Evaluating the Risk of Re-Identification in Data Release Strategies: An Attacker-Centric Approach

Patrick Mesana
HEC Montréal, Canada
patrick.mesana@hec.ca

Pascal Jutras
Purdue University, United States
pjutrasd@purdue.edu

Julien Crowe
National Bank of Canada, Canada
julien.crowe@bnc.ca

Gregory Vial
HEC Montréal, Canada
greogry.vial@hec.ca

Gilles Caporossi
HEC Montréal, Canada
gilles.caporossi@hec.ca

Abstract

In this methodological paper, we introduce a novel approach to evaluate the risk of re-identification of individuals associated with data release strategies, including data redaction, data anonymization and data synthesis. More precisely, our approach simulates an attacker performing singling-out attacks as outlined in data protection regulations, and scores attacks based on the linkability of records and the information gain obtained by the attacker. Additionally, we further enhance our approach by simulating attacks as a cooperative game. In this game, the value of the attackers' information resources is determined using Shapley value borrowed from game theory. We also demonstrate the effectiveness of our approach using the Adult Income Census (AIC) dataset before discussing the economic implications associated with a privacy breach. Our work contributes to research and practice on the pressing need to better understand and evaluate the inherent trade-offs that exist between data privacy and utility in organizations.

Keywords: Privacy, Anonymization, Data Synthesis, Adversarial Agents, Risk of Re-identification.

1. Introduction

According to the General Data Protection Regulation (GDPR) and other modern data protection laws, personal information has a broad definition that includes any data that can identify an individual, either directly or indirectly, through the linking of multiple pieces of information. Entities serving as data custodians carry the obligation to protect their customers' privacy in the eventuality that data is released (*e.g.*, when it is shared with partners or with the public via an open data initiative, or through unlawful access to personal data - internal or external). One strategy that is often used to minimize privacy risks owing to its relative simplicity is data redaction, which consists in removing specific fields or rows. An alternative approach is data anonymization, which involves transforming personal data with the goal of altering sensitive information while preserving its utility. For example, generalization-based transformations can be employed to reduce the uniqueness of individual records, thus reducing the possibility of re-identification. Among the most well-known models for achieving this is k-anonymity (Sweeney, 2002). Finally, another strategy is to synthesize data using generative models trained on real data. This approach has been gaining popularity in recent years (Gootjes-Dreesbach et al., 2020; Park et al., 2018; Wan et al., 2017; Xu et al., 2019).

Organizations are generally interested in anonymous data because most regulations no longer consider them as personal or sensitive information. However, in practice, it is widely acknowledged that all three strategies inevitably carry a risk of re-identification embodying the idea that there is an inherent trade-off between data privacy and data utility for organizations. Consistent with this idea, current regulations acknowledge the ability of organizations to use approaches that significantly reduce, rather than eliminate altogether, the risk of re-identification. In addition to such legal requirements, an organization may have other motives for mitigating this risk. For instance, this could be driven by ethical standards, a commitment to customer trust or a desire to uphold strong data governance practices while still being able to gain significant business value from personal data. These considerations are in line with the proactive principle of "privacy by design" (Cavoukian, 2009).

URI: https://hdl.handle.net/10125/106505
978-0-9981331-7-1

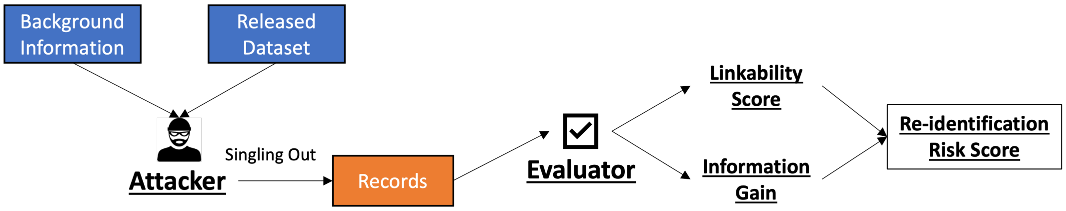

Figure 1. Re-identification risk evaluation of an attack scenario

Seeking to contribute to this important area of interest both in academia and in practice, we propose a risk assessment approach aimed at evaluating the potential threat of an attacker attempting to single out an individual's records within a released dataset. For example, an attacker could be an external entity gaining unauthorized access to a compromised dataset, or an internal employee who might unknowingly or intentionally search someone, perhaps unaware of the associated privacy implications. Singling-out is a notion that has been discussed in the GDPR and translated into mathematical terms (Cohen and Nissim, 2020; Francis et al., 2019). It refers to the potential to isolate certain records or all records that correspond to a single individual within a given dataset. This concept is presumed to form the foundation of the risk associated with re-identification. In our approach, the attacker leverages information resources at their disposal along with the released dataset to isolate a specific group of records. Subsequently, an evaluator scores the attack based on two factors: (1) the degree to which singled-out records can be linked and (2) the amount of information an attacker could gain through the attack (Figure 1). Unlike existing methodologies, we contend that these factors can be perceived as components within an attacker's valuation function, which concurrently serves as a singular re-identification risk score for an individual.

We estimate each score by simulating multiple attack scenarios, which can be modelled as independent attacks or as a collaborative game. The latter configuration enables us to employ the Shapley value to assess the value of an information resource to an attacker. The Shapley value is a concept from cooperative game theory developed by Shapley (1953) that has found applications in various fields, including economics, political science, and computer science (Roth, 1988), among others. In a cooperative game setup, players form a coalition or group and work together to achieve a common goal. Here, the goal of the adversarial player(s) is to re-identify and retrieve as much information as possible on individuals. When each player is assigned a type of resource to attack a released dataset, we propose that Shapley values can provide explanations for the risk of re-identification.

In our case study, we primarily explore tabular data given its widespread use as a format of confidential information that underpins many of an organization's key decisions. We conduct our experiments on the Adult Income Census (AIC) dataset, chosen for its representativeness. We assess two principal strategies for data release: the first is anonymizing the data using a k-anonymity method and the second is synthesizing data using deep learning models. The findings illustrate that both data anonymization and data synthesis yield comparable tradeoffs using our risk evaluation. In addition, the study also provides insights into which features of the AIC are particularly valuable for attackers.

Our primary contribution is made to the literature on data privacy. Specifically, our proposed measure for re-identification risk scores defines a privacy strength metric associated with data releases. This metric can be used alongside an existing data utility metric to better understand the privacy-utility trade-off. Our risk-based approach also has implications for practice as it does not focus solely on worst-case scenarios, providing insights on all individuals associated with a data release strategy. Finally, our approach is compatible with economic analyses that can guide organizations in selecting the right data release strategy.

The remainder of this paper is structured as follows. We begin with a succinct overview of the literature on privacy threats and associated risks. In Section 3, we outline the elements of our risk-based approach for handling tabular data, discussing its key assumptions and measures. Section 4 explains how game theory and the concept of Shapley value can be harnessed to investigate re-identification. In Section 5, we showcase the practical application of our approach through a case study using the AIC dataset. Finally, we summarize our key contributions and provide concluding remarks.

2. Background and Related Work

In 2006, when a major streaming company published a so-called “anonymized dataset” for a public online contest with the goal of improving their recommendation engine, they believed they had sufficiently reduced the risk of re-identification, given the nature of the data being made public. Narayanan and Shmatikov (2008) managed to re-identify a small subset of the individuals who were also included in the IMDB open database. The following class action lawsuit caused the company to settle for 9 million dollars in 2011. Other cases of re-identification attacks on published datasets have been documented (Henriksen-Bulmer and Jeary, 2016), although details on their economic implications for organizations often remain scarce. Nevertheless, there is a general consensus within the privacy community that the residual risk of re-identification associated with the use of data anonymization techniques, regardless of the type of technique being used, depends on the probability of success of re-identification attacks. To minimize this risk, several approaches have been proposed, including generalization-based transformations such as k-anonymity, as well as synthetic data.

K-anonymity (Sweeney, 2002) can reduce the risk of re-identification if measured by the probability of an attacker guessing the record belonging to the right individual. Indeed, when a dataset is k-anonymized, the attacker can only find groups of k indistinguishable records, reducing the expectation of successful guesses to at best $1/k$. Despite its advantages, it has been demonstrated that this method fails to encompass the entirety of privacy risks, as is the case when the risk associated with attribute disclosure is significant. Indeed, attribute disclosure can be viewed as a form of reconstruction attack in which the attacker’s objective is to reconstruct sensitive information about individuals without necessarily re-identifying them. In addition, it is important to distinguish between the information one wishes to protect and any inferences that can be drawn from the dataset (Li and Li, 2009). While our objective is to establish a definition of privacy safeguarding individuals, it is equally important to ensure that this definition allows for insightful analysis (Dwork and Roth, 2014). This shows that the risk of re-identification encompasses more than identifying a specific record and extends beyond mere attribute inference.

As a technique that is gaining attention in the industry, the use of synthetic data involves the generation of new data by a model that is trained on existing data. While it might seem counter-intuitive to attempt re-identification attacks on synthetic data, given the perception that its “fake” nature implies reduced re-identification risk, the reality is more complex. It is unclear whether anonymization is worse than synthetization in this context. However, like any other semantic privacy-preserving technique, it is crucial to acknowledge that there remains a possibility for leakage of personal information, for instance through reconstruction and membership attacks (Dwork et al., 2017). A membership attack can be understood as an attack in which the adversarial agent guesses that an individual was a member of a training dataset based on an observed output. There exist membership attacks specifically designed for data synthesis, such as distance-based attacks presented by Hilprecht et al. (2019). While synthetic data, therefore, holds promises in theory to preserve data privacy, recent findings suggest that it may not be a panacea.

Data protection laws that have been adopted over the past few years seek to regulate the use of personal data by organizations. Among those, the GDPR is considered the most comprehensive regulatory framework, and it has effectively influenced many aspects related to the governance, management and use of personal data by organizations, as well as data protection laws in other jurisdictions (*e.g.*, Canada). Within this context, the GDPR mentions anonymization as a valid technique for irreversibly transforming data, albeit without explicitly incorporating data synthesis as one of the possible ways to achieve this objective. One potential reason for this shortcoming is that data synthesis remained largely within the realm of research when the GDPR was adopted in 2016. Nevertheless, the GDPR highlights three main types of risk factors associated with data releases that should be mitigated, regardless of the nature of the data release strategy itself: “singling-out”, as formally defined by Cohen and Nissim (2020), involves isolating an individual or a specific record within a dataset; “linkability” refers to the ability to link records from the same individual across different datasets; and “inference” refers to the attribute disclosure risk discussed above. Giomi et al. (2022) have developed a unified framework for measuring these three factors separately to quantify the degree of risk associated with synthetic data. They argue that in synthetic data, linkability arises from the statistical similarities between the synthetic data and the original data. Stadler and colleagues also have portrayed membership attacks as a linkability risk factor to compare data synthesis against data anonymization (Stadler et al., 2022).

Measuring uniqueness has been used as an alternative method to assess the risk of re-identification (Skinner and Holmes, 1998). For instance, de Montjoye et al. (2015) demonstrate that four spatiotemporal points are sufficient to uniquely re-identify 90% of individuals in a dataset containing three months of credit card records for 1.1 million people. Dankar et al. (2012) build on the work of Skinner and Elliot (2002) to evaluate the risk of re-identification in 6 datasets by estimating the uniqueness of individuals. Their method for estimating uniqueness is based on the assumption that the population originates from a larger, super-population, thus turning it into a question of parameter estimation. This could be seen as a way to measure the concept of singling-out. However, this approach often showcases bias, as the risk evaluator is limited in terms of the amount of information available on the data distribution. Additionally, attackers generally have only partial access to background knowledge, indicating that uniqueness is not the sole determining factor in singling-out individuals in datasets.

Overall, research on data privacy has greatly contributed to our understanding of the nature and the negative impacts of the risks associated with the re-identification of individuals as well as various techniques available to try and mitigate these risks. Notwithstanding, we observe that there is still a need to further develop approaches that can reconcile the demands of data protection regulations, the need for organizations to generate business value using data, and the rapid technological advances that allow attackers to perform re-identification attacks at low cost. In particular, we argue that decision-makers would benefit from an increased ability to assess the trade-offs between data privacy and data utility in the context of data releases. To help fill this important gap, the next sections detail the constituting elements of our approach in the context of tabular data. Our approach is built on the idea that as managers, decision-makers need to make decisions regarding data releases on a frequent basis. To help them make these decisions in an informed manner without remaining stuck in a state of “paralysis by analysis” that can hinder value creation, we draw from the scientific literature on data privacy and game theory to formally quantify the degree of risk associated with a data release strategy.

3. Re-Identification Risk Evaluation for Tabular Data

Our approach starts with attackers seeking to single out individuals in a released dataset T to obtain the best possible score per individual. We assume there is an original dataset D and that the type of attack is the same whether T was produced through data anonymization or data synthesis. What changes is the background information an attacker has access to. In the literature, ”background information” typically refers to the auxiliary data or knowledge that an attacker may have. In this paper, we extend this concept and call it ”information resources,” or ”resources” in short, because when combined these can provide attackers with more re-identification power. For instance, an attacker who possesses both the “age’ and the “zip code” of an individual has more resources compared to an attacker with only the ”age” information.

We now describe the type of singling-out attack implemented in our approach. The singling-out process involves isolating one record or a small group of records. The symbol α will be used interchangeably with the singling-out function that takes T and r as inputs and returns a subset C of records in T, such as $C = \alpha(r, T)$. Obtaining a subset C instead of just a single record is consistent with scenarios in which T is a k-anonymized dataset in which each anonymized record has at least one duplicate. Because the attacker looks at records in T that are similar to his resources r to re-identify one individual, we select the records that minimize the distance between r and any t contained within T.

$$\alpha(r, T) = \min_{t \in T} d(r, t) \tag{1}$$

We use the unnormalized Gower distance as our base method as this distance measure can handle a combination of continuous and categorical features that are characteristic of tabular data. We vectorize all resources to compute distances. A key advantage of the vectorization of the resources is that we can use neighborhood search algorithms to speed up the simulation of attack scenarios. In particular, instead of computing a complete similarity matrix between vectors r and t, we can use a subset of the matrix. We used a BallTree (Omohundro, 1989), mainly for its computational efficiency, to estimate the neighborhood of r in T.

Having discussed the singling-out function employed by the attacker, we now turn to the role of the evaluator. To simplify notations, we refer to an attack scenario s as an attacker α utilizing their resources r to single out a specific individual within T. The evaluator has the responsibility to score attacks, which serves also as a risk evaluation of any plausible attack scenario s. We assume that the evaluator is aware of the data transformation strategy employed. Records that have been singled out by the attacker constitute a risk only if they can be linked to a real individual. In our framework, the evaluator estimates a linkability score, denoted L_s, of an attack scenario. Similar to other membership risk studies (Giomi et al., 2022;

Hilprecht et al., 2019; Houssiau et al., 2022; Lu et al., 2019), our linkability score is based on the distance between a matched record and the original record of the individual. We start by estimating a relative average distance radius μ of the closest neighbors of o. Only the singled out records within the μ radius are considered possible members of D, see Figure 2 as an illustration. We denote the set of possible members as Ω, in which $\Omega = \{t \in C : d(o, t) \leq \mu\}$. We go further by estimating the uniqueness of o. The rationale is that the uniqueness of an individual should come into play in the measure of linkability. If the attacker knows that an individual is unique and lacks close neighbors, they can more easily link the individual with their resources and deduce accurate information, thus posing a greater risk.

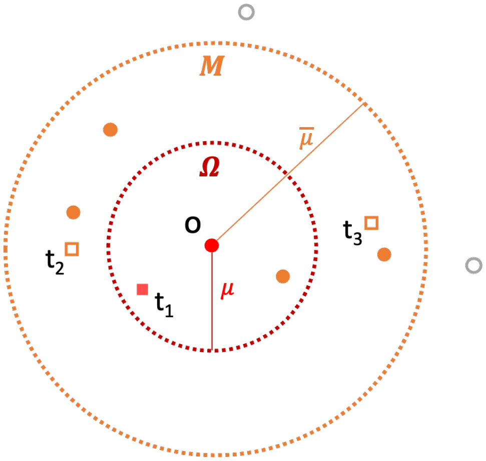

Figure 2. Attacker singles out 3 data points, represented by the squares t_1, t_2 and t_3. Only the data points inside Ω are considered possible members. Here, only t_1 meets that criterion. Original data points in M are used to estimate the uniqueness of the targeted individual o.

As we have seen in the background and related work section, there are many ways to estimate the uniqueness of a record. We chose to estimate the relative density of o using the average neighbourhood radius $\bar{\mu}$. The way $\bar{\mu}$ is selected as a direct impact on the linkability score, as it embeds assumptions of the evaluator regarding the membership risk. We denote the set of possible members as M, in which $M = \{t \in C : d(o, t) \leq \bar{\mu}\}$. The uniqueness of a record o, denoted ρ, is proportional to the size of the set M. To ensure its value falls within the range of 0 and 1, we perform normalization. Thus, the linkability score L_s of an attack scenario s is the percentage of singled out records that are close enough to the original record o, the records in set Ω, multiplied by the complement of the uniqueness ρ.

$$L_s = \frac{|\Omega|}{|C|} \cdot (1 - \rho) \tag{2}$$

The evaluator also has to determine if an attack scenario leads to an information gain, denoted I_s. We claim that I_s depends on the reconstruction loss of each possible match t in C. Let o be the original record of an individual in Dm $A' \subseteq A$ be the set of attributes to be reconstructed and A the set of all attributes. We express the reconstruction loss, denoted $E(t)$, of a matched record t, as an unormalized Gower distance between $o[A']$ and $t[A']$.

$$E(t) = \sum_{i \in A'} \zeta_i \cdot |o_i - t_i| \tag{3}$$

We denote ζ_i as the information gain per distance unit on the i-th feature. By default, we consider all ζ_i to be equal. We can now express the average information gain I_s of scenario s as the following formula:

$$\begin{aligned} I_s &= \frac{1}{|C|} \sum_{t \in C} \gamma_s \cdot \left(1 - \frac{E(t)}{\sum_{i \in A} \zeta_i}\right), \\ \text{in which } \gamma_s &= \frac{\sum_{i \in A'} \zeta_i}{\sum_{i \in A} \zeta_i}. \end{aligned} \tag{4}$$

When the reconstruction loss is 0, the information gain is directly proportional to the sum of the normalized weights of the attributes to reconstruct. The default assumption is that the more attributes the attacker can reconstruct, the more information they can gain. Following that logic, when there are no attributes to reconstruct, the information gain is 0.

Finally, we can define the re-identification risk score RRS of an individual.

Definition 1 (Re-identification risk score)**.** *Let S be the set of plausible attack scenarios on an individual. Each scenario $s \in S$ has a linkability score, denoted by L_s, and an information gain, denoted by I_s, such that $0 \leq L_s \leq 1$ and $0 \leq I_s \leq 1$. The constant ϵ allows the evaluator to give a minimum weight to linkability even when I_s is 0.*

$$RRS = \frac{1}{|S|} \sum_{s \in S} L_s \cdot max\{I_s, \epsilon\} \tag{5}$$

The RRS of an individual measures how much we estimate the singled out records in T to be linked to the individual, moderated by how much information can be gained by the attacker. Formula 5 suggests that the information gain is taken into account only when a record is linkable. Furthermore, if the information gain for a particular scenario s is equal to 0 but the records are linkable, the associated risk is equal to δ. To score a release strategy, we can simply compute the expected re-identification risk score of all individuals present in dataset D.

4. Re-identification Shapley Value

Attack scenarios can be performed individually and independently to evaluate the RRS of each person. Nevertheless, we purport that modelling our problem as a re-identification game can provide additional insights. The entire simulation can be perceived as consisting of multiple attackers, each of them possessing their own resources and focusing on a specific group of individuals. Consider for instance a scenario in which multiple adversarial agents conduct attacks simultaneously, with the evaluator providing risk scores for each individual. The re-identification risk scores can be combined to calculate the score of a sample of individuals. We can repeat this process until this score converges. We create resources by using D to sample possible subsets of resources. As attackers' resources often overlap, this process is akin to a simulation that employs sampling with replacement.

Let us now consider the simulation as a cooperative game involving attackers with varying resources who collaborate to re-identify and obtain as much information as possible on individuals. Let α_1 and α_2 represent two attackers with resources r_i and r_j respectively, targeting the same individual. The combined attacker set $\alpha_{1,2}$, acts as a "super" attacker with combined resources (r_i, r_j). There are now three possible attack scenarios: s_1, s_2, and $s_{1,2}$. The scenarios s_1 and s_2 correspond to attacks using resources r_i and r_j, respectively, while $s_{1,2}$ involves attacking with the combined resources (r_i, r_j). We assume that the combined attacker set $\alpha_{1,2}$ is always stronger than any individual attacker, such that $RRS_{s_{1,2}} \geq RRS_{s_1}$ and $RRS_{s_{1,2}} \geq RRS_{s_2}$. This assumption is consistent with the additive property of the value of information, suggesting that increased resources lead to higher re-identifying capabilities. This also implies that increasing resources never negatively impact re-identification, which we acknowledge may not hold true in some instances. However, in practice, this simplistic assumption appears to work well, primarily because released datasets tend to possess non-redundant features and preserve good utility.

To better understand the motives underpinning potential cooperation in the context of re-identification attacks, we now turn to the evaluation of the rewards associated with the success of an attack, as well as the question of how attackers should split the gains in such an instance. To answer this question, we borrow from the concept of Shapley value, which provides a way to allocate total gains or costs among players in cooperative games. In our context, we adapt this concept and refer to it as Re-identification Shapley Value (RSV):

Definition 2 (Re-identification Shapley Value). *Let's consider a re-identification game with a player set α_N of size n, let ϕ the valuation function of the re-identification of an individual and $S \subseteq \alpha_N \backslash \{i\}$ a subset of players that does not include player i. The re-identification Shapley value $\psi_i(v)$ of player i is defined as follows:*

$$\psi_i(\phi) = \frac{1}{n} \sum_{S} \left(\begin{array}{c} n-1 \\ |S| \end{array} \right)^{-1} \phi(S \cup \{i\}) - \phi(S) \quad (6)$$

If the valuation function is the RRS and if we assign specific resources to an attacker, the RSVs serve as an equitable measurement of how these resources are valuable for re-identification, from the point of view of the evaluator. One of the interesting properties of Shapley values is linearity. More precisely, the linearity axiom states that for any payoff function v that is a linear combination of two other payoff functions u and w, the Shapley values of v equal the corresponding linear combination of the Shapley values of u and w. If we apply this axiom to RSVs, we can combine the Shapley values of attackers when they collaborate to re-identify more than one individual. If each attacker is responsible for one type of resource corresponding to a feature in the dataset T, the RSVs for all individuals in D can be interpreted as the most valuable resources or features to re-identify individuals in T.

In machine learning, Shapley value has been used for measuring the value of features at inference. In particular, the popular SHAP (SHapley Additive exPlanations) framework (Lundberg and Lee, 2017), is particularly notable for its use of Shapley value. It is used to fairly distribute the contribution of each feature to the prediction of a model. Our definition of RSVs is consistent with these prior ideas as we seek to explain the RRS by measuring the contribution of resources in the re-identification process.

5. Case Study

To evaluate our approach, we use the Adult Income Census (AIC) public dataset. The AIC dataset contains sensitive attributes such as age, race, sex, marital status and native country, which can potentially be used to re-identify individuals. In addition, every feature of the AIC dataset can be considered as a potential identifying information resource, or an attribute to infer, making it readily usable by privacy researchers. Initially, we produced several anonymized variations of the dataset by implementing the Mondrian anonymization algorithm (LeFevre et al., 2006) with k-anonymity parameters $k = 5$, $k = 25$ and $k = 100$. We also used the Synthetic Data Vault library (Patki et al., 2016) to generate two synthetic datasets using respectively the CTGAN and TVAE implementations. The parameters were chosen based on common practices found in existing literature, rather than attempting to optimize them specifically for the AIC dataset. All strategies are compared to the original dataset.

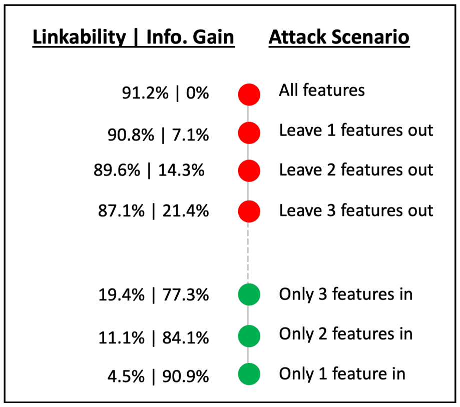

Figure 3. Average linkability scores and information gains of attacks on the original AIC dataset grouped by attack scenario types.

For each dataset, multiple simulations were performed, each involving batches of 100 individuals and ran for 200 iterations. One way to differentiate between simulations on a single dataset is by attack scenario type. For example, we ran a simulation to estimate the risk of re-identification in a scenario in which an attacker has access to all features as resources (Figure 3). Although this may not be entirely realistic, it can provide us with a boundary on how easy it is to link a record in the dataset. Unsurprisingly, when an attacker has access to all resources, and attacks the original AIC dataset, the average linkability score is 91.2% (Figure 3). However, it is important to note that the attacker has no attributes to infer in that case. It becomes more interesting when the attacker has fewer resources. For instance, we find that the average linkability is 19.4% when they have access to 3 features.

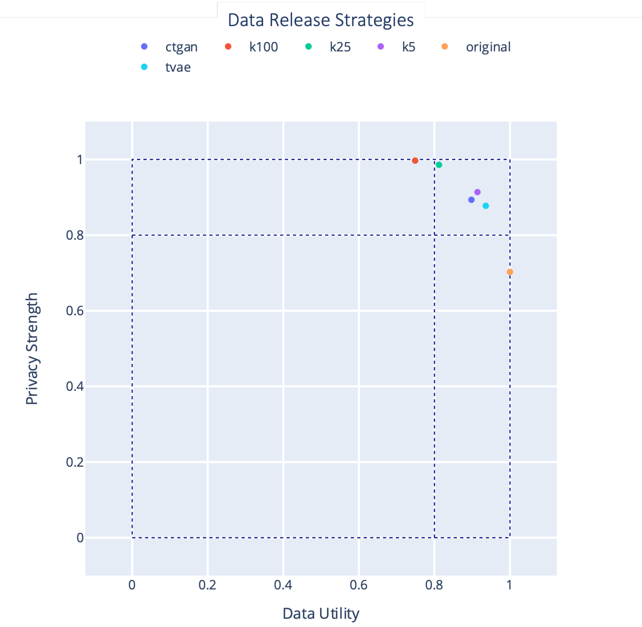

Figure 4. Trade-off between Data Utility and Privacy Strength (1 - Average RRS) of AIC Data Release Strategies. The threshold to determine what is High Protection/Utility is arbitrarily chosen to be 0.8.

While it would ultimately be up to decision-makers within an organization to determine the plausibility of the different scenarios to consider, our approach remains versatile and adaptable to different conditions and contexts. We illustrate how one may compute the score of a given strategy by averaging all scenarios. In a real-world setting, it would also be important to assign a minimum weight to the linkability scores. In our experiments, we arbitrarily chose $\epsilon = 0.5$. The results for all data strategies are presented in Figure 4 and Table 1. To compute the utility score of a dataset, we used a Kolmogorov-Smirnov test (Massey Jr, 1951) to compare the distribution of each dataset. We opted for this test because it is a non-parametric method that makes no assumptions about the forms of the underlying distributions being compared. Based on our simulations, the results showed that data anonymization and data synthesis offered similar trade-offs when $k = 5$, bringing further support to the argument that data synthesis does not completely eliminate the risk of re-identification.

We simulated enough attack scenarios to use the re-identification data to estimate the Re-identification Shapley Values for synthetic data using TVAE (Figure 5). It is interesting to note that the capital-gain, capital-loss and native-country features have little value in attacks. We decided to investigate the relationship

Table 1. AIC Re-Identification Simulations Results

Dataset	Linkability (Avg ± Std)	Information Gain (Avg ± Std)	Re-identification Score (Avg ± Std)	Utility Score
Original	$0.562 \pm 1.11e{-}3$	$0.422 \pm 2.19e{-}5$	$0.297 \pm 8.77e{-}4$	1.000
k5	$0.172 \pm 7.49e{-}4$	$0.420 \pm 1.46e{-}5$	$0.086 \pm 3.78e{-}4$	0.911
k25	$0.028 \pm 3.19e{-}4$	$0.419 \pm 1.47e{-}5$	$0.014 \pm 1.63e{-}4$	0.812
k100	$\mathbf{0.006 \pm 1.46e{-}4}$	$\mathbf{0.418 \pm 1.49e{-}5}$	$\mathbf{0.003 \pm 7.54e{-}5}$	0.749
ctgan	$0.213 \pm 8.71e{-}4$	$0.419 \pm 1.49e{-}5$	$0.107 \pm 4.37e{-}4$	0.898
tvae	$0.245 \pm 9.99e{-}4$	$0.419 \pm 1.48e{-}5$	$0.123 \pm 5.02e{-}4$	**0.936**

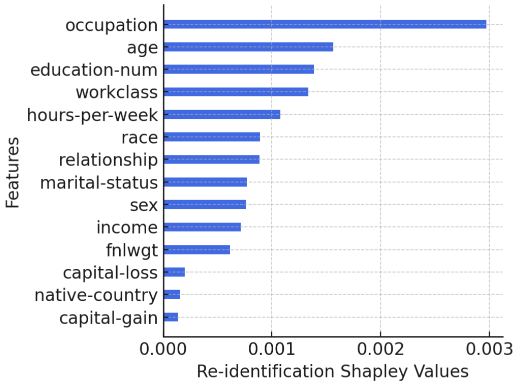

Figure 5. Re-identification Shapley Values of the simulation on synthetic AIC dataset (TVAE)

between the Re-identification Shapley Values (RSVs) and the entropy of the ACI features, suspecting that entropy could serve as a singling-out factor. Upon conducting a correlation analysis using Pearson's correlation coefficient, we found a substantial positive correlation between the two. This indicates that features with higher entropy are often associated with higher RSVs, suggesting a strong linear association between these variables.

6. Discussion

Our experiments on the Adult Income Census (AIC) public dataset demonstrated the practicality of our approach in evaluating the risk of re-identification with respect to different data release strategies. However, our framework has at least four limitations stemming from the assumptions made. These limitations could lead to an underestimation or an overestimation of the risk we want to evaluate. First, in our setup, we framed the risk evaluation as a kind of game where attackers submit records and get a score. Other setups could be explored to evaluate this risk. Second, our approach uses a specific implementation of singling-out attacks. Our use of the unnormalized Gower distance relies on fixed weights associated with resources. A potential improvement to overcome this limitation would be to use a different type of predictor that is not subject to the same constraints, while remaining suitable for use with tabular data. Third, our linkability evaluation method depends heavily on parametrization, which may vary across datasets. As an alternative, an attack such as Shokri's membership attack (Shokri et al., 2017) on synthetic data is pertinent. Similarly, in the evaluation of the information gain, a machine learning model could be employed instead of using distance metrics. Our current implementation of information gain does not consider the potential inferences that could be drawn from the reconstructed information. Finally, we assume attackers have access to resources in the same format as the original dataset, which may lead to an overestimation of the risk. In real-life scenarios, attackers often have aggregated or noisy information. While this means that our approach errs on the side of caution by being more conservative than it may be necessary, future work could simulate different conditions for a more realistic assessment of the risk of re-identification.

An important avenue for future research would be to extend our approach to specifically account for and handle sparse data (*e.g.*, text, transactions, etc.). The presence or absence of local patterns is often abundant due to high dimensionality. It is crucial to understand what makes an individual unique and how to automatically capture these patterns. An attacker could either learn these patterns (*e.g.*, via social engineering) or discover them in the anonymized dataset to re-identify an individual. To illustrate this challenge, consider transaction data. If we simply sum up transaction similarities to target individuals, we will underestimate the risk, as we would not account for the order of transactions and possible behavioural patterns. This limitation of our current approach offers an interesting direction for future research.

Despite these limitations, we believe our approach would be particularly useful for data custodians such as data governance and cybersecurity teams. For example, they could use our approach every time they are about to release data, either internally or to share outside of the organization. Our approach also remains versatile in that we do not explicitly define criteria to decide the right protection and utility thresholds, as these depend on the nature of the data, its classification (*e.g.*, PII), and the difficulty for an attacker to gain access to information resources. To gain further monetary insights into the costs an organization may incur from a partially anonymized or even synthetic dataset breach, a sensitivity analysis can be added to extend our privacy scores. As an example, consider that we estimate that each original record leaked would cost $1 to an organization. If an anonymized dataset contains 10,000 records and its average score is 0.1, we can estimate the breached costs in dollars to be $1,000. Similarly, each point in data utility lost because of privacy-preserving techniques should be taken into account. This is not the focus of this paper, but it is an interesting direction to explore in the future. For example, we could substitute our use of a Kolmogorov-Smirnov test for a utility metric specific to a team in the organization. A drop of 1% in utility for a team could be more costly than for another team. If both the privacy strength and data utility are converted into dollars, they can provide more pragmatic insights for decision-makers. Even if the complete estimation of the monetary costs associated with data breaches may be challenging, recent events attest to the importance of this issue. According to a recent IBM report, the average cost of a data breach in the U.S. is approximately $5 million (IBM, 2022).

Finally, we used Shapley Value for its well-established axiomatic properties, such as linearity, which provide both robustness and interpretability to our analysis. Alternative methods (Condevaux et al., 2022), could offer similar explanatory benefits while being more computationally efficient in certain contexts, particularly with larger datasets.

7. Conclusion

In this work, we have proposed an approach for evaluating the risk of re-identification in tabular data, by considering the risk from the perspective of what would be valuable to an attacker in singling out individuals. Our approach allows us to account for the information resources at the disposal of attackers and their use to single out individuals. Furthermore, we introduced an evaluator to score attacks, which serves as an assessment of the risk of re-identification of individuals based on two factors: the linkability of singled out records and the information gain of an attacker. This evaluation yields a re-identification risk score for each individual, enabling comparisons across different data release strategies. We also introduced the concept of Re-identification Shapley values (RSVs) to estimate the value of information in a privacy attack as a form of cooperative game involving multiple attackers based on our approach. We tested our approach on several anonymized and synthetic versions of the Adult Income Dataset, demonstrating its usefulness in a realistic, albeit fictitious, scenario. While data requirements and their associated impacts remain contextual to the organization in which they are considered, our approach provides valuable insights for data custodians, helping them balance the trade-off between data privacy and data utility.

Acknowledgments: We thank Sébastien Gambs, Genevieve Chafouleas, and Ishika Dhall for insightful comments on an earlier version of this paper.

References

Cavoukian, A. (2009). Privacy by Design - the 7 Foundational Principles. *Office of the Information and Privacy Commissioner*.

Cohen, A., & Nissim, K. (2020). Towards formalizing the GDPR's notion of singling out. *Proceedings of the National Academy of Sciences*, *117*(15), 8344–8352. https://doi.org/10.1073/pnas.1914598117

Condevaux, C., Harispe, S., & Mussard, S. (2022). Fair and efficient alternatives to shapley-based attribution methods. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 309–324.

Dankar, F. K., El Emam, K., Neisa, A., & Roffey, T. (2012). Estimating the re-identification risk of clinical data sets. *BMC Medical Informatics and Decision Making*, *12*(1), 66. https://doi.org/10.1186/1472-6947-12-66

de Montjoye, Y.-A., Radaelli, L., Singh, V. K., & Pentland, A. ". (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, *347*(6221), 536–539. https://doi.org/10.1126/science.1256297

Dwork, C., & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211–407. https://doi.org/10.1561/0400000042

Dwork, C., Smith, A., Steinke, T., & Ullman, J. (2017). Exposed! A Survey of Attacks on Private Data. *Annual Review of Statistics and Its Application (2017)*.

Francis, P., Probst-Eide, S., Obrok, P., Berneanu, C., Juric, S., & Munz, R. (2019, August 21). *Diffix-Birch: Extending Diffix-Aspen*. arXiv: 1806.02075 [cs].

Giomi, M., Boenisch, F., Wehmeyer, C., & Tasnádi, B. (2022, November 18). *A Unified Framework for Quantifying Privacy Risk in Synthetic Data*. arXiv: 2211.10459 [cs].

Gootjes-Dreesbach, L., Sood, M., Sahay, A., Hofmann-Apitius, M., & Fröhlich, H. (2020). Variational Autoencoder Modular Bayesian Networks for Simulation of Heterogeneous Clinical Study Data. *Frontiers in Big Data*, *3*.

Henriksen-Bulmer, J., & Jeary, S. (2016). Re-identification attacks—A systematic literature review. *International Journal of Information Management*, *36*, 1184–1192. https://doi.org/10.1016/j.ijinfomgt.2016.08.002

Hilprecht, B., Härterich, M., & Bernau, D. (2019). Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. *Proceedings on Privacy Enhancing Technologies*, *2019*(4), 232–249. https://doi.org/10.2478/popets-2019-0067

Houssiau, F., Jordon, J., Cohen, S. N., Daniel, O., Elliott, A., Geddes, J., Mole, C., Rangel-Smith, C., & Szpruch, L. (2022, November 11). *TAPAS: A Toolbox for Adversarial Privacy Auditing of Synthetic Data*. arXiv: 2211.06550 [cs].

IBM. (2022). Cost of a data breach 2022. https://www.ibm.com/reports/data-breach

LeFevre, K., DeWitt, D., & Ramakrishnan, R. (2006). Mondrian Multidimensional K-Anonymity. *22nd International Conference on Data Engineering (ICDE'06)*, 25–25. https://doi.org/10.1109/ICDE.2006.101

Li, T., & Li, N. (2009). On the tradeoff between privacy and utility in data publishing. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 517–526. https://doi.org/10.1145/1557019.1557079

Lu, P.-H., Wang, P.-C., & Yu, C.-M. (2019). Empirical Evaluation on Synthetic Data Generation with Generative Adversarial Network. *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*, 1–6. https://doi.org/10.1145/3326467.3326474

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68–78.

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. *2008 IEEE Symposium on Security and Privacy (Sp 2008)*, 111–125. https://doi.org/10.1109/SP.2008.33

Omohundro, S. M. (1989). *Five Balltree Construction Algorithms — PDF — Algorithms And Data Structures — Areas Of Computer Science*.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data Synthesis based on Generative Adversarial Networks. *Proceedings of the VLDB Endowment*, *11*(10), 1071–1083. https://doi.org/10.14778/3231751.3231757

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410. https://doi.org/10.1109/DSAA.2016.49

Roth, A. E. (1988, October 28). *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press.

Shapley, L. S. (1953). Stochastic Games*. *Proceedings of the National Academy of Sciences*, *39*(10), 1095–1100. https://doi.org/10.1073/pnas.39.10.1095

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18. https://doi.org/10.1109/SP.2017.41

Skinner, C. J., & Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 855–867. https://doi.org/10.1111/1467-9868.00365

Skinner, C. J., & Holmes, D. J. (1998). Estimating the Re-identi®cation Risk Per Record in Microdata.

Stadler, T., Oprisanu, B., & Troncoso, C. (2022, January 24). *Synthetic Data – Anonymisation Groundhog Day*. arXiv: 2011.07018 [cs].

Sweeney, L. (2002). K-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570. https://doi.org/10.1142/S0218488502001648

Wan, Z., Zhang, Y., & He, H. (2017). Variational autoencoder based synthetic data generation for imbalanced learning. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–7. https://doi.org/10.1109/SSCI.2017.8285168

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019, October 27). *Modeling Tabular data using Conditional GAN*. arXiv: 1907.00503 [cs, stat]. https://doi.org/10.48550/arXiv.1907.00503