A decision-support framework for data anonymization with application to machine learning processes

Loredana Caruccio[a], Domenico Desiato[a,*], Giuseppe Polese[a], Genoveffa Tortora[a], Nicola Zannone[b]

[a]*Department of Computer Science, University of Salerno, via Giovanni Paolo II n.132, 84084 Fisciano (SA), Italy*
[b]*Eindhoven University of Technology, Eindhoven, Netherlands*

Abstract

The application of machine learning techniques to large and distributed data archives might result in the disclosure of sensitive information about the data subjects. Data often contain sensitive identifiable information, and even if these are protected, the excessive processing capabilities of current machine learning techniques might facilitate the identification of individuals, raising privacy concerns. To this end, we propose a decision-support framework for data anonymization, which relies on a novel approach that exploits data correlations, expressed in terms of relaxed functional dependencies (RFDs) to identify data anonymization strategies providing suitable trade-offs between privacy and data utility. Moreover, we investigate how to generate anonymization strategies that leverage multiple data correlations simultaneously to increase the utility of anonymized datasets. In addition, our framework provides support in the selection of the anonymization strategy to apply by enabling an understanding of the trade-offs between privacy and data utility offered by the obtained strategies. Experiments on real-life datasets show that our approach achieves promising results in terms of data utility while guaranteeing the desired privacy level, and it allows data owners to select anonymization strategies balancing their privacy and data utility requirements.[1]

*Corresponding author
Email addresses: lcaruccio@unisa.it (Loredana Caruccio), ddesiato@unisa.it (Domenico Desiato), gpolese@unisa.it (Giuseppe Polese), tortora@unisa.it (Genoveffa Tortora), n.zannone@tue.nl (Nicola Zannone)

[1]This is a post-peer-review, pre-copyedit version published in Information Sciences Journal, Elsevier. 613: 1-32 (2022). The final authenticated version is available online at: https://doi.org/10.1016/j.ins.2022.09.004

Keywords: Privacy preserving machine learning, k-anonymity, Relaxed functional dependencies, Generalization strategies

1. Introduction

The increasing amounts of data available together with the advances in information technology have brought several benefits and opened new opportunities for the industry, individuals, and society. In particular, Big Data analytics has enabled the development of increasingly sophisticated applications ranging from personalized medicine and e-commerce to crowd management and fraud detection (Meijaard et al., 2020). However, these applications have also introduced new privacy and ethical challenges (Rathore et al., 2017). Big Data typically holds large amounts of personally identifiable information (e.g., criminal records, shopping habits, credit and medical history, and driving records), which can enable mass surveillance and profiling programs and raise several privacy issues (Koshley et al., 2017; Genga et al., 2022; Caruccio et al., 2020c; Ding et al., 2020).

To prevent these issues arising, data protection and privacy frameworks usually define strict requirements on the collection and processing of personally identifiable information (Guarda & Zannone, 2009; Riva et al., 2020). For instance, the General Data Protection Regulation (GDPR) [2] requires organizations to collect, process, and share personal data only for legitimate and lawful purposes, and to periodically identify privacy risks that can affect the data subjects.

Employing all the measures and procedures for the protection of personally identifiable information, as required by data protection regulations and, especially, by the GDPR, can be expensive for organizations. Thus, many organizations need to ensure that the personal data they collect for data analytics are sufficiently anonymized to reduce the associated compliance burdens (Caruccio et al., 2020a).[3] To this end, they often eliminate any unique identifier for each user when collecting personal data. However,

[2] General Data Protection Regulation - Final version of the Regulation URL: `http://data.consilium.1125europa.eu/doc/document/ST-5419-2016-INIT/en/pdf`.

[3] Notice that the principles of the GDPR do not apply to anonymized information, i.e., information from which the data subject is no longer identifiable.

this in itself may not solve the problem, since removing unique identifiers might not be sufficient to guarantee data anonymity (Zigomitros et al., 2020). In fact, anonymized data could be de-anonymized through cross-referencing with data gathered from other sources (Ni et al., 2022; Veeningen et al., 2014). Moreover, the application of machine learning techniques to anonymized data might still lead to the disclosure of sensitive and confidential information about data subjects, thanks to the power of current predictive models. On the other hand, we might still want to enable machine learning and data analytics processes to extract useful knowledge and insights from data while avoiding the disclosure of sensitive information. Thus, the challenge is to devise anonymization techniques that do not allow re-identification of individuals by using machine learning techniques on anonymized data (Zigomitros et al., 2020).

To anonymize data within data sharing and analytics contexts, several techniques relying on cryptography, randomization, perturbation, etc. have been proposed (Li et al., 2020; Pramanik et al., 2021). In this work, we focus on anonymization techniques based on generalization. The latter consists of replacing attribute values with more generalized ones so as to make the records in a dataset indistinguishable from each other (Sweeney, 2002). While protecting the privacy of individual records in the data, the application of generalization results in information loss, which affects the utility of the data for subsequent analysis (Ni et al., 2022). Therefore, existing solutions typically propose approaches to satisfy anonymity constraints that minimize information loss or to find a trade-off between privacy and data utility requirements (Esmeel et al., 2020). However, they often do not account for correlations in the data when applying generalization strategies, which can excessively penalize data utility.

To overcome these problems and derive a complete anonymization process, we propose a novel decision-support framework for data anonymization that *(i)* exploits (multiple) data correlations, represented as relaxed functional dependencies (RFDs) (Caruccio et al., 2021), in order to define generalization strategies that guarantee the required level of privacy and *(ii)* supports the entity responsible for the anonymization of the data (e.g., the data owner) in balancing privacy and data utility requirements. The main novelty of this work lies in the definition of a set of techniques enabling the exploitation of RFDs for data anonymization. In particular, our approach relies on RFDs to

capture data correlations over the complete set of data (instead of restricting the scope to correlations between single attributes and the class attribute), while also considering the level of generalization for each attribute. The extracted RFDs are then used as a baseline to derive local generalization schemes where each attribute can be generalized at a different level of granularity. In order to increase the data utility of the anonymized data, we also investigate how to combine RFDs, thus exploiting multiple data correlations, to devise anonymization strategies that account for a larger number of attributes while ensuring the required level of privacy. To measure the privacy level and data utility provided on a given dataset by an anonymization strategy, we use well-known metrics. In particular, we measure the privacy level of anonymized datasets based on the k-anonymity model proposed in (Sweeney, 2002), and use classification accuracy and information gain as measures for data utility. Several anonymization strategies could potentially satisfy the minimum privacy level required by the data owner. Our approach also offers data owners and other stakeholders a framework to guide them in the selection of the anonymization strategy to apply, by facilitating understanding of the trade-off between privacy and data utility. To this end, we employ a multi-objective optimization method based on Pareto optimality (Petchrompo et al., 2022) to assist data owners in selecting optimal anonymization strategies according to their privacy and data utility requirements.

We performed several experiments using publicly available datasets to demonstrate the applicability of our approach to achieve anonymization in data sharing contexts. Results show that there exists indeed a trade-off between privacy and data utility when anonymizing a dataset. The experiments demonstrate that the proposed approach provides an effective way to assist data owners in identifying anonymization strategies that guarantee (at least) the desired level of anonymity while reducing the loss of data utility caused by generalization. In particular, combining generalization rules and, thus, exploiting multiple data correlations in the definition of anonymization strategies makes it possible to achieve higher data utility compared to using single RFDs extracted from data.

The contribution of the work can be summarized as follows:

- We propose an approach to devise anonymization strategies that exploit correlations

in the data, specified in terms of RFDs, to limit the loss of data utility when anonymized datasets are used for classification activities.

- We show how to exploit multiple data correlations in anonymization strategies to achieve the highest possible data utility for the level of privacy requested by the data owner.
- We provide guidelines to help the data owner identify the anonymization strategies providing an optimal trade-off between privacy and data utility.
- We have performed extensive experiments to evaluate our approach and compare it with existing anonymization techniques using three real-life datasets.

The remainder of the paper is organized as follows. Section 2 introduces background concepts on relaxed functional dependencies and anonymization. Section 3 reviews related work. Section 4 presents the problem statement and Section 5 describes the proposed approach. Section 6 presents experiments and Section 7 discusses our findings. Finally, Section 8 concludes the paper and provides directions for future work.

2. Background

This section introduces background concepts used throughout the paper, such as those related to relaxed functional dependencies (RFD) and k-anonymity. To this end, let us first recall some basic concepts of relational databases.

A relational database schema $\mathcal{R}$ is defined as a collection of relation schemas (R_1,. . ., R_n), where each R_i is defined over a fixed set of attributes $attr(R_i)$, whereas $attr(\mathcal{R}) = \bigcup_{R_i \in \mathcal{R}} attr(R_i)$. Each attribute A_k has associated a domain $dom(A_k)$, which can be finite or infinite. A relation instance (or simply a relation) r_i of R_i is a set of tuples t such that for each attribute $A_k \in attr(R_i)$, $t[A_k] \in dom(A_k)$, $\forall\, t \in r_i$, where $t[A_k]$ represents the projection of t onto A_k, also denoted with $\Pi_{A_k}(t)$. A database instance r of a database schema $\mathcal{R}$ is a collection of relation instances (r_1,. . .,r_n), with r_i relation instance of R_i and $R_i \in \mathcal{R}$.

2.1. Relaxed Functional Dependencies

Several types of data dependencies have been defined and studied in the literature, including functional, join, and multivalued dependencies. In this work, we consider relaxed functional dependencies, an extension of functional dependencies (FDs).

Definition 1. (Functional dependency). Let R be a relation schema of a relational database schema $\mathcal{R}$, a functional dependency (FD) φ between two sets of attributes $X, Y \subseteq attr(R)$, denoted by $X \rightarrow Y$, specifies a constraint on the tuples that can form a relation instance r of R: $X \rightarrow Y$ iff for every pair of tuples t_1, t_2 in r, whenever $t_1[X] = t_2[X]$, then $t_1[Y] = t_2[Y]$. The two sets of attributes X and Y are also called Left-Hand-Side (LHS) and Right-Hand-Side (RHS), resp., of φ.

Relaxed Functional Dependencies (RFDs) extend FDs by relaxing some constraints of their definition. In particular, they might relax on the *attribute comparison* method or on the fact that the dependency must be valid on the entire database (relaxation on the *extent*). Next, we discuss the relaxation on the attribute comparison method only, since our approach relies on RFDs belonging to this category.

Relaxation on the attribute comparison method. This kind of relaxation allows the use of an approximate tuple comparison operator, say $\approx$, instead of the "equality" operator used in the FD definition. In order to define the type of attribute comparison method that is used within an RFD, we need to introduce the concept of *similarity constraint*.

Definition 2. (Similarity constraint). Given an attribute A with domain $\mathbb{D}$ and a threshold α, let $\phi[A] : \mathbb{D} \times \mathbb{D} \rightarrow \mathbb{R}$ be a function evaluating the similarity between two values in $\mathbb{D}$. A similarity constraint ϕ associated to attribute A, also denoted as $A_{\leq\alpha}$, indicates that a pair of values $a_1, a_2 \in \mathbb{D}$ can be considered similar if and only if $\phi[A](a_1, a_2) \leq \alpha$.

As an example, the function ϕ can be defined in terms of a similarity metric $\approx$, like for instance the edit or the Jaro distance (Elmagarmid et al., 2007), such that, given two values $a_1, a_2 \in A$, $a_1 \approx a_2$ holds iff a_1 and a_2 are "close" enough w.r.t. a predefined threshold α.

The concept of *similarity constraint* can be generalized in terms of *set of similarity constraints* defined over a set of attributes $X = \{A_1, \ldots, A_k\}$, and it is denoted as $\Phi = \{A_{1 \leq \alpha_1}, \ldots, A_{k \leq \alpha_k}\}$.

Based on the relaxation criterion introduced above, we provide a general definition of RFD:

Definition 3. (Relaxed functional dependency). Let R be a relation schema of a relational database schema $\mathcal{R}$ and r a relation instance of R, a relaxed functional dependency (RFD) ϱ on R, denoted by

$$X_{\Phi_1} \rightarrow Y_{\Phi_2} \tag{1}$$

where

- $X, Y \subseteq attr(R)$, with $X \cap Y = \emptyset$,
- Φ_1 and Φ_2 sets of similarity constraints on X and Y, respectively,

is said to be valid on r, or equivalently, r *satisfies* ϱ (denoted by $r \models \varrho$), iff for each pair of tuples t_1 and t_2 of r, if Φ_1 is true for each constraint $A_{\leq \alpha} \in \Phi_1$, then Φ_2 is true for each constraint $B_{\leq \beta} \in \Phi_2$.

In other words, if $t_1[X]$ and $t_2[X]$ agree with the constraints specified by Φ_1, then $t_1[Y]$ and $t_2[Y]$ must agree with the constraints specified by Φ_2.

Now, we introduce the notion of Roll-up dependency (RUD), which represents the specific type of RFD relaxing on the attribute comparison used in our approach. It maps the similarity constraints by means of the order relation defined in terms of generalization hierarchies (Calders et al., 2002). In particular, a generalization hierarchy contains several levels, on which an order relation $\preceq$ can be defined.

Given the layered structure of a generalization hierarchy, a relation schema R is defined as a set of attribute-level pairs, from which a "generalization" schema (hereafter called *genschema*) can be built by replacing a level l with a level l', with $l \prec l'$, and/or by entirely omitting certain attributes from R.

Given a genschema G of a relation schema R and an instance r of R, two tuples t_1, t_2 of r are said to be α-equivalent iff t_1 and t_2 become equal after rolling up their attribute values at most as many levels as the ones specified by α.

Definition 4. (Roll-up dependency). Let G be a genschema of a relation schema R and $X, Y \subseteq attr(R)$, a roll-up dependency (RUD) $X_{\Phi_1} \rightarrow Y_{\Phi_2}$ is valid on an instance r of R, if and only if for each tuple pair (t_1, t_2) of r, if $\Pi_X(t_1)$ and $\Pi_X(t_2)$ are α-equivalent, then also $\Pi_Y(t_1)$ and $\Pi_Y(t_2)$ must be α-equivalent.

2.2. K-anonymity

K-anonymity is a largely used anonymization technique, which has been introduced to reduce the risk of re-identification of anonymized data (Samarati & Sweeney, 1998). Some pieces of information in the data may not be unique identifiers by themselves, but their combination yields a unique identifier (Zigomitros et al., 2020). These pieces of information are typically referred to as *quasi-identifiers*. K-anonymity requires that quasi-identifiers appear in the data at least k times.

Definition 5. (k-anonymity). Let r be an instance of a relation schema $R = \{A_1, \ldots, A_n\}$, and $Q \subseteq attr(R)$, then $r_Q = \Pi_Q(r)$ is said to satisfy *k-anonymity* if for each tuple $t_Q \in r_Q$ there exist at least k tuples $t_i \in r$, with $1 \leq i \leq k$, such that $\Pi_Q(t_i) = t_Q$.

3. Related work

A large body of research has investigated how to train a classifier while preserving the privacy of individual records. Existing solutions can be categorized into two main classes: approaches based on cryptographic techniques, in which the classifier model is securely computed (Sheikhalishahi & Zannone, 2020), and anonymization techniques, in which data are perturbed before they are disclosed. Several anonymization techniques have been proposed over the years to enable the sharing of sensitive data (Majeed & Lee, 2021). The first proposed technique is k-anonymity (Samarati & Sweeney, 1998), which requires each record in the data to be indistinguishable from at least $k-1$ other records (cf. Section 2.2).

Although k-anonymity protects against identity disclosure, it fails to guarantee an adequate level of protection with respect to the disclosure of sensitive attributes. This has led to the definition of several anonymization techniques, e.g. ℓ-diversity, t-closeness, m-confidentiality p-probabilistic (see (Zigomitros et al., 2020) for a survey), which account for the semantic closeness and distribution of the values of sensitive attributes. More recently, differential privacy has been proposed to limit the disclosure of private information of individual records by introducing noise during the training of the classification model (Domingo-Ferrer et al., 2021). However, these techniques cannot be directly employed in our approach because, although they provide a more robust approach (compared to k-anonymity) for data perturbation, they do not offer a metric to measure the privacy level of a given dataset. Nevertheless, these techniques can be employed on top of our approach to provide additional privacy guarantees before the generalized dataset is disclosed.

k-anonymity is usually achieved using generalization (i.e., replacing attribute values with more generalized values, typically defined in an attribute taxonomy), and suppression (i.e., deleting/masking attribute values) (Majeed & Lee, 2021). In particular, different generalization strategies and schemes have been proposed. For instance, existing approaches use domain generalization hierarchies (DGH), in which attribute values are generalized by suppressing some parts of them (e.g., a digit in the ZIP code), or value generalization hierarchies (VGH), in which attribute values are aggregated into classes. Generalization can be applied to the data globally or locally (Zigomitros et al., 2020), where global schemes use the same generalization for all attributes (i.e., all attributes are generalized at the same level), and local schemes allow applying a different generalization for each attribute. While protecting the privacy of individual records in the data, the application of generalization results in information loss (Esmeel et al., 2020). For example, generalization strategies, especially those based on DGH, might not preserve correlations in the original data. Similarly, the use of global generalization schemes can result in a dataset that is too coarse-grained for further analysis, particularly when the attributes in the dataset exhibit different susceptibility to generalization. Therefore, in this work, we target local generalization strategies based on VGH.

Finding an optimal solution for the k-anonymity problem is, in general, NP-hard (LeFevre et al., 2006a). This has spurred the design of polynomial algorithms able to

find "good-enough" solutions for real-life datasets. Table 1 provides the characteristics of interest of existing techniques compared with those of our approach. In particular, we consider the approach used to determine the anonymization strategy (i.e., greedy, heuristic, and so on), the privacy model employed (i.e., k-anonymity, l-diversity, and so on), the anonymization techniques (i.e., generalization, suppression, and so on), the supported attribute type (i.e., numerical and/or categorical), the usage of attribute taxonomies, and the employed utility metrics (i.e., information gain, accuracy, and so on).

Some techniques aim to anonymize a dataset without taking into account its subsequent use. For instance, Optimal Lattice anonymization (OLA) (El Emam et al., 2009) exploits generalization and suppression to achieve k-anonymity by searching for an optimal node in a lattice structure representing possible generalization steps. LeFevre et al. (2006a) propose Mondrian, a top-down algorithm for achieving k-anonymity by partitioning the attribute domain space into multidimensional regions. The algorithm uses the highest generalization of quasi-identifiers as a starting point and, then, recursively specializes them into partitions by applying multidimensional cuts until no further cuts are available. Mondrian has been extended to exploit value generalization hierarchies (LeFevre et al., 2006b) and to support l-diversity (Ashkouti et al., 2021). Bild et al. (2018) present a data anonymization algorithm that provides k-anonymity and differential privacy guarantees. This algorithm uses attribute taxonomies and a randomization approach, implemented via sampling, to meet differential privacy. More specifically, the search strategy employs a (randomized) best-first search through the generalization hierarchies, by using a score calculated according to given data quality metrics (i.e., information loss, discernibility, and group size) in order to release a randomized version of a given dataset. Another well-known anonymization approach is top-down greedy (TDG), proposed in (Xu et al., 2006). It iteratively performs a binary data partitioning in combination with a heuristic to split the data into equivalence classes, and it uses normalized certainty penalty (NCP) as a data quality metric to assess the information loss caused by anonymization.

The approaches mentioned above rely on greedy and/or heuristic based solutions to satisfy k-anonymity, and possibly, other privacy models. Other approaches rely on different generalization techniques, such as clustering, and/or exploit data properties, such as

	Approach	Privacy guarantee	Anonymization technique	Attribute type	Attribute taxonomy	Data utility metric
(El Emam et al., 2009)	Greedy	k-anonymity	Generalization, Suppression	Numerical, Categorical	Yes	Information loss
(LeFevre et al., 2006a)	Greedy	k-anonymity	Generalization	Numerical	No	Discernibility metric
(LeFevre et al., 2006b)	Greedy	k-anonymity, l-diversity	Generalization	Numerical, Categorical	Yes	Entropy, Accuracy
(Ashkouti et al., 2021)	Greedy	k-anonymity, l-diversity	Generalization	Numerical, Categorical	Yes	Information loss
(Bild et al., 2018)	Greedy	k-anonymity, differential privacy	Randomization, Generalization	Numerical, Categorical	Yes	Information loss
(Xu et al., 2006)	Greedy, Heuristic	k-anonymity	Generalization	Categorical	Yes	Information loss
(Lin & Wei, 2008)	Clustering	k-anonymity	Generalization	Numerical, Categorical	No	Information loss
(Yan et al., 2021)	Clustering	k-anonymity	Generalization, Suppression	Categorical	Yes	Information loss
(Song et al., 2009)	Association Generalization	k-anonymity	Generalization	Numerical, Categorical	No	Information loss
(Friedman et al., 2008)	Greedy	k-anonymity	Suppression	Numerical, Categorical	No	Accuracy error, Information gain
(Fung et al., 2005)	Greedy	k-anonymity	Generalization	Numerical, Categorical	Yes	Accuracy error, Information gain
(Raj & D'Souza, 2021)	Greedy	k-anonymity	Generalization	Numerical, Categorical	Yes	Accuracy error, Information gain
(Kisilevich et al., 2010)	Greedy	k-anonymity	Suppression, Swapping	Numerical, Categorical	No	Accuracy, Information loss
(Eom et al., 2020)	Greedy	k-anonymity	Suppression	Categorical	No	Information loss, Accuracy
(Wang et al., 2020)	Heuristic	k-anonymity	Generalization, Suppression	Categorical	Yes	Accuracy, F-measure, Information loss
(Liu et al., 2019)	Heuristic	k-anonymity, l-diversity, t-closeness	Re-sampling	Numerical, Categorical	No	Information loss, Accuracy
Our approach	Relaxed Functional Dependencies	k-anonymity	Generalization	Numerical, Categorical	Yes	Accuracy, Information gain

Table 1: Related work categorized w.r.t. criteria of interests

functional dependencies. For instance, in (Lin & Wei, 2008), records are partitioned into equivalence classes by exploiting clustering, aiming to satisfy k-anonymity. At each iteration, the algorithm randomly extracts one record from the dataset and determines other closest $k-1$ records relying on the NCP distance function, which form an equivalence class with the extracted record. Yan et al. (2021) propose a weighted k-member clustering algorithm able to achieve k-anonymity for records encompassing both numerical and cat-

egorical attributes. This algorithm leverages a weighting stage and a series of weighting indicators to evaluate the outlyingness of records, facilitating the filtering of outliers and improving the clustering quality. Finally, Song et al. (2009) propose k-multiset dependency (K-MSD), an algorithm that uses association generalization (AG), i.e., a function mapping attribute values into their generalized versions, to provide k-anonymized datasets on which FDs are preserved. In particular, k-anonymity is considered as a kind of data dependency and is achieved by specifying K-MSDs among attributes. All these approaches aim to find an anonymization strategy that is "optimal" with respect to well-known and/or ad-hoc data quality measures and that satisfies a given level of privacy. In addition, viewing k-anonymity as a kind of functional dependency, the approach in (Song et al., 2009) guarantees k-anonymity by preserving data correlations expressed as FDs. In contrast, our approach uses data correlations, expressed as RFDs, also to guide the identification of suitable anonymization strategies, and not to merely define integrity constraints.

A number of approaches specifically target anonymity within classification contexts. Since the k-anonymity satisfiability still remains a complex problem, brute-force solutions have been only applied to specific application scenarios (Esmeel et al., 2020). More general approaches typically rely on approximate solutions that are able to provide good results in terms of classification accuracy. For instance, a general approach to achieve anonymization within various data mining problems, such as classification, association rule mining, and clustering, is proposed in (Friedman et al., 2008). This approach constructs a classification model similar to the well-known ID3 decision tree induction algorithm, and iteratively splits the data by selecting, among all attributes, the one achieving the highest gain (for a specific gain function, e.g., Information Gain or the Gini Index). Similarly, several approaches based on a top-down specialization strategy have been proposed (Fung et al., 2005; Raj & D'Souza, 2021). They aim to achieve anonymity while preserving its usefulness in classification by applying generalization steps in a top-down fashion, and by using a generalization taxonomy for categorical attributes and intervals for continuous ones. Then, these approaches employ information gain and anonymity loss as data quality measures to evaluate the effectiveness of the obtained generalization strategy. Kisilevich et al. (2010) propose an approach relying on both suppression and swapping to preserve anonymity in the context of classification.

Their approach leverages an existing classification tree induction algorithm, trained on quasi-identifiers, by manipulating tree leaves to achieve k-anonymity, and measuring the data utility by means of the information loss measure. On the other hand, Eom et al. (2020) propose a surrogate vector-based model to classify anonymized trajectory datasets. This model reduces the data dimension significantly, and prunes unnecessary candidate sequences through a length-based frequent pattern tree (LFP-Tree) to improve data utility while satisfying k-anonymity. To manage the k-anonymity satisfiability over high dimensional data, Wang et al. (2020) propose a novel heuristic method based on local recording. The approach vertically divides raw data into disjoint subsets to be anonymized, and exploits the k-anonymity requirements together with attribute correlations to guarantee a suitable level of data utility, measured by accuracy, F-measure, and information loss. Finally, Liu et al. (2019) propose a more general privacy-preserving method that uses conditional probability distribution to predict sensitive attribute values to be replaced, and relies on k-anonymity, l-diversity, and t-closeness to minimize differences in data distribution between the original and the re-sampled dataset, which has been then evaluated in terms of accuracy computed after applying several classification models.

The approaches discussed above either solve a different problem or tackle only partially the ones addressed in our proposal. In particular, all surveyed approaches aim to derive only one generalization strategy guaranteeing a given level of anonymity, and those targeting the classification domain also verify the achieved accuracy, but mostly a posteriori. Another limitation of the surveyed approaches lies in the fact that specialization steps are defined over a single attribute at a time, hence neglecting possible data correlations that would allow the simultaneous evaluation of multiple attributes for the definition of the generalization strategy. The only work exploiting data dependencies (Song et al., 2009) merely uses them as constraints to be verified upon the application of the K-MSD algorithm, but not to identify possible anonymization strategies. Moreover, as highlighted in Table 1, several approaches only support the anonymization of a single type of data, either categorical or numerical. Finally, several approaches do not rely on attribute taxonomies for generalization. Although the definition of these taxonomies requires some initial effort, approaches that do not employ them require

a computationally expensive pre-processing step, yielding possible distortions in the data and/or bias during classification processes. Moreover, their performances are often influenced by the dataset dimensionality.

In this work, we propose a decision-support framework for data anonymization that addresses the identified limitations. Differently from previous anonymization techniques that aim to define a single anonymization strategy satisfying a given level of anonymity, our framework addresses a more general problem. In particular, it provides data owners with an understanding of the trade-offs between privacy and data utility when anonymizing their datasets. The main novelty of the proposed framework lies in the usage of RFDs to directly evaluate combinations of attributes, together with possible generalizations over the data, also embedding a priori criteria to preserve classification accuracy while searching strategies guaranteeing k-anonymity. In particular, it uses RFDs extracted from the data to define a collection of candidate generalization configurations for data anonymization and leverages the Pareto principle to identify those configurations that provide an optimal trade-off between privacy and data utility. In the next section, we introduce the problem of data anonymization in classification processes and, then, we present our framework in Section 5.

4. Problem statement

Classification models capture correlations between the attributes of individuals and a class value, and are often used to predict the class value for any unseen new observation. Classification models are built from a training dataset, which might contain sensitive information. This information could be inferred from the classification model by exploiting the correlations encoded in the model (Majeed & Lee, 2021). To this end, training data are usually anonymized by removing identifiable information before the classifier is trained. However, data can still be re-identified using quasi-identifiers (Zigomitros et al., 2020).

Example 1. Let us consider the sample dataset in Table 2, which is extracted from the Adult dataset.[4] Each tuple describes an individual, where `age`, `workclass`, `fnlwgt`,

[4] `https://www.openml.org/d/179`

	age	workclass	fnlwgt	education	maritial-status	occupation	relationship	sex	capital-gain	classes
t_1	39	State-gov	77516	Bachelors	Never-married	Adm-clerical	Not-in-family	Male	2174	>50K
t_2	50	Self-emp-not-inc	83311	Bachelors	Married-civ-spouse	Exec-managerial	Husband	Male	0	>50K
t_3	38	Private	215646	HS-grad	Divorced	Handlers-cleaners	Not-in-family	Male	0	<=50K
t_4	53	Private	234721	11th	Married-civ-spouse	Handlers-cleaners	Husband	Male	0	<=50K
t_5	37	Private	159449	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Female	0	>50K
t_6	37	Private	284582	Masters	Married-civ-spouse	Exec-managerial	Wife	Female	0	<=50K
t_7	49	Private	160187	9th	Married-spouse-absent	Other-service	Not-in-family	Female	0	>50K
t_8	52	Self-emp-not-inc	209642	HS-grad	Married-civ-spouse	Exec-managerial	Husband	Male	0	<=50K
t_9	38	Private	45781	Masters	Never-married	Prof-specialty	Not-in-family	Female	14084	>50K
t_{10}	49	Private	159449	Bachelors	Married-civ-spouse	Exec-managerial	Husband	Male	5178	>50K

Table 2: A sample dataset containing users' information.

`education`, `maritial-status`, `occupation`, `relationship`, `sex`, and `capital gain` are attributes characterizing her, whereas attribute `classes` indicates whether her annual income is greater or lower than $50K$. From this sample dataset it is possible to narrow down tuple t_1 to a specific individual by looking, for instance, at the `age` attribute, as this is the only tuple for which `age` is equal to 39.

This simple example shows that only removing identifiable information from a dataset might not be sufficient to guarantee anonymization. Anonymized data can be re-identified by linking the data by means of other data sources (Veeningen et al., 2014; Goldstein & Shlomo, 2020). Therefore, before disclosing a dataset containing highly sensitive information, data owners often transform it to reduce the risk that its records can be re-identified. An anonymization model largely used for this is k-anonymity, which requires that at least k individuals in the dataset share the same set of attribute values (cf. Section 2.2 for details).

A common way to achieve k-anonymity is through generalization (Hoogervorst et al., 2019). Intuitively, generalization is used to replace the values in a dataset with more general values. For example, numerical data can be replaced by intervals, whereas categorical attributes can be generalized to higher conceptual values. Hence, the application of generalization results in more tuples to be indistinguishable (i.e., with identical quasi-identifiers), thus contributing to achieve the desired level of k-anonymity.

The values of an attribute can be generalized at a different granularity, providing different levels of generalization and therefore of k-anonymity. Generalization levels can be organized in a hierarchical structure (hereafter called *attribute taxonomy*), which can be used to regulate the level of generalization to be applied to an attribute. In this

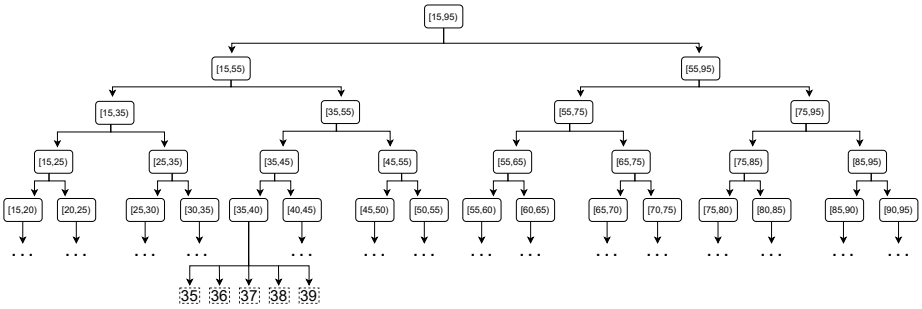

Figure 1: Taxonomy of the age attribute for the dataset in Table 2

work, we assume that every quasi-identifier in the dataset is associated with an attribute taxonomy representing all generalization levels defined for it.

Example 2. Figure 1 shows the taxonomy of the age attribute for the example dataset in Table 2. As shown in the figure, the leaf nodes (level 0) represent the values in Table 2 that can be generalized at different levels. For instance, value 39 can be replaced with interval $[35, 40)$ at level 1, with interval $[35, 45)$ at level 2, and so on. Based on the taxonomy for the attribute age in Figure 1, it is easy to observe that by applying generalization at level 1 for the age attribute on (a projection of) the sample dataset in Table 2 we achieve k-anonymity with $k = 2$ (cf. Table 3(a)), whereas we achieve k-anonymity with $k = 5$ by applying generalization at level 2 (cf. Table 3(b)).

This example shows that by increasing the generalization level of an attribute we can achieve a higher anonymity level (represented by the value of k). Nonetheless, the application of generalization can have a negative impact on data utility. For example, generalization can decrease the performance of a classifier when trained on an anonymized dataset, as generalization might weaken the correlations in the data (Last et al., 2014; Šarčević et al., 2020). Finding suitable generalization strategies that preserve anonymity while not affecting (too much) data utility is not trivial and requires finding a trade-off between anonymity and data utility. This trade-off boils down to determine suitable levels of generalization that guarantee data anonymization while maintaining as much data utility as possible.

In this work, we propose a novel anonymization technique that uses generalization

Table 3: generalization of (a projection of) the dataset in Table 2 over attribute age by considering two generalization levels defined in Figure 1.

(a) Level 1

	age
t_1	[35,40)
t_2	[50,55)
t_3	[35,40)
t_4	[50,55)
t_5	[35,40)
t_6	[35,40)
t_7	[45,50)
t_8	[50,55)
t_9	[35,40)
t_10	[45,50)

(b) Level 2

	age
t_1	[35,45)
t_2	[45,55)
t_3	[35,45)
t_4	[45,55)
t_5	[35,45)
t_6	[35,45)
t_7	[45,55)
t_8	[45,55)
t_9	[35,45)
t_10	[45,55)

and k-anonymity validation to anonymize a dataset while minimizing the loss of data utility. To this end, we exploit data correlations in the dataset, expressed in terms of relaxed functional dependencies (RFDs), as a guideline to define suitable generalization strategies. In the next section, we present our approach that, given a dataset and the attribute taxonomies as input, extracts RFDs that suggest generalization levels ensuring a given level of data anonymization while maintaining as much data utility as possible.

5. A decision-support framework for data anonymization

This section presents a decision-support framework for data anonymization. We show how data correlations, expressed in terms of relaxed functional dependencies (RFDs), can be used to devise strategies for the anonymization of datasets to be used for classification activities. In particular, our approach aims to identify anonymization strategies that comply with the privacy requirements of data owners for the sharing of their datasets while limiting the data utility loss due to the anonymization process. Intuitively, we use RFDs as guidelines to determine which subsets of attributes should be generalized and at which level, in such a way that the resulting anonymized dataset meets (at least) the minimum level of anonymity required by the data owner and, at the same time, its data utility is preserved as much as possible. The application of different

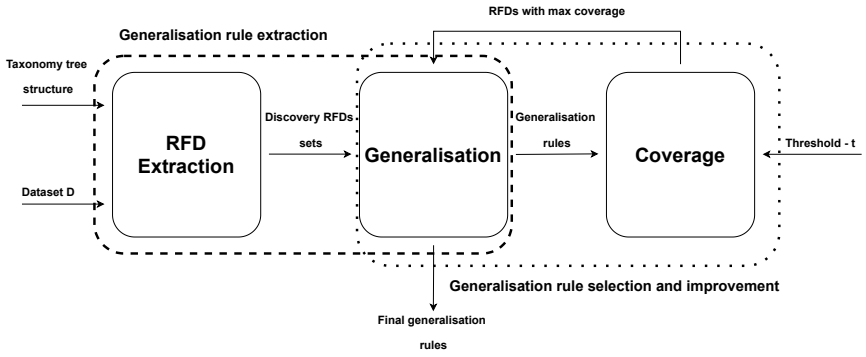

Figure 2: Overview of the approach

RFDs can result in different anonymization strategies, offering different levels of privacy and data utility. Our approach offers data owners and other stakeholders a framework to guide them in the selection of the anonymization strategy to apply. In particular, it enables the understanding of the trade-off between privacy and data utility offered by the obtained strategies by assessing their impact on data utility and their privacy guarantees.

5.1. Overview

Figure 2 shows an overview of our approach. Given an input dataset and a taxonomy of its quasi-identifiers, we first extract generalization rules expressed in terms of RFDs (*RFD Extraction*), and use them to determine which attributes should be generalized and at which level. To assess the quality of a generalization rule, we first apply it to the input dataset to replace attribute values with more general ones, and then compute the anonymity level and the data utility for the resulting generalized dataset (*Generalization*). In a second step, we extend the coverage of the RFDs that satisfy a given level of anonymity by joining generalization rules to increase data utility (*Coverage*). The data anonymization and utility provided by the obtained extended RFDs are then assessed as in the previous step (*Generalization*). The obtained generalization rules provide data owners with a view of which generalization rules can be used to anonymize their datasets and their effects in terms of data utility and anonymization. Next, we present the steps of our approach in detail.

5.2. Generalization rule extraction

The first phase of our approach (represented by the two blocks within the dashed line in Figure 2) aims to extract generalization rules in terms of RFDs and to determine the level of anonymity and data utility they achieve when applied on a dataset. RFDs are extracted from the input dataset, along with the generalization levels (defined with respect to the given attribute taxonomies), by using roll-up dependencies. In this process, all the attributes of the dataset are used for the extraction of RFDs. Recall from Section 2.1 that this is a type of RFD that allows to retrieve not only attribute correlations, but also the generalization level of the attributes, according to a given attribute taxonomy.

During RFD extraction, we only consider RFDs having the classification attribute (i.e., attribute `classes` in the example dataset of Table 2) on the right-hand side, with generalization level equal to 0. This is because we are interested in the generation of anonymized datasets that can be used to train a classification model. Accordingly, our focus is on correlations involving the classification attribute and preserving its original values.

Example 3. The classification attribute `classes` of the dataset in Table 2 can take two values, namely ">50K" and "<=50K". If this attribute is generalized to a single value, for example, *[Any classes]*, all tuples in the dataset will have the same value for it, making the dataset ill-suited to train a classification model.

The obtained RFDs identify which attributes along with their generalization level can allow performing classification activities, based on the data correlations within the dataset. Accordingly, each RFD can be used to produce an anonymized version of the dataset, in which only the attributes involved in the RFD are selected and generalized at the level specified by the RFD itself. This is done by replacing the value of the attributes in the original dataset with those defined in the specified level of the corresponding attribute taxonomy. All attributes that do not occur in the RFD are mapped to the highest level of the corresponding attribute taxonomy, as they are not involved in the correlation defined by the RFD.

Example 4. Suppose that the following RFD is extracted from the dataset of Table 2:

$$\texttt{age}_{\leq 3},\ \texttt{fnlwgt}_{\leq 2} \rightarrow \texttt{classes}_{\leq 0}$$

	age	fnlwgt	Classes
t_1	[35,55)	[0,100000)	>50K
t_2	[35,55)	[0,100000)	>50K
t_3	[35,55)	[200000,300000)	<= 50K
t_4	[35,55)	[200000,300000)	<= 50K
t_5	[35,55)	[100000,200000)	>50K
t_6	[35,55)	[200000,300000)	<= 50K
t_7	[35,55)	[100000,200000)	>50K
t_8	[35,55)	[200000,300000)	<= 50K
t_9	[35,55)	[0,100000)	>50K
t_{10}	[35,55)	[100000,200000)	>50K

Table 4: A sample application scenario of a single RFD.

The right-hand side of the RFD contains the classification attribute `classes`, whereas the left-hand side contains the subset of attributes `age` and `fnlwgt` to be generalized. The generalization level is defined by the values after the tag "$\leq$".

Table 4 shows the dataset resulting from the application of this RFD to the dataset in Table 2. We can observe that the attributes `age` and `fnlwgt` have been generalized by replacing their original values with those defined by the generalization level specified by the RFD (as an example, the taxonomy for attribute `age` is reported in Figure 1). The values of other attributes are generalized to the highest level. For the sake of clarity, we omitted them in Table 4.

Since the extracted RFDs provide different levels of data anonymization and data utility, such levels can be used to determine which RFD(s) should be used for the generation of the generalized dataset. We measure the privacy level offered by an RFD using the k-anonymity model proposed in (Sweeney, 2002), as described in Section 2.2.[5] Accordingly, given a dataset anonymized by applying the generalization rule, we compute the anonymity level provided by the generalized dataset as the minimum number of tuples that are indistinguishable with respect to the quasi-identifiers. It is easy to observe from Table 4 that the application of the RFD presented in Example 4 achieves a k-anonymity

[5]Since we measure privacy in terms of anonymity, the terms "privacy level" and "anonymity level" are used interchangeably in the context of this work.

level with $k = 3$. On the other hand, we measure the data utility of an RFD in terms of classification *accuracy* and *information gain*. Classification accuracy allows us to evaluate the data utility in the context of a classification model, whereas information gain provides us a general measure of data utility, which can be used to evaluate the effect of anonymization on a dataset w.r.t. data entropy.

In summary, this step of the approach returns a list of RFDs along with their anonymity level (measured in terms of k-anonymity) and data utility (measured in terms of accuracy and information gain), as illustrated in the following example.

Example 5. The following RFDs, along with their corresponding data anonymization and data utility measures, are extracted from the dataset in Table 2:

r_1: $[\texttt{age}_{\leq 3},\ \texttt{fnlwgt}_{\leq 2} \rightarrow \texttt{Classes}_{\leq 0}]; k : 3; A : 65; IG : 0.011657;$

r_2: $[\texttt{age}_{\leq 3}, \texttt{gender}_{\leq 1} \rightarrow \texttt{Classes}_{\leq 0}]; k : 4;\ A : 66;\ IG : 0.043581;$

r_3: $[\texttt{workclass}_{\leq 2}, \texttt{capital-gain}_{\leq 3}, \texttt{marital-status}_{\leq 2} \rightarrow \texttt{Classes}_{\leq 0}];$
$k : 3; A : 67; IG : 0.072174;$

r_4: $[\texttt{workclass}_{\leq 2}, \texttt{age}_{\leq 4}, \texttt{marital-status}_{\leq 2} \rightarrow \texttt{Classes}_{\leq 0}]; k : 5;$
$A : 61; IG : 0.007948;$

r_5: $[\texttt{relationship}_{\leq 1}, \texttt{education}_{\leq 2}, \texttt{capital-gain}_{\leq 3} \rightarrow \texttt{Classes}_{\leq 0}]; k : 2;$
$A : 68; IG : 0.079399;$

where k, A, and IG represent the anonymity level, accuracy, and information gain, respectively. We can observe that r_4 achieves the best anonymity level ($k = 5$), but the worst accuracy ($A = 61$). On the other hand, r_5 achieves the best accuracy ($A = 68$), but the worst anonymity level ($k = 2$).

As shown in the previous example, data owners are left with the task to determine which generalization rules should be used for the anonymization of their datasets. This can be a complex task, as a large number of RFDs can be potentially extracted from the dataset itself (Caruccio et al., 2020b,a), and not all of them might satisfy the desired level of anonymity. In addition, RFDs usually capture basic correlations in the data, involving a limited number of attributes and, thus, limiting the data utility that can be

achieved from their application. Increasing the number of attributes on the left-hand side of an RFD will make it possible to involve more attributes in the anonymization of the dataset, and thus, increase its data utility (Ni et al., 2022). However, the use of more attributes could reduce the level of anonymity guaranteed by the generalization rules. Therefore, the data utility can be improved only where, and to the extent that, the minimum level of anonymity required by the data owner is satisfied.

In the next section, we present our approach to identify the generalization rules satisfying a given level of anonymity while maximizing data utility. To this end, we devise an RFD join strategy to increase the length of their left-hand sides, in an attempt to increase the data utility provided by the baseline generalization rules obtained before joining the RFDs.

5.3. Generalization rule selection and improvement

This phase of the approach (represented by the two blocks within the dotted line in Figure 2) aims to generate a set of candidate generalization rules from the RFDs derived in the previous phase of the approach (cf. Section 5.2), which satisfy at least a given level of anonymity and, at the same time, limit the data utility loss due to the anonymization process.

Some RFDs identified in the previous step may not guarantee a level of anonymity that is acceptable for the data owner. In particular, the data owner might define minimum anonymization requirements for a dataset to be shared with other parties. According to the k-anonymity model, we model these requirements as a user-defined threshold t, indicating the minimum anonymity level that the dataset should satisfy in order to be considered for sharing. We use the threshold t to determine whether an RFD provides a sufficient level of anonymity. To check if an RFD is suitable for anonymization, the RFD is applied to the original dataset and the anonymity level k of the obtained anonymized dataset is computed using the k-anonymity model (cf. Section 5.2). If the anonymity level k of the obtained anonymized dataset is equal or greater than the user-defined threshold t, then the RFD satisfies the minimum anonymization requirements, and it is considered in the anonymization process; otherwise, the RFD is discarded.

The RFDs obtained in the previous phase capture only basic correlations in the data,

hence limiting the data utility that can be achieved through their application. To this end, we analyze the attributes involved in the RFDs and define a coverage strategy to increase the number of selected attributes to be used for the anonymization of the dataset. Our strategy compares the RFDs and determines which ones can be combined to improve data utility. The intuition is that joining RFDs allows to account for multiple data correlations simultaneously, hence increasing the number of attributes that can be used. Since combined RFDs have to be valid on the considered dataset, not all RFDs can be combined.

Before presenting the procedure for generating the candidate generalization rules, we introduce the notion of *compatible RFDs*, which specifies when two RFDs can be joined. Intuitively, two RFDs are compatible if and only if their left-hand side attributes are disjoint or occur with the same generalization level, as formalized in Definition 6.

Definition 6 (RFD Compatibility)**.** Let $X_{\Phi} \rightarrow C_{\leq 0}$ and $X'_{\Phi'} \rightarrow C_{\leq 0}$ be two RFDs such that $X = \{A_1, \ldots, A_n\}$, $X' = \{B_1, \ldots, B_m\}$, and each attribute A_i (B_j) is associated with a generalization level ϕ_i (ϕ'_j) in Φ (Φ'). We say that the two RFDs are compatible if and only if:

- $X \cap X' = \emptyset$, or
- $\forall A_i \in X$ and $B_j \in X'$, such that $A_i = B_j \in X \cap X'$, then $\phi_i = \phi'_j$.

Algorithm 1 presents the procedure used to generate the candidate generalization rules. The algorithm takes as input the list of RFDs Z obtained in the previous phase of the approach (cf. Section 5.2), the dataset D with the corresponding attribute taxonomies T, and a threshold t representing the minimum level of anonymity to be satisfied, and returns a list of candidate generalization rules R satisfying at least the required level of anonymity t along with their anonymity level and data utility measures. The algorithm uses three lists which are initialized to the empty set (line 1): R contains the RFDs satisfying the required level of anonymity t together with their anonymity level and data utility measures; Z' is a support list containing the RFDs that satisfy the required level of anonymity t, and W is a support list used to take track of the RFDs to join.

The first block of Algorithm 1 (lines 2 to 12) aims to determine the RFDs that satisfy the required anonymity level t and compute their data utility measures. Each

Algorithm 1 Join procedure

INPUT: Dataset D, taxonomy T, list of RFDs Z, threshold t

OUTPUT: List of generalization rules R

$R := \emptyset; Z' := \emptyset; W := \emptyset$
for each $(e_i \in Z)$ **do**
    $D' \leftarrow$`COMPUTE_generalization(`e_i`,`D`,`T`)`
    $k \leftarrow$`COMPUTE_k(`D'`)`
    if $(t \leq k)$ **then**
        $Z' \leftarrow Z' \cup \{e_i\}$
        $m \leftarrow$`COMPUTE_dataUtility(`D'`)` ▷ $m = (InfoGain, Accuracy)$
        $r \leftarrow (e_i, k, m)$
        $R \leftarrow R \cup \{r\}$
    end if
end for
$W := Z'$
while $W \neq \emptyset$ **do**
    $L := \emptyset$
    for each $(x_i, y_i \in W)$ **do**
        Let $x_i = X_\Phi \rightarrow C_{\leq 0}$
        Let $y_i = Y_{\Phi'} \rightarrow C_{\leq 0}$
        if $(X \cap Y = \emptyset) \vee (\forall a \in X \cap Y$ level$(Y[a]) =$ level$(X[a]))$ **then**
            $c_i = X_\Phi, Y_{\Phi'} \rightarrow C_{\leq 0}$
            $D' \leftarrow$`COMPUTE_generalization(`e_i`,`D`,`T`)`
            $k \leftarrow$`COMPUTE_k(`D'`)`
            if $(t \leq k)$ **then**
                $L \leftarrow L \cup \{c_i\}$
                $m \leftarrow$`COMPUTE_dataUtility(`D'`)`
                $r \leftarrow (c_i, k, m)$
                $R \leftarrow R \cup \{r\}$
            end if
        end if
    end for
    $W \leftarrow L$
end while
return R

RFD in Z is used to create a generalized version of the dataset D using the function `COMPUTE_generalization` (line 3). Then, the anonymity level of the generalized dataset D' is computed through the function `COMPUTE_k` (line 4), as described in Section 5.2. The RFD is then stored along with its anonymity and data utility measures in R (line 9).

Once the RFDs satisfying the required level of anonymity have been identified, the algorithm joins them to improve their data utility measures while guaranteeing that the baseline anonymization requirement t is still satisfied (lines 12 to 31). In particular, the algorithm uses a list W to keep track of which RFDs should be considered at each iteration to create new RFDs, which is initialized to the set Z' (line 13). The RFDs in W are analyzed pairwise (lines 15 to 29): if two RFDs x_i and y_i are compatible (cf. Definition 6), a new RFD c_i is created by joining them (lines 18-19). The new RFD c_i is then used to create a generalized dataset D' using the function `COMPUTE_generalization`, and the function `COMPUTE_k` is used to compute its anonymity level (lines 20-21). If the anonymity level of D' is greater than the user-defined threshold t, c_i is added to L and the data utility measures of D' are computed though the function `COMPUTE_dataUtility` (lines 22-26). Finally, c_i along with its anonymity level and data utility measures is added to R (line 27). After all rules in W have been analyzed, L contains the generalization rules obtained by combining the RFDs in W, which satisfy the minimum level of anonymity. These rules are used in the next iteration. The algorithm terminates when no generalization rule satisfying the minimum level of anonymity can be created, returning at least the RFDs in Z', and possibly new RFDs along with their anonymization and data utility levels.

It is worth noting that the RFDs analyzed during an iteration are exactly those obtained in the previous iteration (line 30). By doing so, no candidate RFD is missed. In fact, we can observe that: *(i)* a set of RFDs can be combined into a new RFD if and only if each RFD is compatible with the others; and *(ii)* if the combination of two RFDs does not meet the minimum anonymity level, any combination of RFDs that includes those RFDs will not satisfy the minimum anonymity level, and hence, it will be discarded.

Example 6. Consider the RFDs presented in Example 5 and a minimum anonymity level $t = 3$. Algorithm 1 filters the RFDs that do not meet the minimum anonymity level,

hence discarding r_4. The remaining RFDs are then analyzed pairwise, and compatible ones are combined, obtaining:

r_6: $[\texttt{age}_{\leq 3}, \texttt{fnlwgt}_{\leq 2}, \texttt{gender}_{\leq 1} \rightarrow \texttt{Classes}_{\leq 0}]; k:3; A:67; IG:0.074251;$

r_7: $[\texttt{age}_{\leq 3}, \texttt{fnlwgt}_{\leq 2}, \texttt{workclass}_{\leq 2}, \texttt{capital-gain}_{\leq 3}, \texttt{marital-status}_{\leq 2} \rightarrow \texttt{Classes}_{\leq 0}]; k:3; A:70; IG:0.098579;$

r_8: $[\texttt{age}_{\leq 3}, \texttt{gender}_{\leq 1}, \texttt{workclass}_{\leq 2}, \texttt{capital-gain}_{\leq 3}, \texttt{marital-status}_{\leq 2} \rightarrow \texttt{Classes}_{\leq 0}]; k:3; A:71; IG:0.099719;$

r_9: $[\texttt{workclass}_{\leq 2}, \texttt{capital-gain}_{\leq 3}, \texttt{marital-status}_{\leq 2}, \texttt{age}_{\leq 4} \rightarrow \texttt{Classes}_{\leq 0}]; k:3; A:69; IG:0.096718$

It is easy to observe that r_6 is obtained by joining rules r_1 and r_2, r_7 by joining rules r_1 and r_3, r_8 by joining rules r_2 and r_3, and finally, r_9 by joining rules r_3 and r_4. Notice that r_4 is not combined with r_1 and r_2 because they are incompatible: attribute age occurs at generalization level 4 in r_4 and at generalization level 3 in r_1 and r_2. Also, all rules satisfy the minimum anonymity level $k = 3$. Thus, the set of generalization rules $\{r_6, r_7, r_8, r_9\}$ is used in the second iteration.

By combining rules r_6 and r_7 (but also r_6 and r_8, or r_7 and r_8) we obtain rule

r_{10}: $[\texttt{age}_{\leq 3}, \texttt{fnlwgt}_{\leq 2}, \texttt{gender}_{\leq 1}, \texttt{workclass}_{\leq 2}, \texttt{capital-gain}_{\leq 4}, \texttt{marital-status}_{\leq 2} \rightarrow \texttt{Classes}_{\leq 0}]; k:3; A:72; IG:0.109829;$

On the other hand, rule r_9 cannot be merged with any other rule, due to its incompatibility on attribute age. As no new RFD can be created, the procedure returns the set of candidate generalization rules $\{r_1, r_2, r_3, r_4, r_6, r_7, r_8, r_9, r_{10}\}$, which represents all the generalization rules meeting the minimum anonymization requirement.

Algorithm 1 returns a list of candidate generalization rules satisfying the given minimum level of anonymity. They provide a different anonymization and data utility level, allowing the data owner to control the trade-off between these two dimensions. However, the large number of rules that can be potentially returned might hamper the selection of the generalization rule to be used. Identifying the optimal candidate rules can be seen as a multi-objective optimization problem and, thus, we use the notion of

Pareto-optimality and Pareto frontier (Petchrompo et al., 2022) to guide the data owner in the selection of suitable generalization rules.

In Pareto-optimality, the objective function comprises multiple criteria, and the multi-objective optimization problem can be formulated as follows:

$$\max F(X), \quad F(X) = f_1(X), f_2(X), \ldots, f_m(X) \tag{2}$$

where each $f_i(X)$, with $i \in \{1, 2, \ldots, m\}$, is a function determining a different objective, $F(X)$ is the multi-objective function, and X is a solution to the multi-objective optimization problem. A solution X is said to dominate a solution Y, if $f_i(X) \geq f_i(Y), \forall i \in \{1, 2, \ldots, m\}$, and there exists $j \in \{1, 2, \ldots, m\}$ such that $f_j(X) > f_j(Y)$. Solution X is called Pareto optimal if it is not dominated by any other solution. More than one Pareto-optimal solution exists when no solution dominates all the others. The curve or surface composed of the Pareto-optimal solutions is known as the Pareto frontier (Lotov & Miettinen, 2008).

We use the Pareto frontier to identify the generalization rules extracted from Algorithm 1 that are (Pareto) optimal with respect to anonymization and data utility. In this light, our objective functions are represented by the k-anonymity level, classification accuracy, and information gain, and the goal is to find the solutions that are not dominated by other ones. The generalization rules on the Pareto frontier are, thus, the rules that provide the data analyst with the best trade-off between anonymization and data utility requirements.

Example 7. Consider the generalization rules returned in Example 6, which are summarized in Table 5. The generalization rules on the Pareto Frontier are highlighted in gray. A visual representation of the Pareto frontier is shown in Figure 3, where the x-axis represents the accuracy level A, the y-axis the anonymity level k, and the z-axis the information gain IG. The blue points represent the Pareto Frontier, i.e. the generalization rules that are not dominated by any other rule (r_2, r_4 and r_{10}).

6. Experiments

We performed a number of experiments to evaluate the approach proposed in Section 5. In particular, we studied whether joining RFDs and, thus, accounting for a

Rule	Privacy	Accuracy	Information Gain
r_1	3	65	0.011657
r_2	4	66	0.043581
r_3	3	67	0.072174
r_4	5	61	0.007948
r_6	3	67	0.074251
r_7	3	70	0.098579
r_8	3	71	0.099719
r_9	3	69	0.096718
r_{10}	3	72	0.109829

Table 5: generalization rules returned in Example 6 with anonymization and data utility levels.

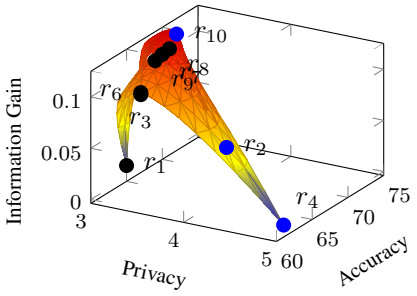

Figure 3: Visual representation of the Pareto Frontier for the generalization rules in Table 5.

larger set of attributes, result in anonymization strategies that allow to obtain anonymized datasets with higher data utility. Moreover, we investigated the trade-off between anonymization and data utility that can be achieved by using generalization rules and how to devise strategies for selecting the generalization rules to be used for data anonymization. More specifically, our experiments were driven by the following research questions:

RQ1: What is the impact of combining generalization rules on data utility?

RQ2: Which trade-off between privacy and data utility can be achieved by using generalization rules?

RQ3: How much effort is required by a data owner to identify the generalization rule to apply?

An assumption underlying our work is that combining generalization rules allows achieving a higher data utility as it allows exploiting multiple data correlations simultaneously (cf. Section 5.3). The first research question (**RQ1**) aims to test our hypothesis and provide insights on the impact that combined generalization rules produce on the data utility. **RQ2** aims to assess the trade-off between anonymization and data utility that can be achieved using generalization rules. In particular, we are interested in understanding how the enforcement of a given anonymity level impacts data utility, also in comparison with other anonymization algorithms. A large number of generalization rules could potentially

Datasets	#Rows	#Attributes	Attribute types
Electricity	45312	8	Numeric
Adult	48842	14	Nominal, Numeric
Bank	45211	17	Nominal, Numeric

Table 6: Statistics on the datasets used in the evaluation.

satisfy both anonymization and data utility requirements. This could affect the data owner, who has to decide which generalization rule to apply on her dataset. **RQ3** aims to evaluate the effort required to a data owner to determine the generalization rule to apply for the anonymization of her dataset, in terms of the number of rules returned by our approach. The remainder of this section presents the settings and the results of the experiments.

6.1. Experiment settings

Datasets. To evaluate the proposed approach, we used three real-world datasets, whose characteristics are reported in Table 6.

Electricity Dataset:[6] This dataset comprises records from the Australian New South Wales Electricity Market from May 1996 to December 1998. Each record refers to a period of 30 minutes, and is characterized by 8 numerical attributes, including the day of the week, the timestamp, the South Wales electricity demand, and the Victoria electricity demand. The class label identifies the price change (UP or DOWN) in New South Wales relative to a moving average of the last 24 hours.

Adult Dataset:[7] It describes 48842 individuals using a mix of numeric and categorical attributes (14 attributes in total), such as *age*, *occupation*, and *education*. The *class* attribute represents individuals' income, which has two possible values: '$> 50K$' and '$< 50K$'.

Bank Dataset:[8] It describes 45211 individuals using a mix of numeric and categorical attributes (17 attributes in total), such as *age*, *job*, and *balance*. The data is related to direct marketing campaigns of a Portuguese banking institution based on phone calls. The *class*

[6] `https://datahub.io/machine-learning/electricity`

[7] `https://archive.ics.uci.edu/ml/datasets/Adult?ref=datanews.io`

[8] `https://archive.ics.uci.edu/ml/datasets/bank+marketing`

attribute represents the bank term deposit, which has two possible values: "yes", and "no".

Attribute Taxonomies. Our approach requires the attribute taxonomies for the quasi-identifiers of the given dataset to enable data generalization. We computed the attribute taxonomy for numerical attributes by using a bottom-up approach, whereas for categorical attributes we used a top-down approach based on k-means clustering. Specifically, the generalization levels for numeric attributes were created by ordering the attribute values (i.e., the leaf nodes) in descending order, and by grouping them in sets of size five.[9] Then, at each level, pairs of contiguous sets were grouped to create a new level until a single set, representing the taxonomy's root, was created. On the other hand, k-means was applied over categorical attributes to ensure that similar tuples were grouped together to minimize accuracy loss. In particular, k-means was used to partition the set of all attribute values (the taxonomy's root) into two clusters, and then it was applied recursively to each cluster until no further split was obtained. The last level of the taxonomy (leaf nodes) was generated by creating a node for each attribute value, which was connected to the node representing the cluster containing that value. The final taxonomy was obtained by ensuring that each increase in the generalization level corresponded to an increase in the anonymity level. To this end, generalization levels that produced no improvement in terms of k-anonymity were removed from the taxonomy. Table 7 presents an overview of the size and number of taxonomy levels of each attribute in the Electricity[10], Adult[11], and Bank[12] datasets.

[9]The choice of the group size is justified by the fact that grouping five values provided a suitable trade-off between the improvement of k and information loss for each level of the taxonomy. In fact, choosing a smaller group size would lead to create a complex taxonomy with no improvement in terms of k for several taxonomy levels, whereas a larger size would lead to a taxonomy with few levels that, while providing improvements in terms of k at each level of the taxonomy, incurs higher information loss.

[10]The complete attribute taxonomies for the Electricity dataset are given in `https://raw.githubusercontent.com/dmndes/Taxonomies/main/TaxElectricity`

[11]The complete attribute taxonomies for the Adult dataset are given in `https://raw.githubusercontent.com/dmndes/Taxonomies/main/TaxAdult`

[12]The complete attribute taxonomies for the Bank dataset are given in `https://raw.githubusercontent.com/dmndes/Taxonomies/main/TaxBank`

Table 7: Overview of the attribute taxonomies for the considered datasets.

(a) Electricity

Attribute	Type	Domain size	#Taxonomy levels
date	numeric	934	5
day	numeric	8	3
period	numeric	47	4
nswprice	numeric	4088	7
nswdemand	numeric	5275	6
vicprice	numeric	6203	10
vicdemand	numeric	2845	6
transfer	numeric	1877	10

(b) Adult

Attribute	Type	Domain size	#Taxonomy levels
age	numeric	73	6
workclass	nominal	7	2
fnlwgt	numeric	26740	6
education	nominal	16	3
education-num	numeric	16	4
marital-status	nominal	7	4
occupation	nominal	14	3
relationship	nominal	6	2
race	nominal	5	3
sex	nominal	2	2
capitalgain	numeric	120	6
capitalloss	numeric	96	4
hoursperweek	numeric	95	4
native-country	nominal	40	4

(c) Bank

Attribute	Type	Domain size	#Taxonomy levels
age	numeric	77	6
job	nominal	12	4
marital	nominal	3	3
education	nominal	4	2
default	nominal	2	2
balance	numeric	7168	8
housing	nominal	2	2
loan	nominal	2	2
contact	nominal	3	2
day	numeric	31	4
month	nominal	12	4
duration	numeric	1573	7
campaign	numeric	48	4
pdays	numeric	559	8
previous	numeric	41	2
poutcome	nominal	4	3

RFD Extraction. To extract the RFDs used to generate generalization rules we employed the *DOMINO RFD discovery algorithm* (Caruccio et al., 2021). The advantage of using this algorithm is that it automatically infers not only the RFDs from data, but also their associated thresholds. *DOMINO* extracts RFDs that are valid on the entire dataset, i.e.,

every tuple pair in the dataset should satisfy the RFD similarity constraint in order to be returned by *DOMINO*. This can be too restrictive when discovering roll-up RFDs over generalization taxonomies. Thus, an RFD discovery algorithm tolerating exceptions would be needed. However, the only discovery algorithm for hybrid RFDs existing in the literature is not capable of automatically discovering similarity and coverage thresholds, requesting the user to specify them in input (Caruccio et al., 2020b). Given that in our context the automatic derivation of thresholds is a fundamental requirement, since they represent the generalization levels to be used, we decided to adopt a dataset sampling strategy and use *DOMINO* on a sampled dataset. In this way, *DOMINO* discovers roll-up RFDs that are not valid on the entire original dataset, hence increasing the set of discovered roll-up RFDs. In addition, we adapted *DOMINO* to create a generalization map in which keys represent distance patterns and values represent the number of tuple pairs complying with each pattern. Then, for each attribute in the considered dataset, a distance pattern (computed between each pair of tuples) maps the number of generalizations to use for including two attribute values in the same taxonomy level. In our experiments, we considered the most frequent distance patterns, yielding the coverage of an x-percentage of tuple pairs, with $x \in \{5, 10, 20, 50\}$. Then, we tested the effects of this sampling strategy at the vary of x.

Anonymization & Data Utility Measures. To determine the anonymity level offered by a generalization rule, we apply the generalization rule to the original dataset and compute the minimum number of tuples in the generalized dataset that are indistinguishable with respect to the quasi-identifiers, as described in Section 5.2. It is worth noting that, in the experiments, we considered all attributes in the datasets (except for the class attribute) to be quasi-identifiers. This allows us to consider the worst-case scenario in terms of data quality, in which k-anonymity should be guaranteed with respect to a larger set of attributes.

As discussed in Section 5.2, data utility is measured in terms of classification accuracy and information gain. In the experiment, classification accuracy was computed using *both* the $J48$ decision tree implementation of Weka[13] and Support Vector Machine

[13] `https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/J48.html`

implementation of Scikit-learn[14], whereas information gain was computed using the $J48$ decision tree implementation of Weka. It is worth noting that our approach is general and other machine learning algorithms and implementations could have been used to measure data utility. In general, the choice of the implementation to be used depends on the specific use case and, in particular, on the machine learning algorithm that will be applied by the data requester. To guarantee the reliability of the obtained predictive models, we used 10-fold cross-validation to compute the data utility measures.

Anonymization algorithms used for comparison. To evaluate the trade-off between privacy and data quality offered by the approach presented in Section 5, we also compared the approach with existing anonymization algorithms. For a fair comparison, we selected algorithms that work with attribute taxonomies, rely on generalization strategies for both categorical and numeric attributes, and are based on k-anonymity. In addition, we considered algorithms for which an implementation is available.

Among the existing anonymization algorithms, we selected three that satisfy the criteria above: Basic Mondrian (LeFevre et al., 2006b), TopDown Greedy Anonymization (TopDown) (Xu et al., 2006), and Datafly (Sweeney, 1997). Basic Mondrian is a top-down greedy data anonymization algorithm for relational datasets, which uses a greedy criterion to produce homogeneous partitions of data exploiting weighted entropy to generalize data. We chose Basic Mondrian over Mondrian (LeFevre et al., 2006a) for its support for both categorical and numerical attributes.[15] The TopDown Greedy Anonymization algorithm (TopDown) relies on binary partitioning to iteratively split data into subsets and uses a normalized central penalty (defined in terms of information loss) as the splitting criterion. Finally, Datafly counts the frequency of unique sequences of values over the quasi-identifiers. If a dataset does not meet the desired level of anonymity (in terms of k-anonymity), the algorithm generalizes the attribute having the largest number of distinct values until the anonymity requirements are satisfied. For the algorithms' implementation, we use the publicly accessible GitHub repositories of Basic

[14] `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html`

[15] Mondrian does not provide support for categorical attributes. Categorical attributes have to be transformed to numerical ones and this transformation can have a negative impact on the generalization for some applications.

Mondrian[16], TopDown[17] and Datafly[18].

The anonymization algorithms described above require as input the dataset to be anonymized, the desired anonymity level k, and a taxonomy for the quasi-identifiers. Since in the evaluation of our proposal we consider the complete set of dataset attributes as quasi-identifiers, for a fair comparison, we set up these anonymization algorithms to work with the complete set of attributes as well and used the attribute taxonomies employed to evaluate our proposal. Differently from our approach that returns multiple anonymization strategies (one per each generalization rule) providing different levels of anonymity, the anonymization algorithms presented in (LeFevre et al., 2006b; Xu et al., 2006; Sweeney, 1997) return an anonymized dataset that satisfies the level of anonymity provided as input. To compute the trade-off between anonymization and data utility offered by these algorithms and, thus, to enable the comparison with our approach, we used the levels of anonymity provided by the generalization rules obtained using our approach as the anonymity level to be achieved. This way, we can compute the data utility of the dataset anonymized using the other anonymization algorithms when the same level of anonymity is provided.

6.2. Results

In this section, we present the results of experiments and answer our research questions.

RQ1: What is the impact of combining generalization rules on data utility? This research question aims to evaluate the benefits of combining generalization rules, represented through RFDs, to generate strategies for data anonymization, which maximize data utility while guaranteeing a desired level of privacy. We expected that, on average, the combination of RFDs provides generalization rules with higher data utility compared to those directly extracted from the data. To measure this, we compare such sets of rules in terms of classification accuracy and information gain. In the analysis, we consider all generalization rules that achieve an anonymity level of at least 2 (i.e., $k \geq 2$).

[16] https://github.com/qiyuangong/Basic_Mondrian

[17] https://github.com/qiyuangong/Top_Down_Greedy_Anonymization

[18] https://github.com/fun-personal-projects/datafly

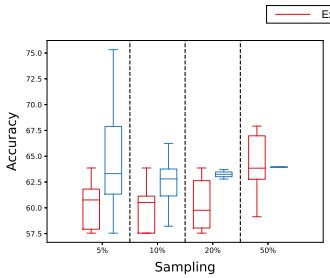

Figure 4: Accuracy achieved by generalization rules extracted directly from RFDS (*Extracted Rules*) and by generalization rules obtained by the combination of RFDS (*Combined Rules*) at the varying of the sampling percentage for the Electricity dataset

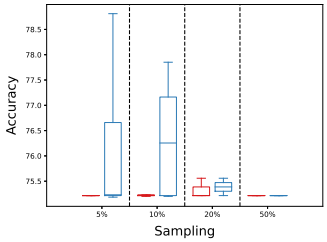

Figure 5: Accuracy achieved by generalization rules extracted directly from RFDS (*Extracted Rules*) and by generalization rules obtained by the combination of RFDS (*Combined Rules*) at the varying of the sampling percentage for the Adult dataset

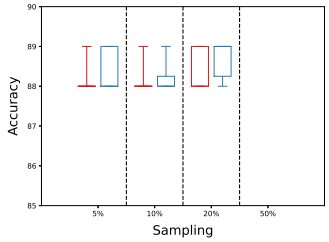

Figure 6: Accuracy achieved by generalization rules extracted directly from RFDS (*Extracted Rules*) and by generalization rules obtained by the combination of RFDS (*Combined Rules*) at the varying of the sampling percentage for the Bank dataset

Figures 4 to 9 show the accuracy that can be achieved using the generalization rules directly extracted from the data (red boxes) and using the combined rules (blue boxes) at the varying of sampling percentage for the Electricity, Adult, and Bank datasets. In particular, Figures 4, 5, and 6 report the accuracy scores obtained by using the ID3 decision tree classifier, whereas Figures 7, 8, and 9 report the accuracy scores obtained by using the SVM classifier.

From Figure 4, we can observe that, for the Electricity dataset, combining generalization rules improves the accuracy obtained using the ID3 decision tree classifier for all sampling percentages, except for the 50% sampling percentage. This is because many generalization rules extracted for this sampling percentage contain the same attributes with different generalization levels and, thus, they are incompatible (for more details see Section 5.3), or their combination violated the privacy requirement over k (i.e., $k < 2$).

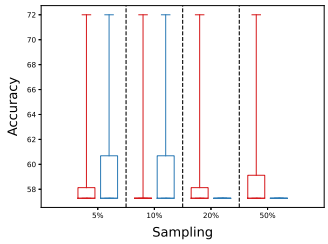

Figure 7: Accuracy achieved by generalization rules extracted directly from RFDs (*Extracted Rules*) and by generalization rules obtained by the combination of RFDs (*Combined Rules*) at the varying of the sampling percentage for the Electricity dataset (SVM)

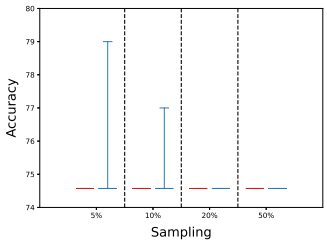

Figure 8: Accuracy achieved by generalization rules extracted directly from RFDs (*Extracted Rules*) and by generalization rules obtained by the combination of RFDs (*Combined Rules*) at the varying of the sampling percentage for the Adult dataset (SVM)

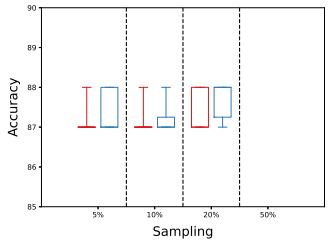

Figure 9: Accuracy achieved by generalization rules extracted directly from RFDs (*Extracted Rules*) and by generalization rules obtained by the combination of RFDs (*Combined Rules*) at the varying of the sampling percentage for the Bank dataset (SVM)

Figure 5 shows similar results for the Adult dataset, although the improvement is less prominent for this dataset. It is worth noting that the accuracy achieved for the Adult dataset, when it is anonymized using generalization rules directly extracted from the data, is already relatively high (over 75% vs. 60% for the Electricity dataset), given that the accuracy achieved on the original data is 85% (vs. 75% for the Electricity dataset). On the other hand, the improvement is very limited for the Bank dataset (cf. Figure 6). It is interesting to observe that, for this dataset, the 50% sampling does not return any generalization rule satisfying the privacy requirement, i.e., for every generalization rule $k < 2$.

When the accuracy is computed using the SVM classifier (cf. Figures 7, 8, and 9), we can observe a lower variability in accuracy on all datasets. Combining RFDs shows some improvement in accuracy when rules are extracted using a small sampling percentage, albeit the improvement is limited especially for the Adult dataset (cf. Figure 8). An

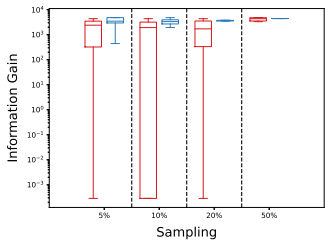

Figure 10: Information gain achieved by generalization rules extracted directly from RFDs (*Extracted Rules*) and by generalization rules obtained by the combination of RFDs (*Combined Rules*) at the varying of the sampling percentage for the Electricity dataset

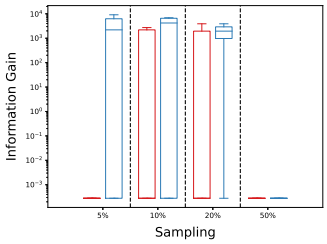

Figure 11: Information gain achieved by generalization rules extracted directly from RFDs (*Extracted Rules*) and by generalization rules obtained by the combination of RFDs (*Combined Rules*) at the varying of the sampling percentage for the Adult dataset

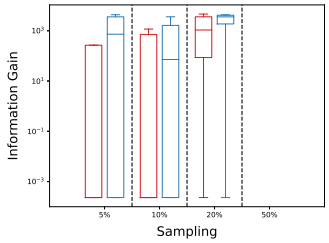

Figure 12: Information gain achieved by generalization rules extracted directly from RFDs (*Extracted Rules*) and by generalization rules obtained by the combination of RFDs (*Combined Rules*) at the varying of the sampling percentage for the Bank dataset

in-depth investigation showed that the low variability is due to the fact that the same classification accuracy is achieved for almost all generalizations rules.

Our experiments also show that combining generalization rules improves information gain for the Electricity, Adult, and Bank datasets, as illustrated in Figures 10, 11, and 12, respectively. Overall, the results confirm our hypothesis and show that considering more correlations in the data simultaneously and, thus, accounting for more attributes in the anonymization process, allows generating anonymized datasets holding a higher data utility.

RQ2: Which trade-off between privacy and data utility can be achieved using generalization rules? We expect that data utility decreases when the anonymity level increases. This is because achieving a higher level of anonymity requires higher generalization levels, leading to less specificity of data. To understand which trade-off

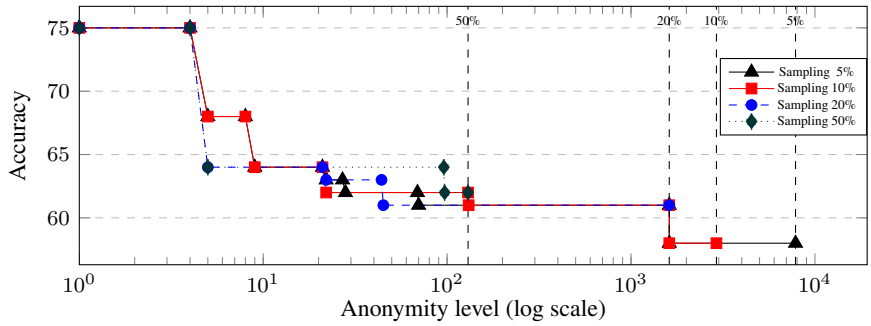

Figure 13: Trade-off between privacy and accuracy for the Electricity dataset (ID3).

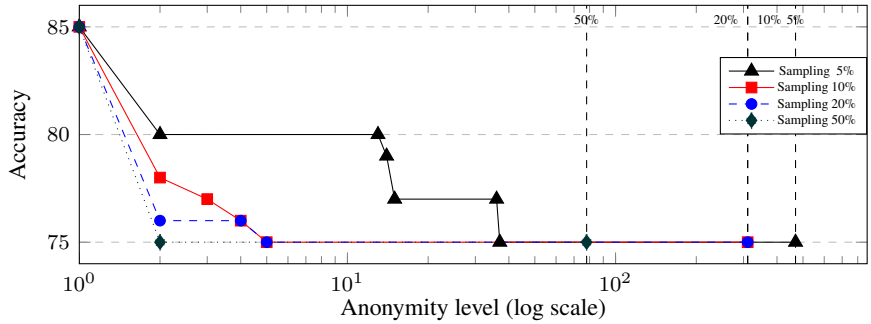

Figure 14: Trade-off between privacy and accuracy for the Adult dataset (ID3).

between privacy and data utility can be achieved, we quantify these effects by showing how accuracy and information gain vary when the anonymity level increases. For the sake of readability, in this section we report only the results when the classification accuracy is computed using the ID3 decision tree classifier and refer to Appendix A.1 for the results when the classification accuracy is computed using the SVM classifier.

Figures 13, 14, and 15 show the trade-off between accuracy and anonymity level for the Electricity, Adult, and Bank datasets, respectively. The x-axis reports the anonymity levels (in log scale), whereas the y-axis reports the best accuracy that can be achieved by applying the generalization rules that satisfy a given anonymity level. The baseline accuracy is obtained over the non-anonymized version of the datasets. Each vertical dashed line in the plots represents the maximum anonymity level that can be achieved using a given sampling percentage (5%, 10%, 20%, and 50%).

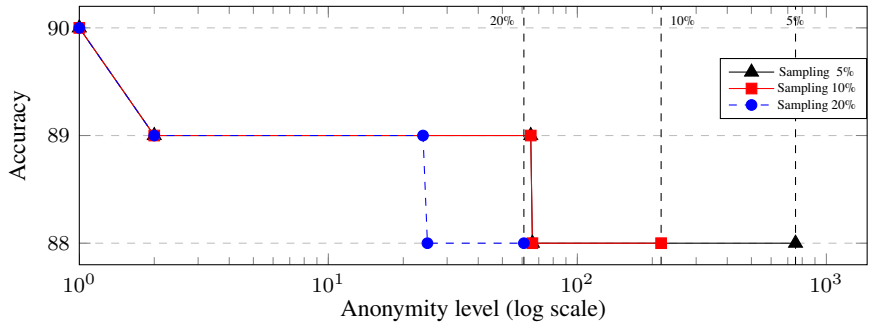

Figure 15: Trade-off between privacy and accuracy for the Bank dataset (ID3).

As expected, we can observe that, for all datasets, the accuracy decreases when the anonymity level increases, and that the highest anonymity level is achieved for the 5% sampling. The latter can be justified by the fact that the 5% sampling not only generates a larger number of generalization rules, but also these rules typically encompass attributes with a higher generalization level. Nevertheless, differences can be noticed in the maximum anonymity level that can be achieved using different sampling percentages for the three datasets. In particular, for the Adult dataset, the maximum anonymity level that can be achieved ranges from 78, for the 50% sampling, to 469, for the 5% sampling (cf. Figure 14). These differences are more notable for the Electricity dataset, where the maximum anonymity level ranges from 130, for the 50% sampling, to 7846, for the 5% sampling (cf. Figure 13). Also, for the Bank dataset the maximum anonymity level that can be achieved ranges from 61, for the 20% sampling, to 752, for the 5% sampling (cf. Figure 15). As previously mentioned, for the Bank dataset, the 50% sampling does not produce any generalization rule that satisfies the privacy requirement over k (i.e., $k \geq 2$).

It is worth noting that for the Electricity dataset all samplings preserve the baseline accuracy for $k \leq 4$. On the other hand, for the Adult dataset, although none of the extracted generalization rules guarantee the baseline accuracy, the loss in accuracy is limited between 5% and 10%. The smaller loss in accuracy for the Adult dataset could be due to the defined attribute taxonomies, which generally have a higher depth than the ones for the Electricity dataset (cf. Table 7). This difference in the attribute taxonomies for the two datasets also affects the number of cut-off points, which is smaller for the

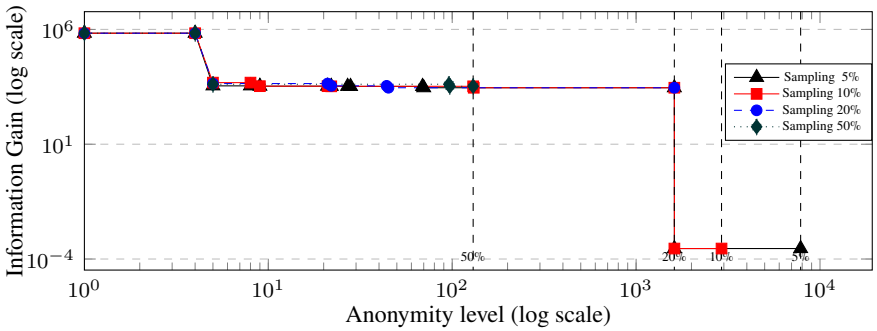

Figure 16: Trade-off between privacy and information gain for the Electricity dataset.

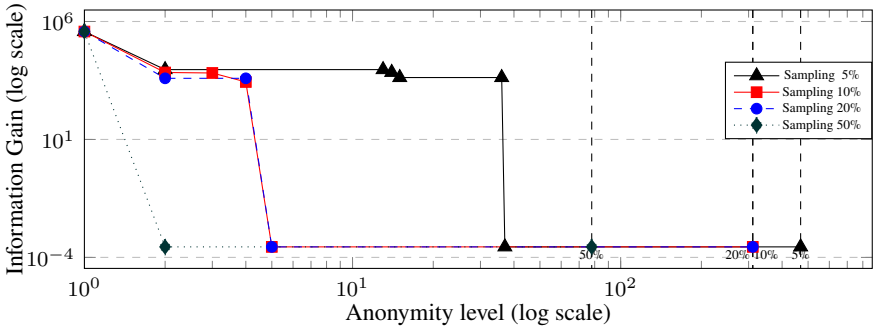

Figure 17: Trade-off between privacy and information gain for the Adult dataset.

Adult dataset. Finally, for the Bank dataset, the loss in accuracy is limited to the 1% w.r.t. the baseline for all sampling percentages.

Figures 16, 17, and 18 show the trade-off between information gain and anonymity level for the Electricity, Adult, and Bank datasets, respectively. Similarly to the results obtained for accuracy, information gain decreases when the anonymity level increases, and the highest anonymity level is achieved for the 5% sampling over all datasets. Moreover, for the Electricity dataset, all samplings preserve the baseline information gain for $k \leq 4$.

However, we can observe that, compared to accuracy, information gain decreases significantly faster, tending to zero at the increase of the anonymity level. In particular, for the Adult dataset, information gain is close to zero already with an anonymity level of 2 ($k = 2$) for the 50% sampling, and with an anonymity level of 5 ($k = 5$) for 10% and 20% samplings. For the 5% sampling, high information gain degrades to a value

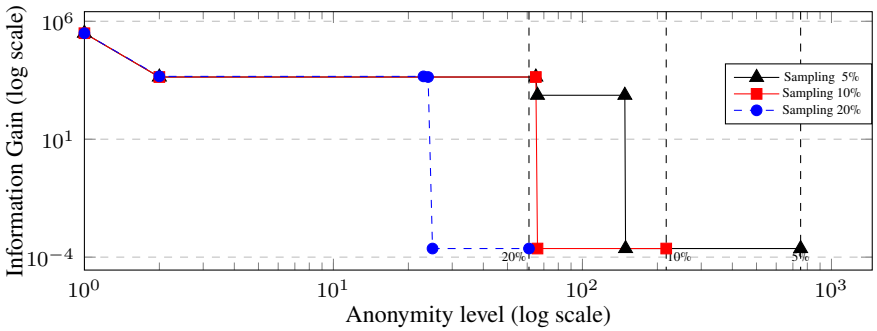

Figure 18: Trade-off between privacy and information gain for the Bank dataset.

close to zero for higher anonymity levels ($k \geq 27$). On the contrary, this effect is less prominent for the Electricity dataset, where a high information gain can be achieved for extremely high anonymity levels ($k \geq 1614$). This is mainly because the Electricity dataset is characterized by several numerical attributes for which many generalization levels were included in their taxonomy. Finally, for the Bank dataset, information gain is close to zero with an anonymity level of 25 ($k = 25$) for the 20% sampling, and with an anonymity level of 66 ($k = 66$) for 10%. For the 5% sampling, high information gain degrades to a value close to zero for higher anonymity levels ($k \geq 149$).

To evaluate the performances of our approach, we also compared it with the three anonymization algorithms presented in Section 6.1. For the comparison, we computed the trade-off between privacy and data utility measures that can be obtained by these algorithms with respect to the one that can be achieved by the generalization rules derived using our approach, as described in Section 6.1. For our approach, we used the generalization rules obtained using a sampling percentage of 5%, as these rules provide the best trade-off between privacy and data utility.

The results are reported in Figures 19, 20, and 21 for the Electricity, Adult, and Bank datasets, respectively, where each line represents the trade-off between anonymity level and classification accuracy (computed using the ID3 classifier) achieved by each of the considered algorithms. For the sake of readability, we refer to Appendix A.2 for the results on the comparative evaluation when the classification accuracy is computed using the SVM classifier. We can observe that our approach always outperforms the other

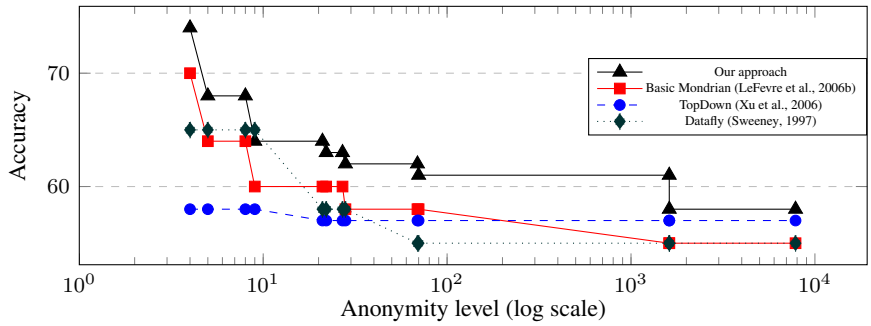

Figure 19: Comparison between anonymization techniques w.r.t. accuracy for the Electricity dataset. (ID3)

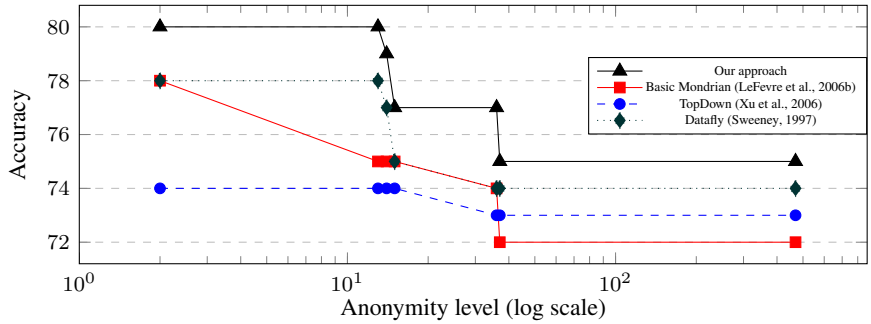

Figure 20: Comparison between anonymization techniques w.r.t. accuracy for the Adult dataset. (ID3)

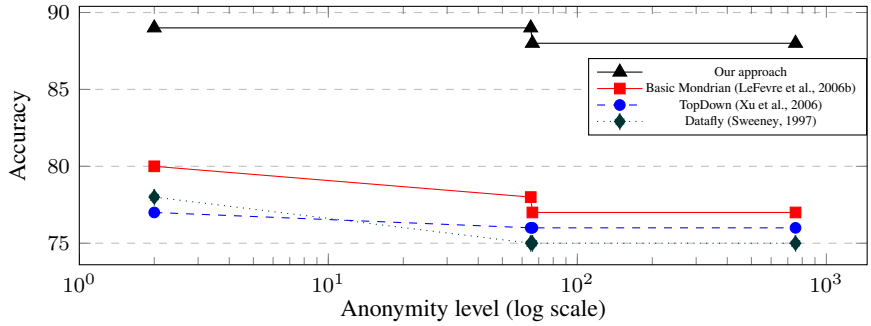

Figure 21: Comparison between anonymization techniques w.r.t. accuracy for the Bank dataset. (ID3)

anonymization algorithms by achieving the best accuracy for each considered value of k. As expected, also the other anonymization techniques show a decreasing trend when the anonymity level increases and obtain the smallest gap in accuracy w.r.t. the

proposed approach for values of k in the range between 30 and 40. Is worth noting that the TopDown algorithm (Xu et al., 2006) achieves very similar accuracy scores at the increasing of the value k, which almost always represent the worst performances w.r.t. scores obtained by the other algorithms. This is mainly due to the fact that, differently from the other algorithms, TopDown performs each generalization step by partitioning values over the complete dataset, resulting in poor performances when the dataset to be anonymized has high dimensionality, as in our experiments.

Figures 22, 23, and 24 report the trade-off between anonymity level and information gain achieved by each of the considered algorithms for the Electricity, Adult, and Bank datasets, respectively. We can observe that also in this case our approach always outperforms the other techniques, even if the difference is less noticeable. For the Electricity dataset, Basic Mondrian maintains an information gain comparable to our approach for anonymity levels lower than 80, whereas Datafly presents a degradation of information gain already when the anonymity level is equal to 20. On the other hand, these anonymization techniques exhibit a trend similar to our approach for the Adult and Bank datasets. Finally, results show that for all considered datasets, as said before, the TopDown algorithm obtains the worst performances compared to all other approaches.

An in-depth investigation of the obtained anonymization strategies shows that for low levels of k, our approach tends to not generalize one attribute compared to the other approaches that generalize each attribute for at least one level of generalization. When the level of k increases, our approach still returns generalization rules in which attributes are less generalized. This explains why our approach outperforms the compared ones in terms of classification accuracy and information gain. We also observed that Basic Mondrian tends to apply less generalization over numerical attributes w.r.t. categorical ones. Conversely, Datafly tends to apply less generalization over categorical attributes. As a result, Basic Mondrian and Datafly often produce anonymization strategies involving different attributes. Finally, the TopDown approach tends to generalize all attributes at high levels, except for k values less than 15 for which some categorical attributes are maintained more specialized.

RQ3: How much effort is required by a data owner to identify the generalization

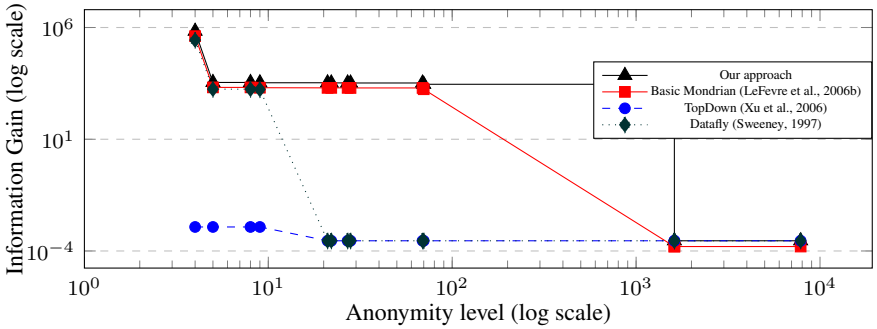

Figure 22: Comparison between anonymization techniques w.r.t. information gain for the Electricity dataset.

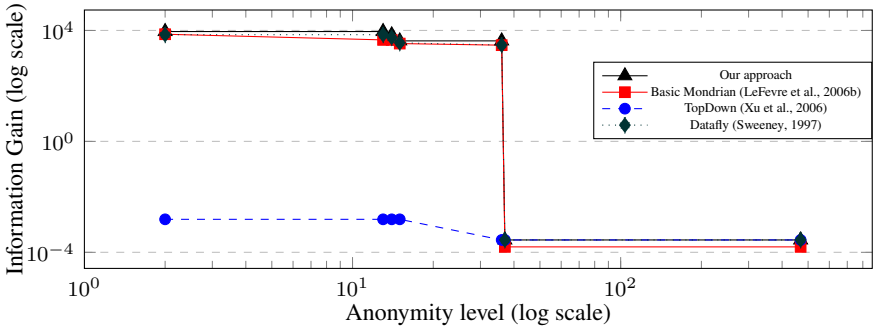

Figure 23: Comparison between anonymization techniques w.r.t. information gain for the Adult dataset.

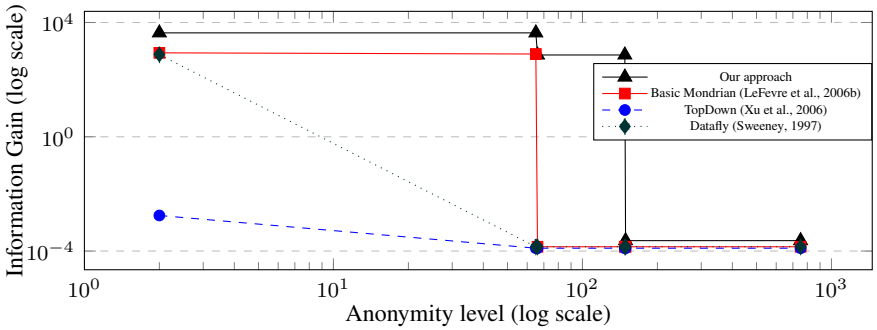

Figure 24: Comparison between anonymization techniques w.r.t. information gain for the Bank dataset.

rule to apply? A large number of generalization rules can be potentially returned by our approach, leaving the data owner with the burden to identify which generalization rule should be applied. To assist the data owner in this task, we employed an approach based

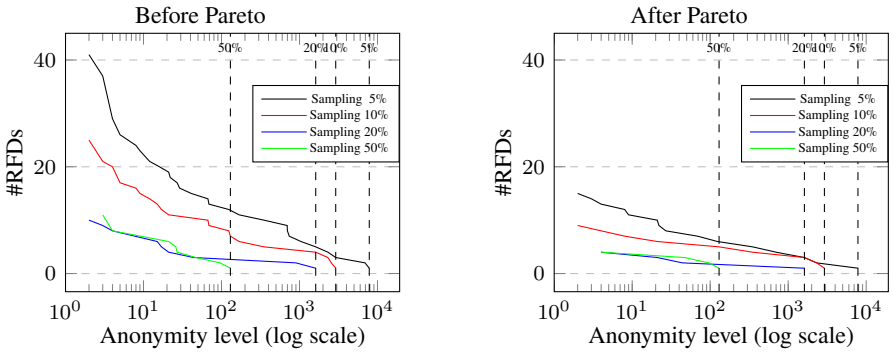

Figure 25: Variation of the number of RFDs at the increase of the anonymity level for the Electricity dataset.

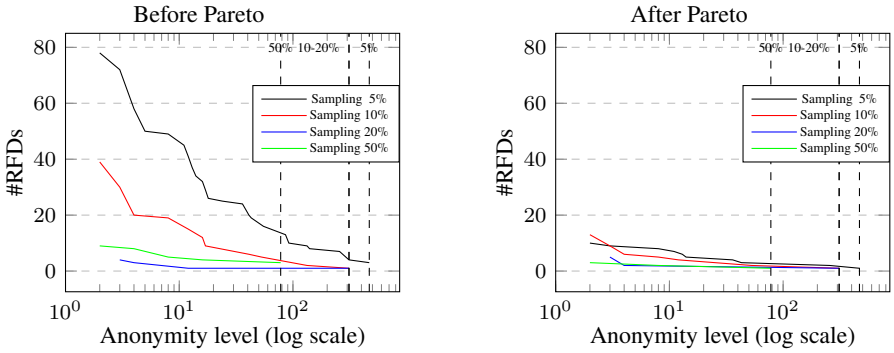

Figure 26: Variation of the number of RFDs at the increase of the anonymity level for the Adult dataset.

on Pareto-optimality to identify those rules providing a suitable trade-off between privacy and data utility (cf. Section 5.3). Next, we evaluate such approach and, in general, the effort required to a data owner to determine the generalization rule to apply, in terms of the number of rules returned by our approach. For the sake of simplicity, we only consider classification accuracy computed using ID3 in the application of Pareto-optimality.

Figures 25, 26, and 27 report the total number of RFDs obtained using our approach at the increase of the anonymity level for each sampling percentage, before (left plot) and after (right plot) the application of Pareto-optimality, for the Electricity, Adult, and Bank datasets, respectively. We can observe that the sampling percentage has a large impact on the number of rules: for all datasets, the use of lower sampling percentages typically results in a larger number of generalization rules. An in-depth analysis (not reported here for lack of space) shows that the number of combined generalization rules is also

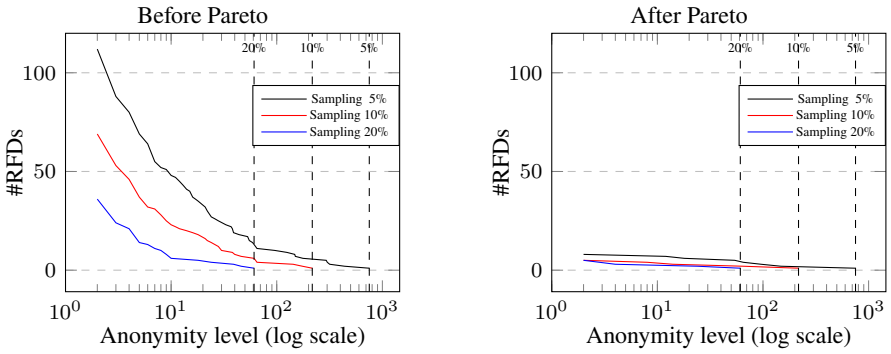

Figure 27: Variation of the number of RFDs at the increase of the anonymity level for the Bank dataset.

higher for lower sampling percentages. This is mainly due to the fact that generalization rules obtained for lower sampling percentages typically involve few attributes, yielding many possibilities to combine them with each other.

The results also show that the application of Pareto-optimality significantly reduces the number of generalization rules to be considered by data owners when anonymizing their datasets. For example, the use of Pareto-optimality yields a reduction of the total number of generalization rules, which achieve at least an anonymity level equal to 2, between 60% and 63% for the Electricity dataset, between 44% and 87% for the Adult dataset, and between 86% and 92% for the Bank dataset, where the largest reduction is obtained for the 5% sampling. When deriving rules using low sampling percentages, Pareto-optimality tends to preserve more combined rules than rules directly extracted from the data, whereas this consideration is reversed when the sampling percentage increases. An inspection of the generalization rules extracted over low sampling percentages showed that these rules typically involve fewer attributes, yielding a larger set of combined rules, among which we have rules that maintain the same anonymity level while providing higher data utility. On the contrary, since rules extracted for high sampling percentages typically contain many attributes, their combination tends to decrease the anonymity level.

The results discussed so far show the capability of Pareto-optimality to significantly reduce the space of candidate generalization rules, with respect to the use of RFDs only. Overall, the candidate generalization rules returned by our approach are in the order of

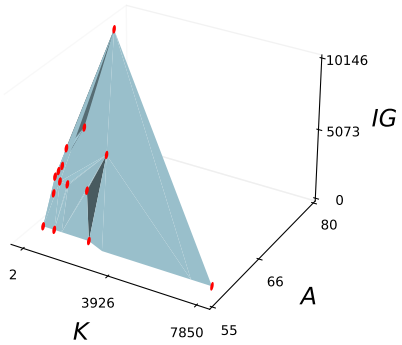

Figure 28: Pareto frontier for Electricity 5%.

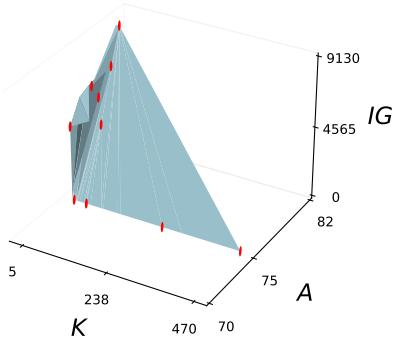

Figure 29: Pareto frontier for Adult 5%.

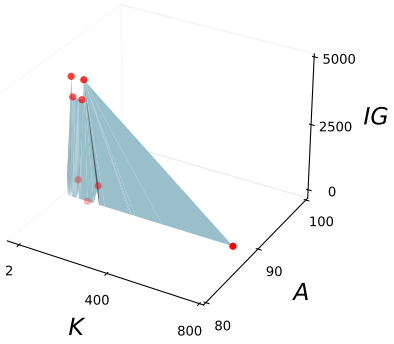

Figure 30: Pareto frontier for Bank 5%.

tens for the Electricity, Adult, and Bank datasets. Nevertheless, exploring these solutions to determine which generalization rule should be applied is, at this point, up to the data owner. Visualizing the Pareto frontier can assist data owners in obtaining an overview of the space of the rules providing a suitable trade-off between data utility and anonymity level and, thus, in effectively carrying out their analysis with respect to their privacy and data utility requirements. As an example, Figures 28, 29, and 30, show the Pareto frontier, represented by the red dots, for the Electricity, Adult, and Bank datasets, when a 5% sampling is used for extracting RFDs. Based on the Pareto frontier, the data owner can determine the expected accuracy and information gain for a given anonymity level and, possibly, ensuring stronger privacy guarantees at the cost of decreasing one of these data utility metrics.

7. Discussion

The proposed approach exploits the notion of RFD to support data owners in the anonymization of their dataset, aiming to let them achieve a given level of privacy while reducing the loss of data utility due to anonymization. In particular, the approach uses RFDs automatically extracted from the data to define possible generalization rules and combines them to achieve a higher data utility. Then, Pareto-optimality is employed to identify those generalization rules that provide a suitable trade-off between privacy and data utility. We evaluated the effectiveness of the approach by measuring: *(i)* the impact of combining RFDs on data utility (**RQ1**), *(ii)* the trade-off between anonymization and data utility (**RQ2**), and *(iii)* the effort required to a data owner to identify the generalization rule to apply (**RQ3**).

Next, we summarize the findings obtained by applying the proposed approach over the considered real-world datasets.

Using RFDs to define generalization rules. Our experiments showed that exploiting data correlations expressed in terms of RFDs provides an effective way to define anonymization strategies that preserve data utility. In particular, the use of roll-up dependencies, i.e., the type of RFDs considered in this work, allows accounting for the generalization levels in the extraction of RFDs, thus directly considering their impact on the attribute to be classified. In our experiments, we used the DOMINO algorithm for the discovery of this type of RFDs from the data (cf. Section 6.1). However, although this algorithm is capable of automatically extracting both the RFDs and their associated similarity thresholds, it only extracts RFDs that are valid on the entire dataset, which can be too restrictive and yield the extraction of few or no RFDs when applied to real-world datasets. On the other hand, algorithms from the literature capable of tolerating exceptions require the data owner to specify thresholds in input, nullifying the benefits of the proposed approach. Thus, to let DOMINO tolerate exceptions, we employed a sampling strategy on input data and applied DOMINO only on the sampled portion of the dataset, yielding a higher number of generalization rules. However, DOMINO returns only one generalization level per attribute, providing a full-domain generalization for that attribute. Although more fine-grained generalization strategies have been proposed in the literature (e.g., subtree generalization, sibling gener-

alization, cell generalization, multidimensional generalization), these strategies cannot be employed in our context because RFDs need to be validated throughout the entire value distribution. Nonetheless, we observed that exploiting data correlations, expressed in terms of roll-up dependencies, for the definition of anonymization strategies, helps preserving the data utility of anonymized datasets and outperforms other anonymization techniques (cf. Section 6.2). An interesting direction for future work is to explore the use of other types of RFDs (Breve et al., 2021) and study their impact on the anonymization process.

Construction of attribute taxonomies. The results of our experiments show that the effectiveness of extracted generalization rules depends on the quality of the attribute taxonomies defining the generalization levels. In particular, we observed that the use of an attribute taxonomy comprising several generalization levels typically leads to a higher number of generalization rules (e.g., leading to the potential of finding more suitable trade-offs between privacy and data utility), from which the data owners can choose for the anonymization of their datasets. In this work, we employed a generalization strategy based on VGH, as this approach better preserves data correlations compared to DGH. In particular, we employed a clustering approach to build the taxonomies of categorical attributes, by applying a binary splitting of data in order to maximize the depth of the taxonomy tree. Although this approach does not return the "genuine" number of clusters, we argue that this is not the main aim when building an attribute taxonomy. For instance, although the elbow method has the potential to provide a more accurate clusterization, its use would lead to attribute taxonomies with a lower depth, significantly limiting the space of possible anonymization strategies and, ultimately, resulting in fewer, coarse-grained strategies. Nevertheless, although the overall results of our approach are promising, an interesting direction for future work is to investigate the application of other data discretization techniques. Moreover, it would be also interesting to experimentally evaluate the impact of the depth of the taxonomy tree on the construction of generalization rules.

Combining generalization rules to improve data quality. We hypothesized that combining generalization rules helps reducing data utility loss in the anonymized dataset, as this approach has the potential of accounting for a larger number of attributes over which the dataset is anonymized (recall that the attributes which do not occur in the

applied generalization rule are removed from the dataset). Experiments confirmed our hypothesis and showed that combined generalization rules always provide a higher data utility than the rules directly extracted from the data (cf. the results for **RQ1** in Section 6.2), thus offering an effective way to minimize data utility loss.

Privacy and data utility metrics. To assess the trade-off between privacy and data utility offered by anonymization strategies we employed a number of metrics to measure data utility and privacy level of an (anonymized) dataset. In particular, our approach uses the k-anonymity model to measure the privacy level guaranteed by datasets, together with accuracy and information gain as data utility measures. While providing an effective measure for data anonymization, k-anonymity is susceptible to several attacks (see Section 3). At the same time, several metrics have been proposed to measure the data utility of anonymized datasets (e.g., precision, recall, F-score, entropy, and Gini index), where the choice of the data utility measure to be used depends on the purpose of the data publishing activities. An interesting direction for future work is to explore the application of other privacy and data utility measures and build a metric framework to assist data owners in determining the anonymization strategies providing the best trade-off between privacy and data utility requirements with respect to the use they intended when publishing data (Eom et al., 2020; Zigomitros et al., 2020). To this end, new classes of RFDs could be potentially exploited in order to properly map anonymization strategies in terms of data correlations. For instance, we are investigating the possibility to use conditional RFDs to provide an anonymization approach meeting the differential privacy strategy.

Selecting anonymization strategies to apply. Our experiments show that the number of obtained generalization rules remains manageable for being analyzed by a human (cf. the results for **RQ3** in Section 6.2). This suggests that our approach can be effective in practice to obtain usable indications of the strategies that can be applied for the anonymization of a dataset. We also show how plotting the generalization rules on the Pareto frontier provides a useful aid to data owners to visualize the achievable trade-off between privacy and data utility, letting them select the one that better fits their privacy and data utility requirements.

8. Conclusion

This work presents a decision-support framework for data anonymization with application to machine learning processes. Our framework relies on a novel approach that leverages the notion of RFD to exploit correlations in the data, aiming to achieve data anonymization while minimizing data utility loss. The approach extracts RFDs from the data to define possible generalization rules and combine them to derive anonymization strategies guaranteeing a higher data utility. Pareto-optimality is then employed to identify those generalization rules that provide optimal trade-offs between privacy and data utility. We evaluated the effectiveness of the proposed approach on three real-world datasets, by evaluating the impact of combining RFDs on data utility, the trade-off between anonymization and data utility, and the efforts required to a data owner to identify the generalization rule to apply. Results show that the proposed approach enables a data owner to identify effective anonymization strategies.

In the future, we plan to investigate some of the directions discussed in Section 7. In particular, we plan to investigate the application of other data utility and privacy metrics and study their impact on the trade-off between anonymization and data utility, as well as the applicability of our approach to other data sharing contexts (Feng et al., 2020). Moreover, we plan to investigate the use of other profiling metadata and types of RFDs to preserve data utility in the anonymization process.

References

Ashkouti, F., Khamforoosh, K., & Sheikhahmadi, A. (2021). DI-Mondrian: Distributed improved mondrian for satisfaction of the l-diversity privacy model using apache spark. *Information Sciences*, *546*, 1–24.

Bild, R., Kuhn, K. A., & Prasser, F. (2018). Safepub: A truthful data anonymization algorithm with strong privacy guarantees. *Proceedings on Privacy Enhancing Technologies*, *2018*, 67–87.

Breve, B., Caruccio, L., Cirillo, S., Deufemia, V., & Polese, G. (2021). Dependency visualization in data stream profiling. *Big Data Research*, *25*, 100240.

Calders, T., Ng, R. T., & Wijsen, J. (2002). Searching for dependencies at multiple abstraction levels. *ACM Transactions Database Systems*, *27*, 229–260.

Caruccio, L., Desiato, D., Polese, G., & Tortora, G. (2020a). GDPR compliant information confidentiality preservation in big data processing. *IEEE Access*, *8*, 205034–205050.

Caruccio, L., Deufemia, V., Naumann, F., & Polese, G. (2021). Discovering relaxed functional dependencies based on multi-attribute dominance. *IEEE Transactions on Knowledge and Data Engineering*, *33*, 3212–3228.

Caruccio, L., Deufemia, V., & Polese, G. (2020b). Mining relaxed functional dependencies from data. *Data Mining and Knowledge Discovery*, *34*, 443–477.

Caruccio, L., Piazza, O., Polese, G., & Tortora, G. (2020c). Secure IoT analytics for fast deterioration detection in emergency rooms. *IEEE Access*, *8*, 215343–215354.

Ding, H., Tian, Y., Peng, C., Zhang, Y., & Xiang, S. (2020). Inference attacks on genomic privacy with an improved HMM and an RCNN model for unrelated individuals. *Information Sciences*, *512*, 207–218.

Domingo-Ferrer, J., Sánchez, D., & Blanco-Justicia, A. (2021). The limits of differential privacy (and its misuse in data release and machine learning). *Communications of the ACM*, *64*, 33–35.

El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J.-P., Walker, M., Chowdhury, S., Vaillancourt, R., Roffey, T., & Bottomley, J. (2009). A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, *16*, 670–682.

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions Knowledge and Data Engineering*, *19*, 1–16.

Eom, C. S.-H., Lee, C. C., Lee, W., & Leung, C. K. (2020). Effective privacy preserving data publishing by vectorization. *Information Sciences*, *527*, 311–328.

Esmeel, T. K., Hasan, M. M., Kabir, M. N., & Firdaus, A. (2020). Balancing data utility versus information loss in data-privacy protection using k-anonymity. In *Conference on Systems, Process and Control* (pp. 158–161). IEEE.

Feng, J., Yang, L. T., Gati, N. J., Xie, X., & Gavuna, B. S. (2020). Privacy-preserving computation in cyber-physical-social systems: A survey of the state-of-the-art and perspectives. *Information Sciences*, *527*, 341–355.

Friedman, A., Wolff, R., & Schuster, A. (2008). Providing k-anonymity in data mining. *The VLDB Journal*, *17*, 789–804.

Fung, B. C., Wang, K., & Yu, P. S. (2005). Top-down specialization for information and privacy preservation. In *Proceedings of International Conference on Data engineering* (pp. 205–216). IEEE Computer Society.

Genga, L., Allodi, L., & Zannone, N. (2022). Association Rule Mining Meets Regression Analysis: An Automated Approach to Unveil Systematic Biases in Decision-Making Processes. *Journal of Cybersecurity and Privacy*, *2*, 191–219.

Goldstein, H., & Shlomo, N. (2020). A probabilistic procedure for anonymisation, for assessing the risk of re-identification and for the analysis of perturbed data sets. *Journal of Official Statistics*, *36*, 89–115.

Guarda, P., & Zannone, N. (2009). Towards the development of privacy-aware systems. *Information and Software Technology*, *51*, 337–350.

Hoogervorst, R., Zhang, Y., Tillem, G., Erkin, Z., & Verwer, S. (2019). Solving bin-packing problems under privacy preservation: Possibilities and trade-offs. *Information Sciences*, *500*, 203–216.

Kisilevich, S., Rokach, L., Elovici, Y., & Shapira, B. (2010). Efficient multidimensional suppression for k-anonymity. *IEEE Transactions on Knowledge and Data Engineering*, *22*, 334–347.

Koshley, D. K., Rani, S., & Halder, R. (2017). Towards generalization of privacy policy specification and property-based information leakage. In *International Conference on Information Systems Security* (pp. 68–87). Springer.

Last, M., Tassa, T., Zhmudyak, A., & Shmueli, E. (2014). Improving accuracy of classification models induced from anonymized datasets. *Information Sciences*, *256*, 138–161.

LeFevre, K., DeWitt, D., & Ramakrishnan, R. (2006a). Mondrian multidimensional k-anonymity. In *Proceedings of International Conference on Data Engineering* (pp. 25–25).

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006b). Workload-aware anonymization. In *Proceedings of SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 277–286). ACM.

Li, J., Kuang, X., Lin, S., Ma, X., & Tang, Y. (2020). Privacy preservation for machine learning training and classification based on homomorphic encryption schemes. *Information Sciences*, *526*, 166–179.

Lin, J.-L., & Wei, M.-C. (2008). An efficient clustering method for k-anonymization. In *Proceedings of International Workshop on Privacy and Anonymity in Information Society* (pp. 46–50). ACM.

Liu, C., Chen, S., Zhou, S., Guan, J., & Ma, Y. (2019). A novel privacy preserving method for data publication. *Information Sciences*, *501*, 421–435.

Lotov, A. V., & Miettinen, K. (2008). Visualizing the pareto frontier. In *Multiobjective optimization* (pp. 213–243). Springer.

Majeed, A., & Lee, S. (2021). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access*, *9*, 8512–8545.

Meijaard, Y. J., Cappers, B. C. M., Mengerink, J. G. M., & Zannone, N. (2020). Predictive analytics to prevent voice over IP international revenue sharing fraud. In *Data and Applications Security and Privacy XXXIV* (pp. 241–260). Springer volume 12122 of *LNCS*.

Ni, C., Cang, L. S., Gope, P., & Min, G. (2022). Data anonymization evaluation for big data and IoT environment. *Information Sciences*, *605*, 381–392.

Petchrompo, S., Coit, D. W., Brintrup, A., Wannakrairot, A., & Parlikad, A. K. (2022). A review of pareto pruning methods for multi-objective optimization. *Computers & Industrial Engineering*, *167*, 108022.

Pramanik, M. I., Lau, R. Y., Hossain, M. S., Rahoman, M. M., Debnath, S. K., Rashed, M. G., & Uddin, M. Z. (2021). Privacy preserving big data analytics: A critical analysis of state-of-the-art. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*, e1387.

Raj, A., & D'Souza, R. (2021). Scalable two-phase top-down specification for big data anonymization using apache pig. In *Advances in Artificial Intelligence and Data Engineering* (pp. 1009–1021). Springer.

Rathore, S., Sharma, P. K., Loia, V., Jeong, Y.-S., & Park, J. H. (2017). Social network security: Issues, challenges, threats, and solutions. *Information sciences*, *421*, 43–69.

Riva, G. M., Vasenev, A., & Zannone, N. (2020). SoK: engineering privacy-aware high-tech systems. In *Proceedings of International Conference on Availability, Reliability and Security* (pp. 19:1–19:10). ACM.

Samarati, P., & Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information. In *Symposium on Principles of Database Systems* (p. 188). ACM.

Šarčević, T., Molnar, D., & Mayer, R. (2020). An analysis of different notions of effectiveness in k-anonymity. In *International Conference on Privacy in Statistical Databases* (pp. 121–135). Springer.

Sheikhalishahi, M., & Zannone, N. (2020). On the comparison of classifiers' construction over private inputs. In *Proceedings of International Conference on Trust, Security and Privacy in Computing and Communications* (pp. 691–698). IEEE.

Song, J., Huang, L., He, Q., Gao, Y., Liu, X., & Li, Y. (2009). Preserving FDs in K-Anonymization by K-MSDs and Association Generalization. In *Proceedings of International Conference on Computational Intelligence and Security* (pp. 565–569). IEEE Computer Society.

Sweeney, L. (1997). Datafly: A system for providing anonymity in medical data. In *Database Security XI: Status and Prospects* (pp. 356–381). Chapman & Hall.

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*, 571–588.

Veeningen, M., Piepoli, A., & Zannone, N. (2014). Are on-line personae really unlinkable? In *Data Privacy Management* (pp. 369–379). Springer volume 8247 of *LNCS*.

Wang, R., Zhu, Y., Chang, C.-C., & Peng, Q. (2020). Privacy-preserving high-dimensional data publishing for classification. *Computers & Security*, *93*, 101785.

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., & Fu, A. W.-C. (2006). Utility-based anonymization using local recoding. In *Proceedings of SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–790). ACM.

Yan, Y., Herman, E. A., Mahmood, A., Feng, T., & Xie, P. (2021). A weighted k-member clustering algorithm for k-anonymization. *Computing*, *103*, 2251–2273.

Zigomitros, A., Casino, F., Solanas, A., & Patsakis, C. (2020). A survey on privacy properties for data publishing of relational data. *IEEE Access*, *8*, 51071–51099.

Appendix A. Additional Evaluations

Appendix A.1. Trade-off between privacy and accuracy (SVM)

To answer **RQ2**, we have also computed the trade-off between privacy and classification accuracy when SVM is used to compute the classification accuracy of the generalized dataset. Figures A.31, A.32 and A.33 report the results for the Electricity, Adult and Bank datasets, respectively. These results are in line with the ones obtained when ID3 is used to compute the classification accuracy of the generalized dataset, reported in Section 6.2.

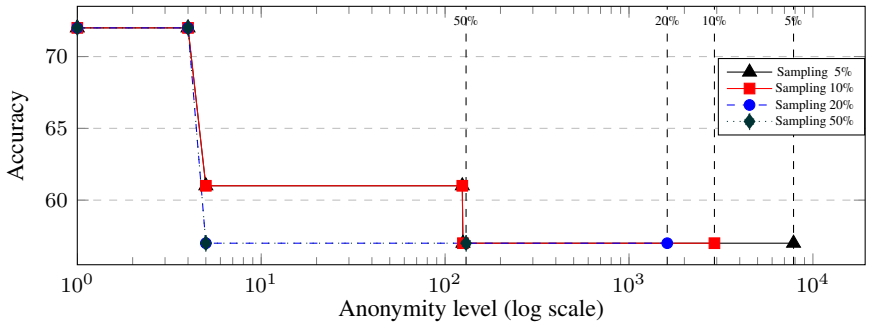

Figure A.31: Trade-off between privacy and accuracy for the Electricity dataset (SVM).

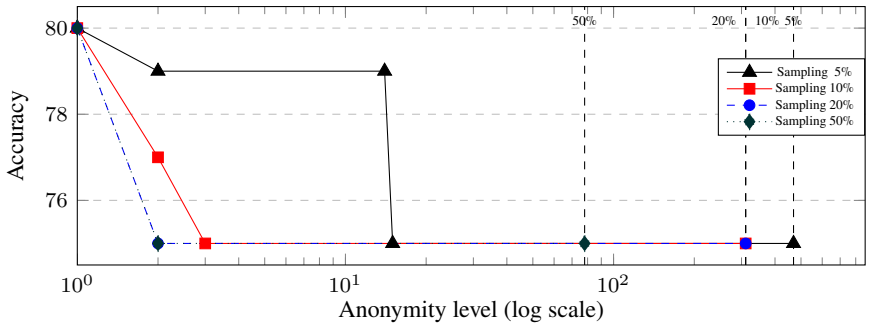

Figure A.32: Trade-off between privacy and accuracy for the Adult dataset (SVM).

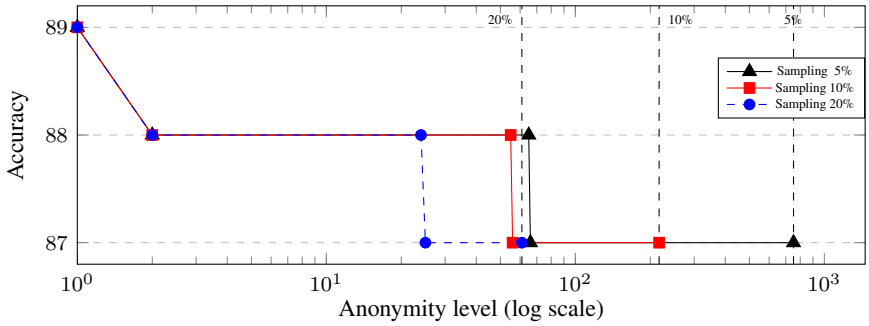

Figure A.33: Trade-off between privacy and accuracy for the Bank dataset (SVM).

Appendix A.2. Comparison between anonymization techniques w.r.t. accuracy (SVM)

We also performed a comparative evaluation between our approach and the anonymization techniques presented in Section 6.1 in the case classification accuracy computed using SVM was used as the metric for data utility. Figures A.34, A.35 and A.36 report the results for the Electricity, Adult and Bank datasets, respectively. The

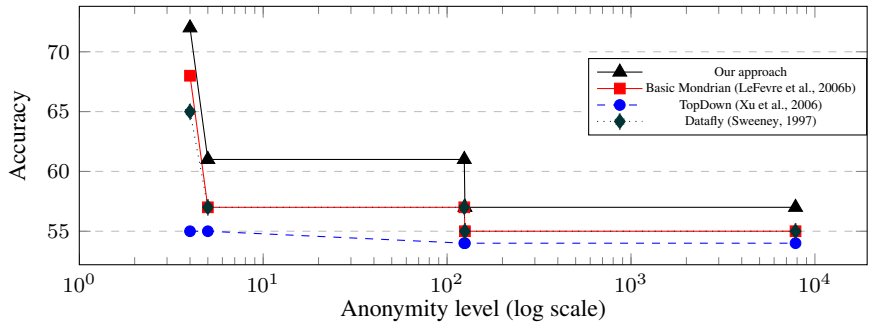

Figure A.34: Comparison between anonymization techniques w.r.t. accuracy for the Electricity dataset (SVM).

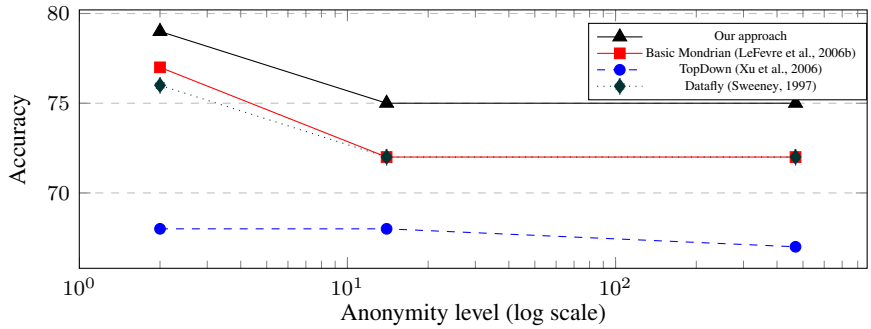

Figure A.35: Comparison between anonymization techniques w.r.t. accuracy for the Adult dataset (SVM).

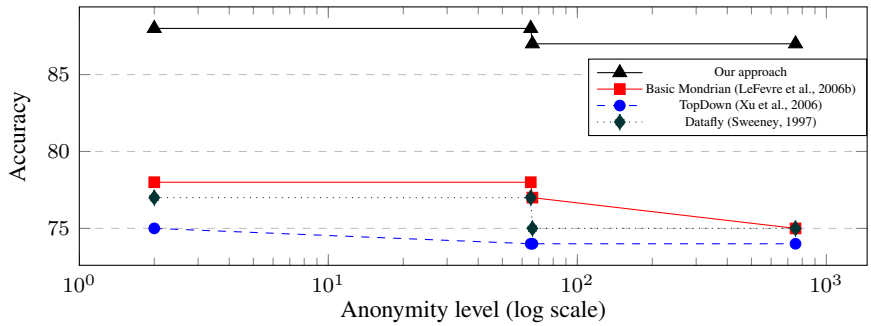

Figure A.36: Comparison between anonymization techniques w.r.t. accuracy for the Bank dataset (SVM).

results show that our approach overcomes the other considers anonymization techniques and that the improvement is even more noticeable than in the case where classification accuracy was computed using ID3 (Figures 19, 20 and 21).