytical value of the original data while ensuring privacy protection. Existing methods measure utility by comparing actual values or statistical properties between the anonymised and original datasets. These assessments are inherently a posteriori, as they evaluate utility after anonymisation and require access to both datasets. Common methods include evaluating the number of missing values, the number of records modified, and contingency table comparisons [39]. Statistical similarity measures (mean, covariance, and correlation), information loss measures, eigenvalue analysis, entropy comparison and distribution distance comparison are also employed. Moreover, multivariate utility measures such as linear and logistic regression, Cronbach's Alpha and Adjusted Cramer's V evaluate changes in accuracy between anonymised and original data. These measures help determine whether anonymised data remains suitable for specific analyses [39], [40], [41], [42]. However, when the original dataset is not available or not accessible for the utility assessment, direct comparisons become infeasible, making absolute utility measurement challenging and establishing it as a distinct research area. Therefore, within the scope of SCORR, data utility measurement is not addressed and will be considered in future work.

H. LEGAL ASPECTS

As the collection and utilisation of personal information continue to grow, concerns regarding privacy breaches and the potential misuse of data have prompted the development of legal frameworks designed to safeguard individuals' rights while allowing the benefits of data analysis. Re-identification risks have been a focal point in various legal frameworks, such as the European Union's General Data Protection Regulation (GDPR) [43], the US Privacy Act [44], and Brazil's Lei Geral de Proteção de Dados Pessoais (LGPD) [45]. The GDPR, in particular, recognises the importance of data anonymisation as a measure to enhance privacy [46]. It encourages organisations to employ anonymisation techniques to minimise the risk of re-identification and the ensuing privacy breaches. While the GDPR acknowledges the challenges associated with achieving true anonymisation, it emphasises the need to balance data utility with privacy protection. Although it does not prescribe specific anonymisation methods, the regulation mandates that techniques must be sufficiently robust to resist re-identification attempts. Organisations are required to assess and mitigate re-identification risks to ensure the security of personal data, thereby upholding individuals' rights in the face of evolving re-identification technologies. The GDPR does not explicitly delineate the criteria that publicly released datasets must meet to comply with its requirements. However, the Article 29 Data Protection Working Party [47] provides practical guidance on ensuring transparency in personal data processing under the GDPR. In this context, privacy models such as k-anonymity, l-diversity, and t-closeness are recommended to mitigate re-identification risks. To implement these models effectively, a minimum privacy parameter value of $k > 10$ is commonly recommended [48], consistent with guidance from the Working Party and the European Medicines Agency (EMA) [49]. Additionally, for t-closeness, the criteria of $t <= 0.5$ have been established to meet compliance [48]. Collectively, these recommendations form a standard framework for enhancing data privacy when releasing datasets to the public.

III. RELATED WORK

Various methods have been proposed to quantify re-identification risk, each tailored to specific data structures and anonymisation techniques. These methods rely on mathematical models designed for distinct attack scenarios, making their applicability highly context-dependent. This section reviews recent studies, categorising them based on their methodological approach and scope.

A. RE-IDENTIFICATION RISK METRICS

Several studies focus on quantifying re-identification risks using statistical and probabilistic models. The work in [25] introduces a Markov Model for calculating re-identification risk within a realistic threat model, where an intruder may have varying levels of knowledge, from a single attribute to full dataset awareness. It also integrates uniqueness, uniformity and correlation characteristics into the final risk score. However, this method does not distinguish between QI and sensitive SAs, as it assumes that their impact on privacy risk is determined by the intruder's prior knowledge and intentions, which limits its applicability for SCORR. Entropy-based approaches have also been explored, such as in [50], where a novel metric estimates re-identification risk by analysing conditional entropy between original and anonymised datasets. This method quantifies uniqueness and models the probability of mapping an intruder's background knowledge into the anonymised dataset. However, its applicability is limited since it can only be utilised when the probability distribution of mapping the original dataset to the anonymised dataset is known, making it unsuitable for SCORR's framework, where only the anonymised dataset is available. In contrast, re-identification algorithmic approaches like those presented in [51] introduce the re-identification ratio, which quantifies risk based on the success rate of various re-identification techniques. The metric measures the proportion of correctly matched records, reflecting both dataset vulnerability and the effectiveness of the re-identification algorithm. However, this approach requires significant computational resources and prior knowledge of re-identification strategies, which are not available in SCORR. Furthermore, a modification of k-anonymity, known as k-anonyMean, was introduced in [51] as an alternative risk metric. It provides the average k value across all equivalence classes, highlighting potential over-anonymisation.

B. ALTERNATIVE RISK ASSESSMENT TECHNIQUES

The study in [52] introduces a copula-based modelling approach that employs synthetic data generation to assess re-identification risks associated with sample-to-population attacks. However, as this method relies on external reference datasets, it is not directly applicable to SCORR's self-contained risk analysis. Other studies focus on evaluating the impact of data sanitisation techniques. The work in [53] explores how different anonymisation strategies, such as recoding, top coding, swapping, and adding noise, influence re-identification probability. By modelling intruder assumptions, this study provides comprehensive guidance on risk mitigation strategies. Dankar et al. [11] emphasise uniqueness as a primary measure of re-identification risk, particularly in clinical datasets. Given the challenge of directly measuring uniqueness, their study evaluates estimation techniques to support health data privacy compliance. However, it assumes full dataset availability rather than limited public releases. Another privacy risk framework, Anonymeter, introduced in [37], focuses on synthetic data. This tool evaluates singling out, linkability, and inference risks using an attack-based approach, where synthetic data is tested against adversarial models. Unlike SCORR, which relies on metric-based estimations, Anonymeter executes active attacks to measure vulnerability, making it less suited for evaluating anonymised datasets in isolation.

C. EXISTING RE-IDENTIFICATION RISK ASSESSMENT TOOLS

Several software tools have been developed to assist in privacy risk analysis and anonymisation. ARX [54], [55] is a widely used open-source tool that provides dataset anonymisation through configurable privacy models such as k-anonymity, l-diversity, and t-closeness. Users can define privacy constraints, apply generalisation and suppression techniques, and evaluate privacy-utility trade-offs. The Re-identifier Risk Ready Reckoner (R4) [26], developed by CSIRO Data61, quantifies risk using uniqueness-based metrics and supports anonymisation via generalisation and perturbation. R4 offers an interactive dashboard and an API for integration into enterprise data management systems. Google's Cloud Data Loss Prevention API [56] includes built-in risk assessment functionalities that calculate k-anonymity, l-diversity, k-Map, and δ-presence. However, it primarily focuses on classification, de-identification and redaction rather than a comprehensive evaluation of dataset re-identification risks.

D. COMPARISON WITH SCORR

SCORR differs from existing approaches by integrating multiple risk assessment models while maintaining computational efficiency and ease of use. Unlike attack-based re-identification models that require extensive computation, SCORR employs metric-based risk estimation, allowing risk assessment without prior knowledge of potential attacks. Moreover, SCORR introduces structured risk categorisation that extends beyond uniqueness, incorporating uniformity and correlation risks. Additionally, existing tools like ARX and R4 provide strong anonymisation capabilities but do not explicitly quantify re-identification risks across multiple attack scenarios. Google Cloud's DLP API supports some risk calculations but lacks a multi-metric risk model. By contrast, SCORR provides a comprehensive, structured, and interpretable approach to risk assessment, allowing data custodians to make informed decisions on dataset release. Table 1 summarises the key differences between SCORR and existing tools, highlighting its broader scope in addressing re-identification risks.

IV. SYSTEM FOR CALCULATING OPEN DATA RE-IDENTIFICATION RISK (SCORR)

A. INITIALIZATION

SCORR is an application with a modern user interface, designed to guide users through each step of the process.

TABLE 1. Comparison of SCORR, ARX, and R4 tools.

Feature/Aspect	SCORR	ARX	R4
Privacy Metrics	k-anonymity, l-diversity, t-closeness, uniqueness-only risk, uniformity-only risk, correlation-only risk, simplified Markov model risk	k-anonymity, l-diversity, t-closeness, δ-presence, δ-likeness, δ-disclosure, k-Map	Uniqueness-only risk
Scalability	Moderate, scalability depends on the performance of the working station	Moderate, scalability varies with the complexity of configurations	High, focuses on computational efficiency
Ease of Use	High, Minimal user interaction (dataset upload, attribute selection) and automated analysis	Moderate, Requires more user intervention and expertise for configuration and result interpretation	High, intuitive dashboard and API integration
Legal Compliance	Includes GDPR-ready risk thresholds, Supports extended metrics for a comprehensive risk analysis	Does not include risk thresholds for legal compliance, Focus only on risk calculation	Does not include risk calculation or risk thresholds for legal compliance
Extended Features (Strengths)	Broad attack coverage with novel risk metrics (uniformity, correlation and simplified Markov model risk), Focus on Open Data publishing, Risk categorisation (low, medium, high), Minimal user interaction	Support different anonymisation techniques to create anonymised datasets	Detailed risk visualisation, Easy integration with APIs
Limitations	Simplified Markov model risk has limited applicability and non-linear thresholds, No exact thresholds for extended metrics	Lacks support for extended metrics like uniformity and correlation risks, Does not include risk thresholds for legal compliance	Limited to uniqueness risk, No focus on legal compliance

An overview of these steps is provided in Figure 2. Upon launching, the first task for the user is to load a dataset in CSV format. The application then automatically identifies and extracts the attributes listed in the dataset's first row. Next, the user is required to classify the dataset, specifying whether it is event-level or person-level. For event-level datasets, the user must also identify the person ID attribute, which uniquely corresponds to individuals within the dataset. Once the attributes are identified, the interface displays them, enabling users to select the QI and SAs that will be the focus of the risk assessment process. SCORR is designed to assess each SA individually, producing specific results and risk analyses for each. If multiple SAs require evaluation, SCORR mandates separate runs for each SA. To assist in the selection of attributes, a dedicated function offers guidance on identifying QI. After completing the selection process, the user's involvement concludes, and the tool proceeds with the subsequent processes autonomously. The application then computes various risk metrics, including both conventional and extended metrics, which provide insights into the potential risk of re-identification within the dataset. By evaluating these metrics, the application offers a detailed analysis of the dataset's privacy risk in re-identification attacks.

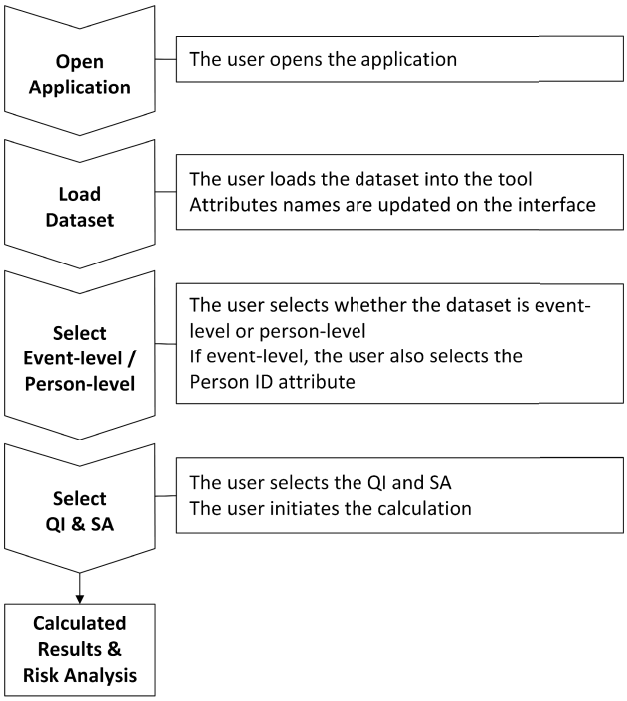


FIGURE 2. Overview of steps followed by the user in SCORR.

B. RISK ANALYSIS

Conventional metrics are employed to ensure compliance with legal data privacy requirements. As discussed in Section II-H, specific criteria for the maximum allowable re-identification risk or the required level of anonymisation before a dataset can be publicly released are not explicitly defined. However, it is suggested that k-anonymity should satisfy the condition $k > 10$, and t-closeness should meet the criterion $t \leq 0.5$ as discussed in section II-H. Based on these parameters, a workflow has been developed to assess compliance with conventional metrics, as shown in Figure 3.

The initial step in this workflow is to verify whether the dataset meets the k-anonymity requirement. If this criterion is not satisfied, the dataset is classified as "Not

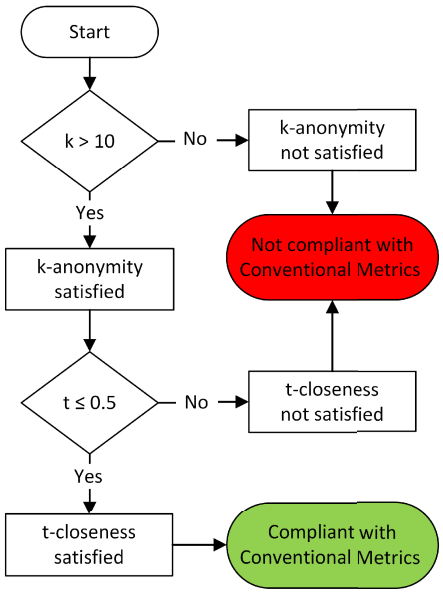


FIGURE 3. Compliance criteria for conventional metrics.

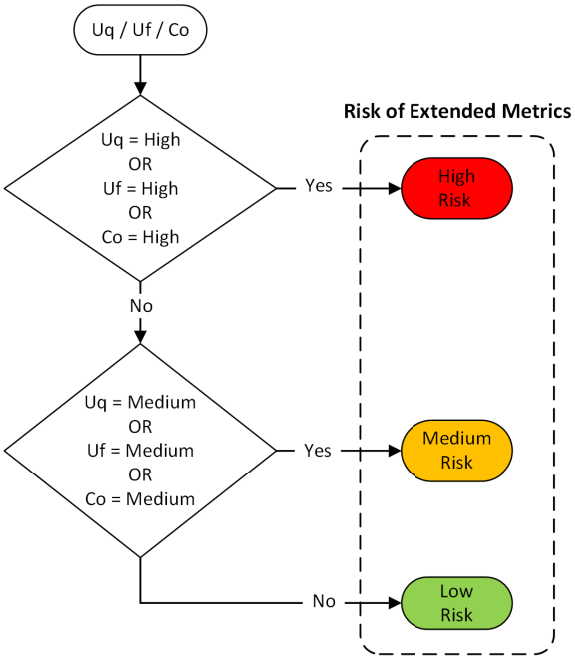


FIGURE 4. Compliance criteria for extended metrics.

Compliant with Conventional Metrics.'' Subsequently, the t-value is checked. Only if the dataset meets this criterion as well is it considered compliant with conventional metrics; otherwise, it is classified as non-compliant. While there is no fixed requirement for l-diversity in this workflow, the maximum achievable l-diversity for a dataset is defined as the minimum between the k-anonymity value and the number of distinct values of the SA. Although l-diversity is not used to determine compliance with conventional metrics in this study, we calculate its value to verify whether the maximum possible l-diversity has been achieved. The trade-off between data utility and privacy is less pronounced in t-closeness compared to l-diversity, which tends to reduce data utility significantly for minimal privacy protection.

Once compliance with conventional metrics has been assessed, extended metrics are evaluated to gain a deeper understanding of the dataset's uniqueness, uniformity, and correlation characteristics. These metrics are scored on a scale from 0 to 1, where a score closer to 0 indicates a lower likelihood of re-identification, and a score closer to 1 suggests a higher re-identification risk. The metrics are computed for all records in the dataset, but only the maximum risk scores are used for analysis, reflecting the worst-case scenario. The score range is divided into three categories: low risk [0 to 0.33], medium risk [0.34 to 0.66], and high risk [0.67 to 1]. All uniqueness, uniformity, and correlation metrics are classified according to these risk categories and combined in a workflow, as shown in Figure 4. If any extended metric falls into the high-risk category, the dataset is categorised as high risk under extended metrics. Similarly, if any metric falls into the medium-risk category, the dataset is classified as medium risk. A dataset is considered low risk only if all three metrics fall into the low-risk category.

It should be noted that the R_{mm} is not linearly related to the actual re-identification risk. Therefore, it cannot be categorised into low, medium, and high risk based on linear thresholds, unlike the uniqueness-only, uniformity-only, and R_{co}. Consequently, although the Markov Model risk is calculated, it is not used to assess dataset compliance. After categorising the risks associated with both conventional and extended metrics, a final risk overview for the dataset is established. If the dataset does not comply with conventional metrics or is deemed high risk based on the extended metrics, it is not approved for public release. If the dataset complies with conventional metrics but presents a medium risk in the extended metrics, it may be released publicly with an acknowledgement of the associated risk. A dataset is eligible for public release if it complies with conventional metrics and exhibits low risk based on extended metrics. This final decision-making process is illustrated in Figure 5.

All metrics can be applied to both categorical and numerical data in person-level and event-level datasets. However, with the exception of t-closeness, numerical data is treated as categorical data with discrete values by ignoring their continuous nature. Thus, numerical data anonymised through generalisation and suppression techniques can be assessed, while numerical data anonymised using perturbation techniques cannot.

V. EVALUATION

For the practical evaluation, we utilised publicly available tabular datasets, specifically the Open University Learning Analytics Dataset (OULAD) [57], Hospital Inpatient Discharges (HID) [58], and the Adult dataset [59]. The

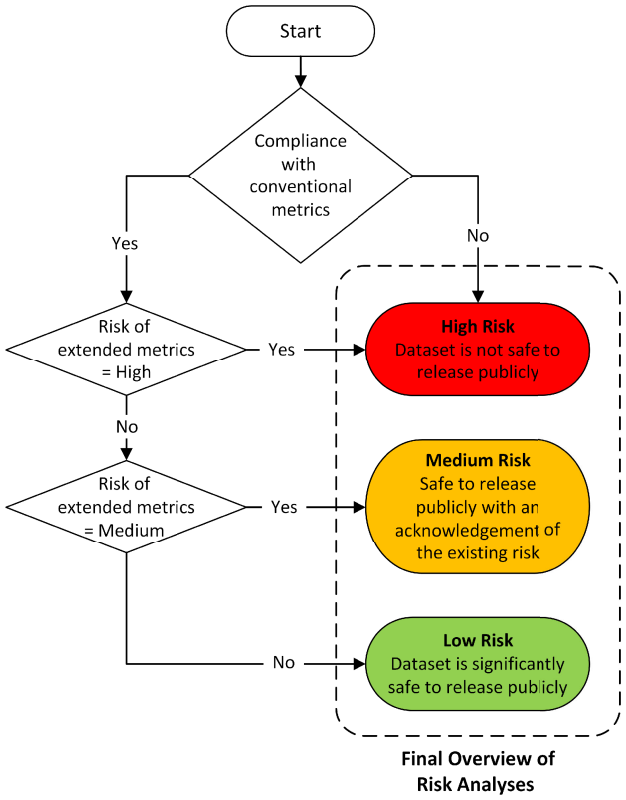


FIGURE 5. Compliance criteria for the dataset.

OULAD dataset, provided by The Open University, which has the largest number of Bachelor's students in the UK, includes data from seven modules of the university's online learning platform. The dataset encompasses student demographics, assessment scores, and interactions with the Virtual Learning Environment, amounting to 32,593 records across 12 attributes. The dataset is event-level, and the attribute "id_student" serves as the unique student identifier (Person ID), while attributes such as "gender", "region", "highest_education", "age_band", and "disability" are treated as QI. The attribute "final_result" is designated as the SA for the evaluation. The HID dataset is a person-level dataset, consisting of anonymised records of hospital discharges in New York State, detailing patient characteristics, diagnoses, treatments, services, and charges. This dataset, which includes 2,367,550 records across 34 attributes, has been previously employed in re-identification risk assessment studies [60] based on the size of equivalence classes. In our study, the attributes "Hospital Country", "Facility Name", "Age Group", "Zip Code – 3 digits", "Gender", "Race", and "Ethnicity" are identified as QI, with the "CCS Diagnosis Description" serving as the SA. The Adult dataset, a person-level dataset sourced from the UC Irvine Machine Learning Repository, was extracted from the 1994 U.S. Census Bureau database by Ronny Kohavi and Barry Becker. It contains 14 attributes, where "age", "working class", "education", "marital status", "occupation", "relationship", "race" and "sex" are selected as QI and "income" is selected as the SA. This dataset comprises 48,842 records and is commonly used to predict whether an individual's annual income exceeds $50,000. To simplify computations, we restricted our analysis to the first 5,000 records of each dataset.

A. PRIVACY METRICS

The datasets were anonymised using ARX [54], [55], ensuring $k > 10$ and $t \leq 0.5$. ARX allows dataset anonymisation based on predefined parameter settings for these metrics. Several anonymised versions were generated, each with different levels of generalisation for QI. However, three progressively generalised versions were selected for further analysis. SCORR was then applied to compute results separately for the original dataset and the three anonymised versions. The experimental setup used in this study is shown in Figure 6 and was applied to all three datasets: OULAD, HID, and the Adult dataset.

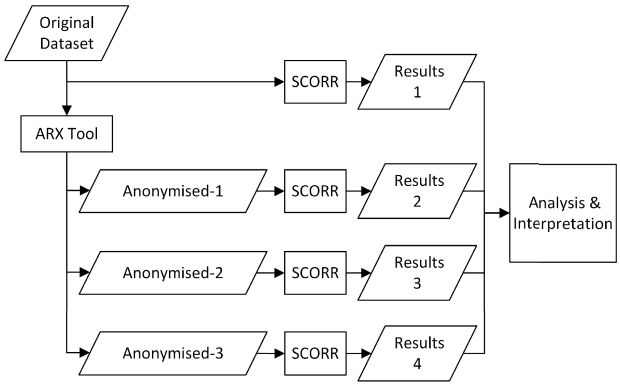


FIGURE 6. Test Setup for privacy metrics of original datasets.

The evaluation of privacy metrics was conducted using two distinct approaches. First, the metrics were assessed based on the privacy characteristics they measure, as detailed in the previous sections. Table 2 summarises these characteristics for each metric. Second, the metrics were tested to evaluate re-identification risk across various publicly available tabular datasets. The results of this analysis are presented and discussed in the following sections.

1) CONVENTIONAL METRICS

This section demonstrates the performance of SCORR in accurately estimating conventional privacy metrics and comparing these results with those obtained from ARX. The metrics evaluated include k-anonymity, l-diversity, and t-closeness. Specifically, k-anonymity is derived from the QI, while l-diversity and t-closeness are calculated based on both QI and SA. The calculation steps are provided in Section II-D. For the evaluations, three anonymised versions of the OULAD dataset were selected, with $k = 12$, $k = 20$, and $k = 25$. Similarly, for the HID dataset, anonymised

TABLE 2. Characteristics addressed by metrics.

Metric	Uniqueness	Uniformity	Correlation	Distribution of SA values
k-anonymity	✓	×	×	×
l-diversity	×	×	×	✓
t-closeness	×	×	×	✓
R_{uq}	✓	×	×	×
R_{uf}	×	✓	×	×
R_{co}	×	×	✓	×
R_{mm}	✓	✓	✓	×

versions with $k = 11$, $k = 14$, and $k = 19$ were chosen, and for the Adult dataset, versions with $k = 11$, $k = 14$, and $k = 20$ were selected. The original datasets exhibited a k-anonymity of 1. All anonymised datasets, except for the HID dataset, reached the target maximum t-closeness value of 0.5. In the HID dataset, however, the t-closeness target was set higher to avoid a further reduction in data utility. Figure 7 presents the summary of conventional privacy metrics across all datasets and anonymisation levels, including the original datasets. In the figure, each colour represents a specific level of anonymisation, with dark blue denoting the original dataset. The figure is organised into three rows: the top row shows results for the OULAD dataset, the middle row corresponds to the HID dataset, and the bottom row represents the Adult dataset.

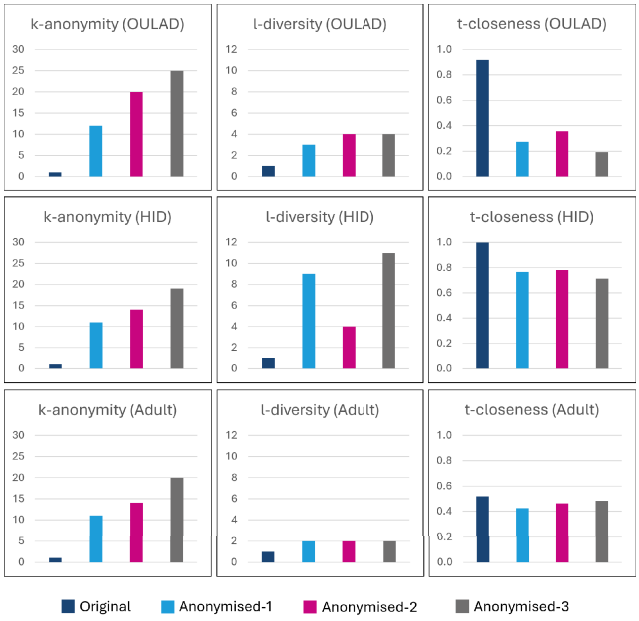


FIGURE 7. Conventional metrics plot for all datasets.

To start with, the privacy metrics calculated by ARX for the original and anonymised datasets match the values determined by SCORR for the same datasets.

The first row shows the results for the OULAD dataset. Three anonymised datasets with k-anonymity values of up to 25 were created. The l-diversity also improves, but only reaches a maximum value of 4 since the selected SA "final_result" has only 4 distinct values. In contrast, the t-closeness decreases significantly from the original to the anonymised dataset, achieving the desired values of less than or equal to 0.5. For the OULAD dataset, it can also be observed that t-closeness increases from Anonymised-1 to Anonymised-2 due to changes in the SA values within equivalent classes in the respective anonymised datasets.

In the second row, displaying the HID dataset, it can be seen that the k-anonymity has increased with each anonymisation level. At the same time, l-diversity improved from the worst value of 1 in the original dataset to 11. The t-closeness value dropped from 1 in the original dataset to 0.713, but could not reach the value below 0.5 as described above. This also shows that higher anonymisation does not necessarily provide a better solution, as can be observed in the transition from Anonymised-1 to Anonymised-2. Here, the l-diversity decreased, and at the same time, the t-closeness value increased. This is because the formation of the equivalence classes and their SA value distribution can be different in anonymised datasets, which will lead to different l-diversity and t-closeness values. In addition, the SA "CCS Diagnosis Description" contains 178 distinct values, meaning that the maximum achievable l-diversity is limited by the k value. However, this maximum l-diversity could not be reached in practice. Anonymising the datasets to achieve the greatest possible l-diversity and a t-closeness of 0.5 or less would lead to a considerable reduction in data utility and was therefore not considered.

In the third row, the results for the Adult dataset can be seen. Closer k values to the HID dataset were used in order to find a suitable balance between data quality and privacy. The l diversity improves as a result of anonymisation, but can only reach a maximum value of 2, as the specified SA "income" only contains the two distinct values of "$\leq 50k$" and "$> 50k$". Similar to the OULAD dataset, the Adult dataset also achieved $t \leq 0.5$. However, these values were already very low before anonymisation, with an initial value of 0.519. t-closeness shows a slight increase from Anonymised-1 to Anonymised-3 while satisfying $t \leq 0.5$ criteria. This happened since ARX does not try to further reduce t-closeness, as it has already been achieved. Furthermore, the formation of the equivalence classes and their sensitive attribute value distribution can be different in anonymised datasets, which will lead to different t-closeness values.

The exact results for all datasets have been summarised below in Table 3

2) EXTENDED METRICS

Extended metrics were calculated for each record in the dataset using the equations presented in section II-E, with only the maximum values considered as global values for evaluation, focusing on the worst-case risk scenario. First,

TABLE 3. Calculated conventional metrics results for all datasets.

Metric	Original	Anonymised-1	Anonymised-2	Anonymised-3
OULAD				
k-anonymity	1	12	20	25
l-diversity	1	3	4	4
t-closeness	0.918	0.273	0.355	0.192
HID				
k-anonymity	1	11	14	19
l-diversity	1	9	4	11
t-closeness	1	0.764	0.781	0.713
ADULT				
k-anonymity	1	11	14	20
l-diversity	1	2	2	2
t-closeness	0.519	0.426	0.462	0.483

R_{uq} and R_{uf} are calculated for the entire QI, followed by R_{uf} for each individual QI attribute. R_{co} is assessed for every combination of a QI attribute with the selected SA, while R_{mm} is computed based on the entire QI in relation to the SA. The datasets are the same as those used in the evaluation before and correspond to the same anonymisation levels.

There is a separate heatmap for each dataset because the datasets have different attributes, which are also individually assessed by R_{uf} and R_{co}. The evaluation of the extended metrics for OULAD can be seen in Figure 8. Each row of the heatmap represents an extended metric for a specific attribute or attribute combination calculated for the Original, Anonymised-1, Anonymised-2 and Anonymised-3 datasets mentioned in the previous section. The last row of the heatmap represents the R_{mm} values.

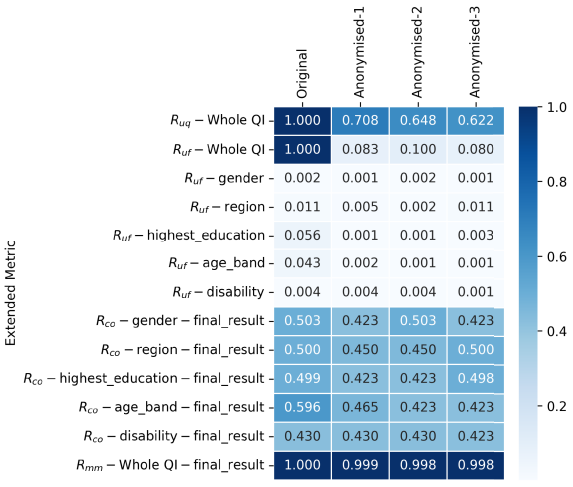


FIGURE 8. Extended metrics heatmap for OULAD Student Info dataset.

It is evident that most of the extended metrics show a decrease in associated risk from the Original to the Anonymised-3 in general, despite some occasional increases. R_{uq} – *Whole QI* decreases from 1 to 0.622 due to lower uniqueness caused by anonymisation, which is also represented by k-anonymity. A significant decrease in R_{uf} – *Whole QI* is observed in the anonymised datasets compared to the original, resulting in an extremely low risk (all below 0.1), with minor increases being negligible. Single attribute R_{uf} is negligible even in the original dataset, where further reduction is not necessary. R_{co} have been slightly reduced in most of the anonymised datasets. However, even after anonymising the dataset, anonymised QI values still show a moderate correlation with the SA values. To further reduce the R_{co}, extreme levels of anonymisation may be required, but this will lead to a highly compromised data utility. R_{mm} exhibits the least reduction in risk. While conventional metrics indicate higher risk reductions and other extended metrics show substantial reductions, the R_{mm} does not significantly reflect this decrease, therefore the risk cannot be categorised into low, medium or high using linear thresholds based on R_{mm}.

In Figure 9, it can be seen that most of the extended metrics in the HID dataset show a decrease in associated risk from the Original to Anonymised-3 in general, despite some occasional increases. R_{uq} – *Whole QI* in Figure 9 decreases from 1 to 0.654 due to lower uniqueness caused by anonymisation, which is also represented by k-anonymity. R_{uf} – *Whole QI* experiences a significant decrease from the Original and ends up with extremely low risk (all below 0.1) in the anonymised datasets, with minor increases being negligible. Individual R_{uf} values also show almost zero uniformity in anonymised datasets. R_{co} of the SA "CCS Diagnosis Description" with Hospital Country, Facility Name, Age Group, Zip Code – 3 digits and Ethnicity have drastically decreased in the anonymised datasets. However, like before R_{co} with "Gender" and "Race" show only a slight reduction because even after the anonymisation, the QI values still show a moderate correlation with the SA values. Achieving a more substantial reduction in these risks would also require a higher level of anonymisation, which would further compromise data utility. Once again, the R_{mm} was not able to represent the risk reduction significantly.

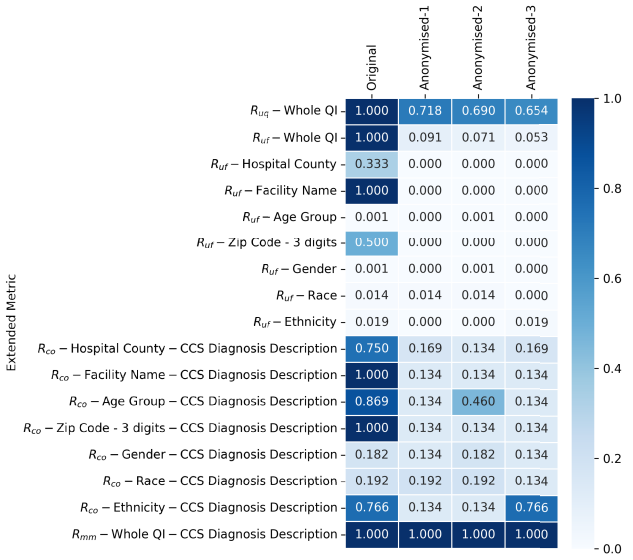


FIGURE 9. Extended metrics heatmap for HID dataset.

The results for the evaluation of the extended metrics in the Adult dataset can be seen in Figure 10. As before, anonymisation leads to significantly reduced values for most extended metrics. With each level of anonymisation, the R_{uq} – *Whole QI* decreases continuously from 1 to 0.648. R_{uf} values of *Whole QI*, ''age'', ''workclass'', ''education'', ''martial.status'' and ''occupation'' are drastically reduced in Anonymised-1 and further decreased in Anonymised-2 & 3. Other individual R_{uf} values were below 0.04 even in the original dataset, and decreased further after the anonymisation. R_{co} however, remains relatively high even after anonymisation. These correlations are important for data analysis and could only be reduced further by increasing anonymisation, which would lead to a further reduction in utility. Similar to the two previous datasets, R_{mm} is hardly impacted by the anonymisation. With each anonymisation level, the R_{mm} value only reduces by 0.01, which is almost negligible.

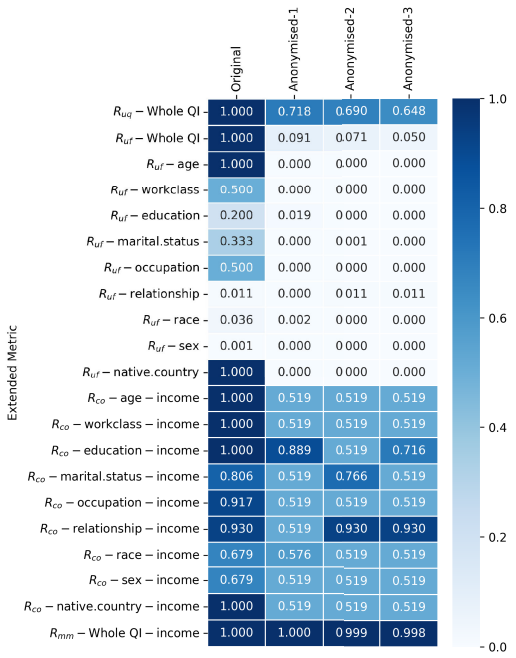


FIGURE 10. Extended metrics heatmap for Adult dataset.

These evaluations imply that a higher level of anonymisation does not always lead to a lower re-identification risk, and thus does not necessarily result in better privacy protection. To generate an anonymised dataset that adheres to both conventional and extended metrics while not compromising its data utility, specific attention must be paid to the level of anonymisation of each element of the QI. It is not enough to check the score of the whole dataset and anonymise more and more until the desired scores are obtained. SCORR is highly applicable in such scenarios since it provides attribute-oriented risk assessment to identify where higher anonymisation is needed and where it is not.

B. SCALABILITY

The scalability of calculating the privacy metric is a crucial factor, as it determines its applicability to larger datasets. In real-world scenarios, the number of attributes and records in a dataset can be enormous, resulting in numerous calculations.

To evaluate scalability, the execution time of each metric was measured by executing SCORR locally on a computer. From a hardware perspective, SCORR was designed with the idea that users can assess their anonymised records locally on their system for re-identification risk. For this reason, hardware specifications were chosen to match the performance of an average working station. More precisely, a 64-bit Windows 11 machine with an Intel i5-1135G7 with 2.40GHz and 8.00 GB of RAM was used.

In the field of time measurement, the programming language and libraries used also play a decisive role. Therefore, the programming language, Python (v. 3.11) was chosen due to its ease of use and widespread use in the field of scientific applications. Pandas (v. 1.5.2), a foundational open-source Python library which plays a crucial role in data science applications, was used for the requirements, particularly for data manipulation and analysis. The tabular data was provided in CSV format, and Pandas compatibility with CSV files ensures seamless integration into data pipelines, streamlining the process of data importing and exporting for ease of use. Finally, the calculations were performed using NumPy (v. 1.24.2), a fundamental open-source Python library known for its optimised array processing capabilities, which enable faster computations and reduced memory consumption.

As shown in Figure 11, each phase of the execution, from booting the application to displaying the results, varies in duration. For the evaluation of scalability, only the time required to calculate each metric, depicted in grey in the figure, was considered, while the time taken for other phases was excluded. When the same metric is calculated for different attributes as R_{uf} risk and R_{co}, the average execution time is considered.

Two different tests were conducted for two datasets. In the first test, the scalability of SCORR against the number of records was evaluated. The test was conducted for different numbers of records while selecting the same QI and SA. The test is repeated 100 times, and average execution times are considered for the evaluation. In the second test, the scalability of SCORR against the number of elements of the QI was evaluated. The test was conducted for the same dataset, starting with selecting one attribute in QI and continuing by selecting multiple attributes of QI, increasing one at a time. The tests are repeated 100 times, and average execution times are considered for the evaluation.

In Figure 12, the scalability of SCORR against the number of records can be seen. The test was conducted on the OULAD dataset on its first 1000, 2000, 3000, 5000, 7000 and 10000 records. During the test, attributes

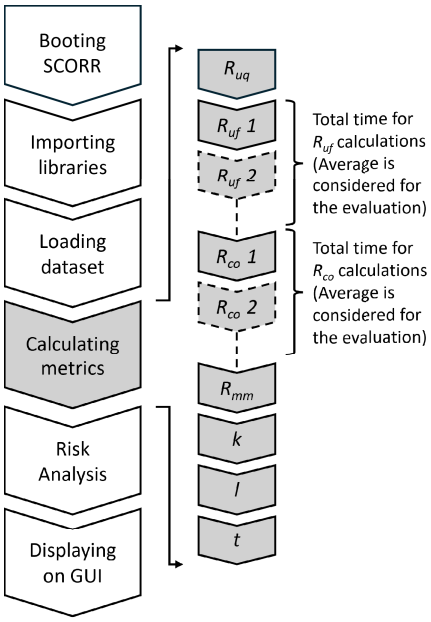


FIGURE 11. Execution phases of SCORR.

of "gender", "region", "highest_education", "age_band" and "disability" were selected as QI and "final_result" as the SA.

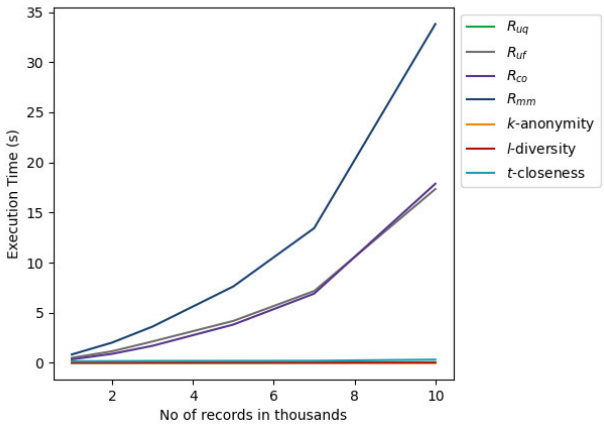


FIGURE 12. Execution time vs no of records - OULAD.

It can be observed that conventional metrics and R_{uq} do not show a significant increase in execution time when the number of records grows, and hence those metrics are highly scalable across large datasets. The reason for this higher scalability can be identified as the simplicity of the metric calculation and implementation with optimised vectorised operations instead of scalar operations in the Python Pandas library. Vectorised operations operate on entire arrays at once, while scalar operations operate on individual elements of an array, taking more execution time. R_{uf}, R_{co} and R_{mm} exponentially increase the execution time against the number of records. This is due to the complexity of the calculation and the impossibility of completely implementing only vectorised operations. R_{mm} limits the scalability of SCORR due to its complexity of calculation. When analysing its scalability versus quality, it is apparent that the R_{mm} lacks adequate performance. Figure 13 shows the scalability of SCORR against the number of elements of the QI. The test was conducted on the HID dataset using its first 5000 records. The test was started with selecting one QI, Hospital County and progressively added the remaining six QI of Facility Name, Age Group, Zip Code (3 digits), Gender, Race and Ethnicity one at a time.

During testing, it was observed that all metrics, except t-closeness, consistently exhibited similar execution times with only minor variations. For the extended metrics, execution time remained constant regardless of the number of elements of the QI because the calculations were based on a combined QI generated in a previous step, which took around 10 ms for any number of attributes. Although t-closeness exhibited a slight increase in execution time, SCORR remained substantially scalable with the number of QI.

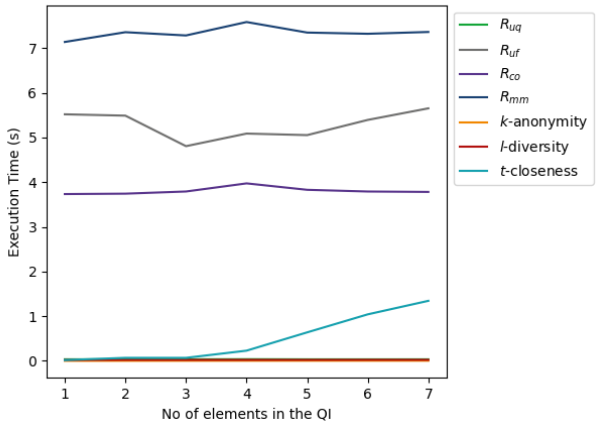


FIGURE 13. Execution time vs no of elements of the QI - HID.

C. USABILITY

To further substantiate the claim of minimal user interaction, we conducted a usability evaluation comparing SCORR with the freely available ARX Tool, a widely used data anonymisation solution. The primary objective was to quantify the number of interactions required from initiation to obtaining risk analysis results for a given dataset. An interaction was defined as a selection made via a click. ARX was chosen for comparison due to its accessibility and established use in the privacy domain. Since ARX offers over 15 selectable options on its main interface, the evaluation focused solely on interactions necessary for risk analysis, utilising shortcuts where possible. Table 4 summarises the number of interactions required from programme launch to the display of risk results. The process is divided into four categories: Setup (pre-import actions), Import (dataset loading), Configuration (attribute classification as DIs, QI, Sensitive, or Non-Sensitive), and

Risk View (displaying computed risk results). This evaluation was applied to all datasets used in this study, comparing SCORR and ARX. The table provides the number of interactions required per dataset for each tool, along with the average interaction count, offering a comparative overview. The results indicate that most interactions remain consistent across datasets, except for configuration, which depends on the number of attributes. In ARX, attributes must be selected individually, with each requiring three interactions after the first. SCORR, however, presents all attributes on a single page, allowing efficient classification via checkboxes, significantly reducing interaction effort. Both tools default unselected attributes to "Non-Sensitive." For Risk View, both tools require a single step to display results. However, ARX divides risk analysis across four tabs, potentially requiring up to three additional interactions based on the metrics of interest. In contrast, SCORR consolidates all risk information on a single scrollable page, eliminating extra navigation. SCORR requires half the interactions for Setup and Import compared to ARX, with the most significant reduction observed in Configuration, where interactions are reduced by 63.63%. Overall, across the three datasets, SCORR reduces total required interactions by 59.38%, demonstrating a significantly more streamlined and efficient process, driven by improvements in both calculation methods and UI/UX.

TABLE 4. Comparison of required interactions for risk analysis: ARX vs. SCORR.

Tool	Data	Setup	Import	Config.	View	Total
ARX	OULAD	2	8	17	1	27
	HID	2	8	23	1	33
	Adult	2	8	26	1	36
	Avg.	2	8	22	1	32
SCORR	OULAD	1	4	6	1	11
	HID	1	4	8	1	13
	Adult	1	4	9	1	14
	Avg.	1	4	8	1	13

VI. CONCLUSION

This paper addressed the challenge of quantifying privacy risks in the re-identification of tabular datasets. We introduced SCORR, a comprehensive scoring system that assesses privacy risks using multiple metrics tailored to different attack types. By integrating conventional metrics (k-anonymity, l-diversity, and t-closeness) with extended metrics (R_{uq}, R_{uf}, R_{co}, and R_{mm}), SCORR provides a holistic risk analysis and supports informed decision-making regarding dataset release. A thorough literature review highlighted the limitations of relying on single-metric approaches, reinforcing the need for a multi-metric risk assessment. SCORR incorporates a risk analysis model where conventional metrics follow threshold-based recommendations, while extended metrics are categorised into low, medium, and high risk. Users receive a visual representation of risks via a risk scale, simplifying interpretation. SCORR was evaluated on three datasets (OULAD, HID, and ADULT), each with progressively anonymised versions. The results confirmed SCORR's accuracy in computing conventional metrics, aligning with ARX outputs. However, greater anonymisation did not always correlate with lower re-identification risk. R_{co} remained high in certain cases, while R_{mm} showed minimal reduction across all datasets, leading to its exclusion from the risk model. Scalability analysis demonstrated that conventional metrics and R_{uq} exhibited minimal computational overhead, while more complex metrics (R_{uf}, R_{co}) required significantly more processing time. The number of elements of the QI did not impact most metrics, except for t-closeness, which scaled linearly beyond a threshold. A usability evaluation showed that SCORR reduces the number of required interactions by nearly 60% compared to ARX, making risk assessment more efficient and accessible. SCORR offers an attribute-oriented privacy assessment that enables data custodians to balance anonymisation and utility effectively. Its modular design allows for the integration of new metrics and refinements, ensuring adaptability to evolving privacy challenges.

VII. FUTURE WORK

Despite its advantages, SCORR has limitations that should be addressed to enhance re-identification risk assessment. Currently, there are no established thresholds for extended risk metrics. While conventional metrics follow literature-based recommendations, extended metrics rely on arbitrary categorisations ($[0 - 0.33]$, $[0.34 - 0.66]$, $[0.67 - 1.00]$). A more informed approach would derive thresholds from analysing published datasets, comparing those successfully re-identified with those that have not. Another limitation is SCORR's assessment of single datasets, whereas real-world re-identification attacks often involve cross-dataset linkages. Future work should incorporate external data sources, as linking anonymised datasets with public or anonymised records significantly increases risk. Additionally, SCORR currently determines attribute-level risk based on the most vulnerable record, but lacks transparency in how individual records contribute. Providing record-level insights would allow targeted risk mitigation while maintaining data quality. Key priorities for future development include defining risk thresholds, integrating cross-dataset analysis, and improving transparency through detailed statistical evaluations. These enhancements will further refine SCORR's effectiveness in real-world anonymisation scenarios.

REFERENCES

[1] P. Huston, V. Edge, and E. Bernier, "Reaping the benefits of open data in public health," *Canada Communicable Disease Rep.*, vol. 45, no. 10, pp. 252–256, Oct. 2019.

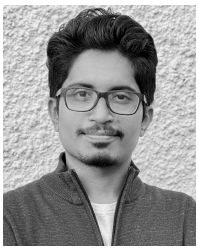
[2] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits, adoption barriers and myths of open data and open government," *Inf. Syst. Manage.*, vol. 29, no. 4, pp. 258–268, Sep. 2012.

[3] F. Kitsios, N. Papachristos, and M. Kamariotou, "Business models for open data ecosystem: Challenges and motivations for entrepreneurship and innovation," in *Proc. IEEE 19th Conf. Bus. Informat. (CBI)*, vol. 1, Jul. 2017, pp. 398–407.

[4] *Open Data and Privacy*. [Online]. Available: https://citizens-guide-open-data.github.io/guide/4-od-and-privacy
[5] T. Scassa, "Privacy and open government," *Future Internet*, vol. 6, no. 2, pp. 397–413, Jun. 2014.
[6] A. Rohunen, J. Markkula, M. Heikkilä, and J. Heikkilä, "Open traffic data for future service innovation-addressing the privacy challenges of driving data," *J. Theor. Appl. Electron. Commerce Res.*, vol. 9, no. 3, pp. 71–89, Sep. 2014.
[7] R. Meijer, P. Conradie, and S. Choenni, "Reconciling contradictions of open data regarding transparency, privacy, security and trust," *J. Theor. Appl. Electron. Commerce Res.*, vol. 9, no. 3, pp. 32–44, Sep. 2014. [Online]. Available: https://www.mdpi.com/0718-1876/9/3/18
[8] M. Beno, K. Figl, J. Umbrich, and A. Polleres, "Open data hopes and fears: Determining the barriers of open data," in *Proc. Conf. E-Democracy Open Government (CeDEM)*, May 2017, pp. 69–81. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8046274
[9] E. Graham-Harrison and C. Cadwalladr, "Revealed: 50 million Facebook profiles harvested for Cambridge analytica in major data breach," *Guardian*, vol. 17, 2018.
[10] H. Hodson. *Revealed: Google AI Has Access to Huge Haul of NHS Patient Data*. [Online]. Available: https://www.newscientist.com/article/2086454-revealed-google-ai-has-access-to-huge-haul-of-nhs-patient-data/
[11] F. K. Dankar, K. El Emam, A. Neisa, and T. Roffey, "Estimating the re-identification risk of clinical data sets," *BMC Med. Informat. Decis. Making*, vol. 12, no. 1, pp. 1–15, Dec. 2012.
[12] L. O. Gostin, L. A. Levit, and S. J. Nass, "Beyond the hipaa privacy rule: Enhancing privacy, improving health through research," 2009.
[13] *Anonymisation: Managing Data Protection Risk Code of Practice*, ICO, 2012.
[14] T. G. Moraes, A. N. L. E. Lemos, A. K. Lopes, C. Moura, and J. R. L. de Pereira, "Open data on the COVID-19 pandemic: Anonymisation as a technical solution for transparency, privacy, and data protection," *Int. Data Privacy Law*, vol. 11, no. 1, pp. 32–47, May 2021.
[15] E. Mackey, "A best practice approach to anonymization," in *Handbook of Research Ethics and Scientific Integrity*, R. Iphofen, Ed., Cham, Switzerland: Springer, 2020, pp. 323–343, doi: 10.1007/978-3-030-16759-2_14.
[16] Atockar. (2014). *Riding With the Stars: Passenger Privacy in the NYC Taxicab Dataset*. [Online]. Available: https://agkn.wordpress.com/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/
[17] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2008, pp. 111–125. [Online]. Available: http://ieeexplore.ieee.org/document/4531148/
[18] M. Arrington. (2006). *AOL Proudly Releases Massive Amounts of Private Data*. [Online]. Available: https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/
[19] B. Lubarsky, "Re-identification of 'anonymized data," *Georgetown Law Technol. Rev.*, 2010. Accessed: Sep. 10, 2021. [Online]. Available: https://wwwgeorgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017
[20] A. Majeed and S. Lee, "Anonymization techniques for privacy preserving data publishing: A comprehensive survey," *IEEE Access*, vol. 9, pp. 8512–8545, 2021.
[21] OCR. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance With the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. Accessed: Aug. 22, 2024. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html
[22] S. Kiyomoto, T. Nakamura, and Y. Miyake, "Towards tracing of K-anonymized datasets," in *Proc. IEEE Trustcom/BigDataSE/ISPA*, vol. 1, Aug. 2015, pp. 1237–1242.
[23] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, p. 3, Jan. 2006.
[24] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond K-anonymity and L-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.
[25] D. Vatsalan, T. Rakotoarivelo, R. Bhaskar, P. Tyler, and D. Ladjal, "Privacy risk quantification in education data using Markov model," *Brit. J. Educ. Technol.*, vol. 53, no. 4, pp. 804–821, Jul. 2022.
[26] *Metrics and Frameworks for Privacy Risk Assessments—CSIRO*. [Online]. Available: https://www.csiro.au/en/research/technology-space/cyber/metrics-and-frameworks-for-privacy-risk-assessments
[27] N. O. Attoh-Okine, "Differential privacy," in *Big Data and Differential Privacy: Analysis Strategies for Railway Track Engineering*. Cham, Switzerland: Springer, 2017, pp. 241–247.
[28] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. 3rd Theory Cryptogr. Conf. Theory Cryptogr. (TCC)*, New York, NY, USA. Cham, Switzerland: Springer, Jan. 2006, pp. 265–284.
[29] A. Gadotti, L. Rocher, F. Houssiau, A.-M. Creţu, and Y.-A. de Montjoye, "Anonymization: The imperfect science of using data while preserving privacy," *Sci. Adv.*, vol. 10, no. 29, p. 7053, Jul. 2024.
[30] *Tumult Labs The Fundamental Trilemma of Synthetic Data Generation*. [Online]. Available: https://www.tmlt.io/resources/fundamental-trilemma-synthetic-data-generation
[31] L. Yao, Z. Chen, H. Hu, G. Wu, and B. Wu, "Privacy preservation for trajectory publication based on differential privacy," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 3, pp. 1–21, Jun. 2022.
[32] A. Majeed and S. O. Hwang, "Differential privacy and K-anonymity-based privacy preserving data publishing scheme with minimal loss of statistical information," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 3, pp. 3753–3765, Jun. 2024.
[33] C. McKay Bowen and J. Snoke, "Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge," 2019, *arXiv:1911.12704*.
[34] *The Global Synthetic Dataset*. [Online]. Available: https://www.ctdatacollaborative.org/page/global-synthetic-dataset
[35] S. Hod and R. Canetti, "Differentially private release of Israel's national registry of live births," 2024, *arXiv:2405.00267*.
[36] A. Kiran, P. Rubini, and S. S. Kumar, "Comprehensive review of privacy, utility, and fairness offered by synthetic data," *IEEE Access*, vol. 13, pp. 15795–15811, 2025.
[37] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnádi, "A unified framework for quantifying privacy risk in synthetic data," 2022, *arXiv:2211.10459*.
[38] P. A. Osorio-Marulanda, G. Epelde, M. Hernandez, I. Isasa, N. M. Reyes, and A. B. Iraola, "Privacy mechanisms and evaluation metrics for synthetic data generation: A systematic review," *IEEE Access*, vol. 12, pp. 88048–88074, 2024.
[39] *Measuring Utility and Information Loss—SDC Practice Guide Documentation*. [Online]. Available: https://sdcpractice.readthedocs.io/en/latest/utility.html
[40] S. Kocar et al., "A universal global measure of univariate and bivariate data utility for anonymised microdata," Tech. Rep., 2018.
[41] W. Lixia and H. Jianmin, "Utility evaluation of K-anonymous data by microaggregation," in *Proc. ISECS Int. Colloq. Comput., Commun., Control, Manage.*, Aug. 2009, pp. 381–384.
[42] N. Rajeshwari and C. Sowmyarani, "Data utility measures—A survey," in *Proc. 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, Jul. 2016, pp. 722–725.
[43] *GDPR Archives-gdpr.eu*. [Online]. Available: https://gdpr.eu/tag/gdpr/?cn-reloaded=1
[44] *Office of Privacy and Civil Liberties | Privacy Act of 1974*. [Online]. Available: https://www.justice.gov/opcl/privacy-act-1974
[45] *L13709*. [Online]. Available: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm
[46] *Privacy By Design—General Data Protection Regulation (GDPR)*. [Online]. Available: https://gdpr-info.eu/issues/privacy-by-design/
[47] (2014). *Article 29 Data Protection Working Party*. [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
[48] C. E. M. Jakob, F. Kohlmayer, T. Meurers, J. J. Vehreschild, and F. Prasser, "Design and evaluation of a data anonymization pipeline to promote open science on COVID-19," *Scientific Data*, vol. 7, no. 1, p. 435, Dec. 2020.
[49] (2018). *External Guidance on the Implementation of the European Medicines Agency Policy on the Publication of Clinical Data for Medicinal Products for Human Use*. [Online]. Available: https://www.ema.europa.eu/contact
[50] M. Bezzi, "An entropy based method for measuring anonymity," in *Proc. 3rd Int. Conf. Secur. Privacy Commun. Netw. Workshops (SecureComm)*, 2007, pp. 28–32.

[51] H. Kikuchi, T. Yamaguchi, K. Hamada, Y. Yamaoka, H. Oguri, and J. Sakuma, "Ice and fire: Quantifying the risk of re-identification and utility in data anonymization," in *Proc. IEEE 30th Int. Conf. Adv. Inf. Netw. Appl. (AINA)*, Mar. 2016, pp. 1035–1042.
[52] Y. Jiang, L. Mosquera, B. Jiang, L. Kong, and K. E. Emam, "Measuring re-identification risk using a synthetic estimator to enable data sharing," *PLoS ONE*, vol. 17, no. 6, Jun. 2022, Art. no. e0269097.
[53] J. P. Reiter, "Estimating risks of identification disclosure in microdata," *J. Amer. Stat. Assoc.*, vol. 100, no. 472, pp. 1103–1112, Dec. 2005, doi: 10.1198/016214505000000619.
[54] *ARX-data Anonymization Tool | A Comprehensive Software for Privacy-preserving Microdata Publishing*. [Online]. Available: https://arx.deidentifier.org/
[55] F. Praßer, F. Kohlmayer, R. Lautenschläger, and K. A. Kuhn, "ARX-a comprehensive tool for anonymizing biomedical data," in *Proc. AMIA Annu. Symp.*, Jan. 2014, pp. 984–93.
[56] *Measuring Re-identification and Disclosure Risk Sensitive Data Protection Documentation Google Cloud*. [Online]. Available: https://cloud.google.com/sensitive-data-protection/docs/compute-risk-analysis
[57] *Open University Learning Analytics Dataset (OULAD) | Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/anlgrbz/student-demographics-online-education-dataoulad
[58] (2014). *Hospital Inpatient Discharges (SPARCS De-identified): 2014 | State of New York*. [Online]. Available: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/rmwa-zns4
[59] R. K. B. Becker. *Adult*. [Online]. Available: https://archive.ics.uci.edu/dataset/2
[60] Y. J. Lee and K. H. Lee, "What are the optimum quasi-identifiers to re-identify medical records?" in *Proc. 20th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2018, pp. 1025–1033.

JAKOB FOLZ received the B.Eng. and M.Sc. degrees in applied computer science from the Deggendorf Institute of Technology (DIT), in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree with the University of West Bohemia. He is currently a Research Assistant with DIT's Technology Campus Vilshofen (TCV), specializing in artificial intelligence, data protection, and cybersecurity. His research interests include encrypted data manipulation and privacy preservation. This includes exploring anonymization methods and modern cryptographic techniques.



MANJITHA D. VIDANALAGE received the B.Sc. and master's Diploma degrees in electrical engineering from the University of Moratuwa, Sri Lanka, in 2016 and 2018, respectively, and the M.Sc. degree in electrical and information technology from the Deggendorf Institute of Technology (DIT), Germany, in 2023. He is currently a Research Assistant with the DIT's Technology Campus Vilshofen (TCV), specializing in personal privacy and open data.

ROBERT AUFSCHLÄGER received the M.Sc. degree in computer science and in computational mathematics from the University of Passau, in 2021. He was a Research Assistant with the University of Passau. In 2022, he moved to the Technology Campus Vilshofen (TCV), Deggendorf Institute of Technology (DIT). He is currently a Research Assistant and a Doctoral Student. His research interests include privacy-preserving machine learning, open data, and synthetic data.

AMAR ALMAINI received the B.Sc. and M.Sc. degrees in computer science from the Technical University of Munich, in 2011 and 2014, respectively, and the Ph.D. degree from the School of Computing, Engineering and the Built Environment, Edinburgh Napier University, U.K., in 2024. Previously, he was with Focus GmbH and Societe Generale Securities Services as a Software Developer. He is currently a Postdoctoral Researcher with the Deggendorf Institute of Technology. His main research interests include software-defined networking, resource management techniques, and machine learning.

MICHAEL HEIGL received the Ph.D. degree in computer science and engineering from the University of West Bohemia. He is currently a Professor with the Faculty of Applied Computer Science, Deggendorf Institute of Technology (DIT). He is one of the scientific directors of the Technology Transfer Centre for Digital Security, Vilshofen Technology Campus (TCV), DIT, and one of the leaders of the inter-faculty institute ProtectIT conducting application-oriented research on the protection of industrial automation technology, critical infrastructures, automotive and avionics systems, and IoT-devices. His research interests include network communication detection and prevention mechanisms, especially applying artificial intelligence to improve cybersecurity. His findings have been published in international open access journals and IEEE conference proceedings.

DALIBOR FIALA received the Graduate degree from the University of West Bohemia, Pilsen, Czech Republic, in 2003, specialized in software engineering, and the joint Ph.D. degree from UWB and Louis Pasteur University, Strasbourg, France. He spent a semester with the Queen's University of Belfast, U.K., in 2001. He defended his doctoral thesis under joint supervision, in 2007. From 2007 to 2009, he was a Software Engineer with Gefasoft AG, Munich, Germany. Currently, he is an Associate Professor with the Department of Computer Science and Engineering, UWB. His research interests include data mining, web mining, information retrieval, informetrics, and information science.

MARTIN SCHRAMM received the Ph.D. degree from the Faculty of Science & Engineering, University of Limerick (UL), Ireland, in 2016. He is currently a Professor with the Faculty of Computer Science, Deggendorf Institute of Technology (DIT). He also works as one of the Director of the Institute for the Protection of Industrial Technologies (ProtectIT) and the Technology Transfer Centre for Digital Security, Vilshofen Technology Campus (TCV). He leads application-oriented research in securing industrial automation technology, critical infrastructures, automotive and avionic systems, and IoT-devices funded by German Federal Ministry of Education and Research and Bavarian Ministry of Science and Art. His findings have been published in international open access journals and IEEE conference proceedings.