Received: 6 August 2019 | Revised: 27 January 2020 | Accepted: 28 January 2020

DOI: 10.1002/spe.2812

RESEARCH ARTICLE

WILEY

Flexible data anonymization using ARX—Current status and challenges ahead

Fabian Prasser[1] | **Johanna Eicher**[2] | **Helmut Spengler**[2] | **Raffael Bild**[2] | **Klaus A. Kuhn**[2]

[1] Medical Informatics Lab, Berlin Institute of Health (BIH) and Charité Universitätsmedizin Berlin, Berlin, Germany

[2] School of Medicine, Institute of Medical Informatics, Statistics and Epidemiology, Technical University of Munich, Munich, Germany

Correspondence
Fabian Prasser, Medical Informatics Lab, Berlin Institute of Health (BIH), Charité Universitätsmedizin Berlin,
Charitéplatz 1, 10117 Berlin, Germany.
Email: fabian.prasser@charite.de

Summary

The race for innovation has turned into a race for data. Rapid developments of new technologies, especially in the field of artificial intelligence, are accompanied by new ways of accessing, integrating, and analyzing sensitive personal data. Examples include financial transactions, social network activities, location traces, and medical records. As a consequence, adequate and careful privacy management has become a significant challenge. New data protection regulations, for example in the EU and China, are direct responses to these developments. Data anonymization is an important building block of data protection concepts, as it allows to reduce privacy risks by altering data. The development of anonymization tools involves significant challenges, however. For instance, the effectiveness of different anonymization techniques depends on context, and thus tools need to support a large set of methods to ensure that the usefulness of data is not overly affected by risk-reducing transformations. In spite of these requirements, existing solutions typically only support a small set of methods. In this work, we describe how we have extended an open source data anonymization tool to support almost arbitrary combinations of a wide range of techniques in a scalable manner. We then review the spectrum of methods supported and discuss their compatibility within the novel framework. The results of an extensive experimental comparison show that our approach outperforms related solutions in terms of scalability and output data quality—while supporting a much broader range of techniques. Finally, we discuss practical experiences with ARX and present remaining issues and challenges ahead.

KEYWORDS

data anonymization, de-identification, privacy, security, software tools

Fabian Prasser and Johanna Eicher contributed equally to this study.

1 | INTRODUCTION

In the era of big data processing and artificial intelligence, the race for innovation has become a race for data. The spectrum of personal data that is collected electronically covers almost all aspects of our lives. Important examples of sensitive personal information include financial transactions, data about activities in social networks, location traces collected via mobile phone networks and medical records.[1] These data bear a tremendous potential for modern technologies to enable progress in a wide range of fields, such as economics, science, and public security. Possible applications vary from product recommender systems to health care decision support, computational criminology, and terrorism informatics.[2,3] Yet, in order to unlock this potential, data often need to be published, shared with third parties or reused for other purposes than the ones for which it was originally collected. This is a challenging task, as privacy concerns and restrictions imposed by national and international data protection laws, for example, the US Health Insurance Portability and Accountability Act (HIPAA),[4] the European General Data Protection Regulation (GDPR),[5] or the Chinese national standard on the protection of personal information,[6] need to be considered.

Data privacy can be addressed on multiple levels. The *Five Safes* framework describes one approach to conceptualize relevant safeguards in data management processes.[7] First, it can be important to ensure that *projects* are safe, which for example requires organizational measures that ensure that data use is appropriate. Second, it can be important to ensure that *people* working with the data are safe and trustworthy, for example by using strong authentication and authorization measures. Third, the *data* itself can be made safe, meaning that risks of re-identification are reduced to an acceptable minimum. Fourth, safe *settings* can be set up to reduce the risk of privacy breaches during processing, for example, by means of cryptographic protocols for secure multiparty computation.[8] Finally, the disclosure risk of *output* data can also be controlled to ensure that results do not leak sensitive personal information.

Data anonymization is an important building block for achieving safe input and output data. The basic idea is to transform data in such a way that privacy risks are reduced while the reduction of risks is balanced against a reduction of data utility.[9-13] Several high-profile re-identification attacks have demonstrated that this is a complex task requiring tool support.[14,15] For instance, simply removing directly identifying attributes, such as names or social security numbers, will typically not be enough to prevent privacy breaches.[16-18] More formal approaches are required, which employ mathematical and statistical models for quantifying risks and the impact of anonymization on data usefulness. Moreover, complex algorithms must be employed to balance both aspects in a scalable manner. We note that formal data anonymization is different from basic techniques of data masking or random data generation.[19] In this work, we focus on non-interactive microdata anonymization, which means that protected records are created from the records of an input dataset[11] and we do not cover interactive query anonymization, as, for example, implemented by PINQ[20] or Airavat.[21]

1.1 | Background

The obvious first step in any data anonymization process is to remove all direct identifiers of individuals.[11] The next—and far more challenging—step is to modify the dataset in a way that reduces the risk that an attacker is able to successfully link identified or identifiable individuals to one or multiple records or other sensitive information contained in the dataset.[17,22] In this process, the risk of such privacy breaches is quantified by mathematical or statistical *privacy models* (typically involving a threshold for what level of risk is deemed acceptable) and the utility of output data is quantified by a *utility model*. Figure 1 shows an abstract overview of an anonymization algorithm: A procedure searches through the space of all possible outputs, which is defined by one or multiple data *transformation models*, to find a solution which fulfills the risk thresholds specified for the privacy model and at the same time provides optimal output according to the utility model.

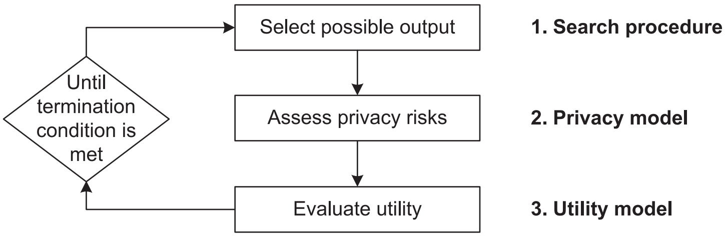
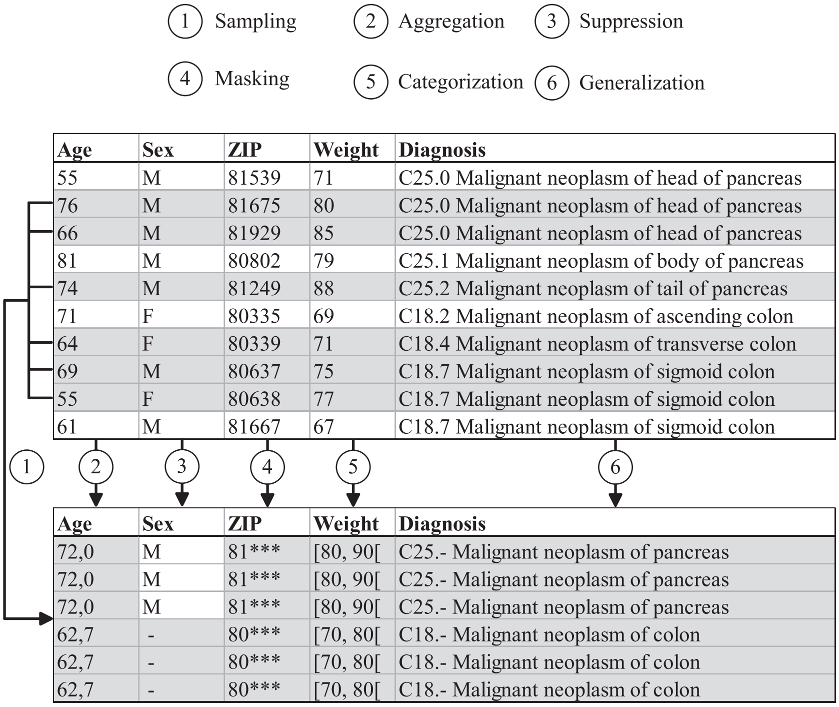


FIGURE 1 Abstract process implemented by data anonymization algorithms. A search procedure traverses the space of possible outputs while privacy models are used for assessing privacy risks and utility is evaluated using a utility model

FIGURE 2 Example of data transformation methods. A variety of transformation techniques typically need to be combined with each other to effectively anonymize a dataset



Age	Sex	ZIP	Weight	Diagnosis
55	M	81539	71	C25.0 Malignant neoplasm of head of pancreas
76	M	81675	80	C25.0 Malignant neoplasm of head of pancreas
66	M	81929	85	C25.0 Malignant neoplasm of head of pancreas
81	M	80802	79	C25.1 Malignant neoplasm of body of pancreas
74	M	81249	88	C25.2 Malignant neoplasm of tail of pancreas
71	F	80335	69	C18.2 Malignant neoplasm of ascending colon
64	F	80339	71	C18.4 Malignant neoplasm of transverse colon
69	M	80637	75	C18.7 Malignant neoplasm of sigmoid colon
55	F	80638	77	C18.7 Malignant neoplasm of sigmoid colon
61	M	81667	67	C18.7 Malignant neoplasm of sigmoid colon

Age	Sex	ZIP	Weight	Diagnosis
72,0	M	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
72,0	M	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
72,0	M	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
62,7	-	80***	[70, 80[	C18.- Malignant neoplasm of colon
62,7	-	80***	[70, 80[	C18.- Malignant neoplasm of colon
62,7	-	80***	[70, 80[	C18.- Malignant neoplasm of colon

An example using a combination of multiple transformation models is shown in Figure 2. As can be seen, a transformation might involve procedures such as taking a random sample of the records from the input dataset, aggregating numerical values and replacing them by their mean, suppressing individual values, masking parts of strings, categorizing numerical attributes, and generalizing categorical attributes. To reduce the risk of successful linkage attacks or the confidence an attacker might have in the correctness of linkage, these transformations may reduce the fidelity of data or introduce uncertainty by introducing noise.

Obviously, anonymization algorithms that support such complex transformation schemes cannot be implemented by simply searching the space of all potential output datasets for an optimal solution, as the search spaces are typically far too large. As a consequence, a wide range of heuristic strategies[23,24] and sophisticated clustering algorithms[25-29] have been developed. We emphasize, however, the importance of keeping the abstract model of data anonymization procedures implementing a specific combination of risk, utility, and transformation models in mind. For example, previous algorithms are typically only able to implement a specific combination of selected models, which severely limits their practical applicability.

As a consequence, the range of publicly available open source solutions is surprisingly small. It is well known that the effectiveness of different anonymization techniques highly depends on context, which includes the dimensionality, volume, and statistical properties of data.[12,30,31] Other important aspects that need to be considered include which types of applications or analyses the data are to be used for, whether the data will be released publicly or with additional access control and whether the data are tabular or have longitudinal or transactional characteristics. To ensure that anonymization software can be utilized for different application scenarios, different algorithms, and different methods for transforming data and quantifying reductions in usefulness must therefore be supported.[11] Moreover, many anonymization techniques involve significant computational complexity[32] which makes it challenging to implement them in a scalable manner.

1.2 | Related work

The current landscape of open source anonymization software basically consists of three types of solutions:

- First, there are tools originating from the computer science community (typically research prototypes), such as the *UTD Anonymization Toolbox*,[33] the *Cornell Anonymization Toolkit*,[34] *TIAMAT*,[35] *Anamnesia*[36] or *SECRETA*[37] and source code published as supplementary material to articles (eg, References 38 and 39). These solutions are able to automatically enforce privacy guarantees specified by users a priori. However, they usually only support a limited set of privacy models and focus on specific privacy and data transformation models.

- Second, there are tools originating from the statistics community, with *sdcMicro*[40] and μ-Argus[41] being the most prominent examples. These tools implement a more manual approach which enables them to support a wider variety of methods for measuring risks, transforming data, and analyzing the usefulness of output data. Privacy risks are typically quantified after transformations have been applied (a posteriori), which leads to an interactive anonymization process involving repeated and incremental transformations of a dataset.
- More recently, a wide range of commercial solutions has become available, often as a result to the requirements laid out in the GDPR. These closed-source tools focus on commercial markets. Typically, little is known about the underlying algorithms and they are not available for experimental evaluations and comparisons.

The ARX Data Anonymization Tool positions itself between these extremes with the aim of providing open software achieving a high degree of automation while at the same time providing supporting a wide range of techniques. In the past, various individual features and functionalities of ARX have been described in specific publications. Examples include anonymization methods based on statistical models,[42] game-theory,[43] differential privacy,[44] and an initial version of ARX's support for privacy-preserving data mining.[45] In addition, we have published two overview articles about ARX over the course of the years. The first article, which was published in 2014, focused on version 2.2.0 of ARX[46] while the second article, which was published in 2015, covered version 3.0.0 and introduced the application programming interface.[47] However, previous versions of ARX provided only limited support for complex data transformation models. We addressed this limitation in the work described in this article.

1.3 | Contributions

In the data anonymization space, it is of significant importance to distinguish between privacy models, transformation models, utility models, and anonymization algorithms. In general, a wide range of models needs to be supported to be able to address different real-world anonymization problems. However, prior algorithms typically only support a specific combination of methods. While previous versions of ARX already supported multiple privacy and utility models, only a small set of transformation techniques was available. In this work, we present a novel approach that has been implemented into the software to support (almost) arbitrary combinations of privacy and utility models with a wide range of data transformation techniques while preserving scalability.

We first present the core design principles that enable ARX to support multiple techniques for measuring privacy risks as well as output data utility while providing computational efficiency. Second, we present a novel approach for extending this design to significantly improve its genericity and flexibility regarding supported transformation methods. Next, we review the spectrum of methods supported and discuss their compatibility within the enhanced anonymization framework of the software. Then we present an extensive experimental comparison with related software. Our results show that ARX often outperforms other solutions in terms of scalability and—at the same time—output data quality, all while supporting a much broader spectrum of techniques. Finally, we discuss practical experiences with ARX, present remaining challenges and outline how we plan to address them in future work.

2 | FLEXIBLE DATA ANONYMIZATION IN ARX

2.1 | Basic design

At its core, ARX uses a highly efficient *globally-optimal search algorithm* for transforming data with *full-domain generalization* and *record suppression*. The transformation of attribute values is implemented through domain generalization hierarchies, which represent valid transformations that can be applied to individual-level values. Two examples are shown in Figure 3. Here, values of an attribute "age" are transformed into intervals with decreasing precision over increasing levels of generalization. Values of the attribute “sex“ can only be suppressed. We note that assigning generalization level zero to an attribute leaves its values unchanged. In ARX, generalization hierarchies can be specified by the user or created

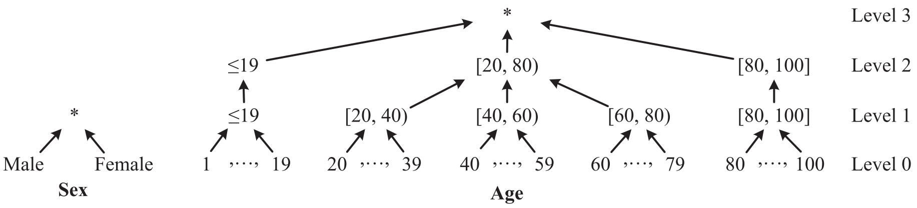
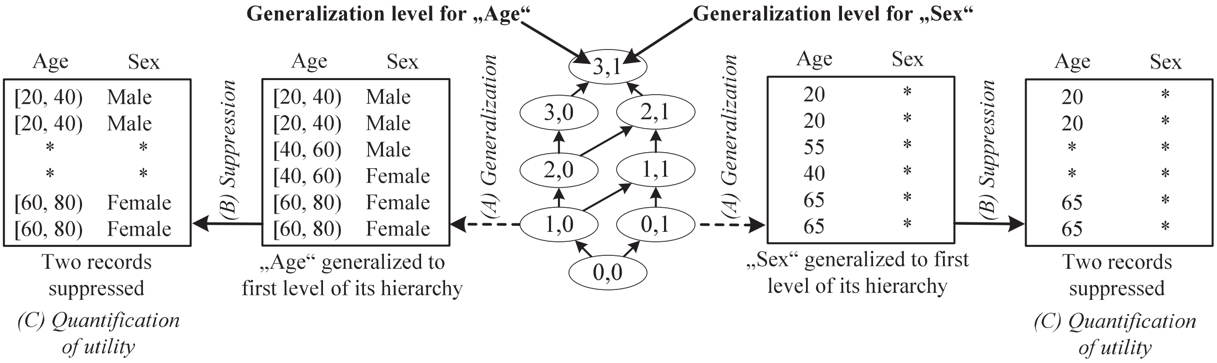


FIGURE 3 Examples of domain generalization hierarchies. The hierarchy to the left specifies possible generalizations of values for the attributes sex and the hierarchy to the right specifies generalizations for the attribute age

automatically for categorical and continuous attributes. In the latter case, this is accomplished by specifying functions for performing on-the-fly categorization of the value domain (eg, creating a grouping of heights or weights).

With full-domain generalization, all values of an attribute are transformed to the same generalization level in all records.[11] The set of all possible combinations of generalization levels for all attributes forms a *generalization lattice*, where each element is called a *generalization scheme*. The generalization lattice for the example hierarchies from Figure 3 together with an example dataset to which various generalization schemes have been applied is shown in Figure 4. Each node represents a single generalization scheme, which defines generalization levels for all attributes in the dataset. An arrow between two schemes indicates that they differ by exactly one generalization level. The transformation (0,0) represents the original dataset whereas the transformation (3,1) represents the dataset which results from maximal generalization. Referring to the overview from Figure 1, the optimal scheme from the lattice can be determined by going through all schemes one-by-one. In each step, the generalization scheme is applied (Step A), all records that do not adhere to the privacy requirements are suppressed (Step B) and the utility of the resulting output dataset is calculated (Step C). In the end, the optimal solution (ie, the output dataset with the highest utility) is returned. In the example, the privacy requirement is k-anonymity with $k = 2$, which means that each record must be indistinguishable from at least one other record (see Section 2.3 for more details on privacy models). In both output datasets created through generalization, the records three and four violate the privacy requirement and thus they have to be suppressed. After this, output data utility is measured to enable selecting the optimal solution. A simple utility model would be the number of cells that have not been suppressed (ie, that have a value different from "*"). In this case, the output dataset on the left would have a utility of eight while the output dataset on the right would have a utility of four. In practice, more sophisticated utility models are typically used, as is described in Section 2.3.

Anonymization algorithms using full-domain generalization are among the oldest approaches that have been developed in the field. Well-known examples include globally-optimal algorithms, such as Incognito[48] or *OLA*[49] and heuristic algorithms for data of higher dimensionality, such as *DataFly*.[23] ARX implements its own algorithms, *Flash* and *Lightning*, that significantly outperform prior approaches in the low-dimensional[50] as well as the high-dimensional setting,[12] respectively. Both algorithms make heavy use of ARX's compressed in-memory data representation[47] and advanced pruning-strategies.[31] Moreover, ARX employs a specialized record-suppression strategy that enables the software to suppress individual records for a specific generalization scheme, even when the privacy model used can only be evaluated


Generalization level for „Age"
Generalization level for „Sex"
3,1
3,0
2,1
2,0
1,1
1,0
0,1
0,0
(B) Suppression
(A) Generalization
(A) Generalization
(B) Suppression
Two records suppressed
„Age" generalized to first level of its hierarchy
„Sex" generalized to first level of its hierarchy
Two records suppressed
(C) Quantification of utility
(C) Quantification of utility

Age	Sex
[20, 40)	Male
[20, 40)	Male
*	*
*	*
[60, 80)	Female
[60, 80)	Female

Age	Sex
[20, 40)	Male
[20, 40)	Male
[40, 60)	Male
[40, 60)	Female
[60, 80)	Female
[60, 80)	Female

Age	Sex
20	*
20	*
55	*
40	*
65	*
65	*

Age	Sex
20	*
20	*
*	*
*	*
65	*
65	*

FIGURE 4 Example of full-domain generalization. It shows a generalization lattice and the results of applying two generalization schemes to a dataset followed by the suppression of records that do not adhere to the privacy requirements

for the overall dataset[13] (an example is *average re-identification risk*; further privacy models supported by ARX will be described in Section 2.3).

An important advantage of this class of search-based algorithms is that they are generic, which means that a wide range of privacy and utility models can be plugged into the system. The most important downside is that they are very inflexible in terms of supported transformation schemes and global generalization does not adjust well to the multidimensional distribution of data. This typically results in significant reductions to the quality of output data.

2.2 | Implementing advanced transformation methods

To overcome these limitations, we developed an approach for using the scalable basic algorithms of ARX as building blocks for implementing a wider range of more flexible transformation models. The basic idea is to iteratively apply the full-domain generalization algorithm to different subsets of an input dataset, resulting in different generalization schemes being used for the different subsets.

Horizontal partitioning strategy: What is needed for this purpose, is a partitioning strategy that reduces the overall degree of generalization applied. Such a strategy can be constructed using the basic algorithms provided by the software as follows. ARX supports the specification of a *limit on the number of suppressed records*. Moreover, records that have been suppressed may either be considered when calculating the overall utility of a transformed output dataset or they may be *ignored entirely* (ie, when calculating data utility, suppressed records are considered to be unmodified). To automatically partition and anonymize a dataset with n records, users only need to specify a *limit on the maximal number of partitions* (p) that can be created. From this limit, the minimal number of records in each partition can be derived ($n_p = \frac{n}{p}$). As is illustrated in Figure 5, ARX then sets the suppression limit accordingly and anonymizes the dataset while ignoring the impact of record suppression on data utility. This process is then iteratively repeated for the records that have been suppressed in the previous step until less than n_p suppressed records remain.

Vertical partitioning strategy (ie, grouping or clustering): To also support data aggregation, we developed a clustering strategy that is also based upon ARX's core algorithms as follows. The basic idea is to use the generalization scheme computed in each iteration not to transform the dataset, but to determine the clusters of values that need to become indistinguishable. In a subsequent *postprocessing step*, attributes of records within these clusters are then made indistinguishable by applying *aggregation functions* to the values from the input dataset of selected attributes (hence, vertical partitions) within each cluster (returning, eg, the mean or dynamic intervals). Vertical partitioning is performed automatically by ARX for attributes for which the user has configured aggregate functions. Further details on the horizontal as well as the vertical partitioning strategy, including pseudocode and examples, are provided in Appendices A and B.

As a result of the implementation of these two partitioning approaches, the software now supports combinations of four different types of transformation methods, which are listed in Table 1. With the new horizontal partitioning strategy, ARX can be configured to apply the same transformation scheme to all records in a dataset (*full-domain generalization*) or to apply different transformation schemes to different subsets of the records (*multi-dimensional generalization*).[51] The maximal number of transformations that may be used can be specified. ARX always guarantees that identical records in the input dataset will be transformed identically. With the new vertical partitioning strategy, hierarchies can also be

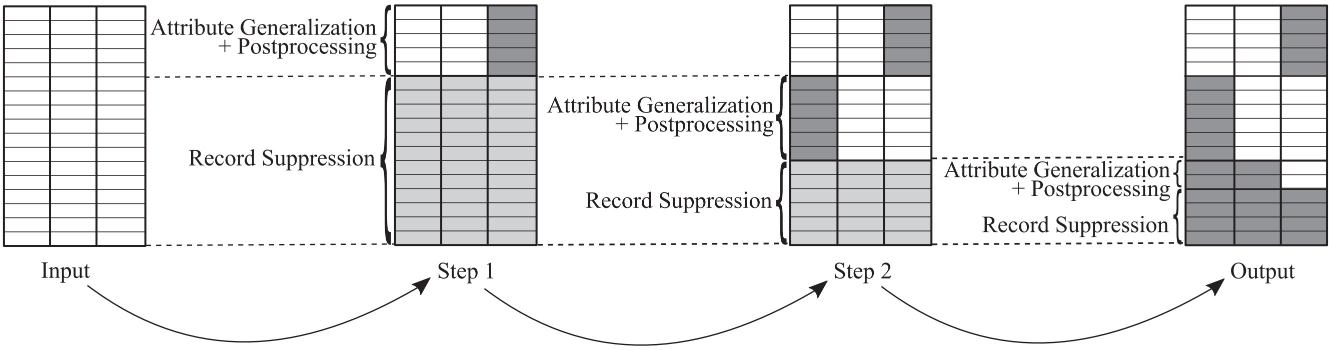

Attribute Generalization + Postprocessing
Record Suppression
Attribute Generalization + Postprocessing
Record Suppression
Attribute Generalization + Postprocessing
Record Suppression
Input
Step 1
Step 2
Output

FIGURE 5 Recursive application of the core transformation process for horizontal partitioning. ARX is able to apply full-domain generalization of attribute values followed by record suppression recursively to different subsets of a dataset

TABLE 1 Overview of transformation models supported by ARX

Transformation model		Type of attribute		Supported in prior versions
Type	**Implementation**	**Categorical**	**Numeric**	
Generalization	Multi-dimensional generalization	✓	✓	—
	Full-domain generalization	✓	✓	✓
	Top- and bottom-coding	—	✓	✓
	Categorization	—	✓	✓
Suppression	Cell-level	✓	✓	–
	Attribute-level	✓	✓	✓
	Record-level	✓	✓	✓
Sampling	Random	✓	✓	✓
	By query	✓	✓	✓
Microaggregation	Arithmetic and geometric mean	—	✓	–
	Median and mode	✓	✓	–
	Set	✓	✓	–
	Interval	—	✓	–

used to form clusters in which sets of attribute values can then be transformed into a common value by user-specified aggregation functions. Here, we have implemented support for the arithmetic and geometric mean, intervals, sets as well as median and mode for numerical attributes and sets, median and mode for categorical attributes. With the addition of the two partitioning schemes, we were able to extend ARX with six new transformation methods.

If transformation rules have been specified that only enable a suppression of values, a global transformation process will result in *attribute suppression*, while a local transformation process will result in a cell suppression scheme.[52] Independently of the specific transformation models specified, ARX may return a solution in which some of the records have been suppressed (typically only a tiny fraction). Generalization hierarchies can also be represented as functions, which can be used to perform on-the-fly categorization of continuous attributes during anonymization. Top- and bottom-coding can be implemented by using hierarchies that truncate values exceeding a user-specified range. We note that ARX contains multiple methods and wizards to automatically or semi-automatically construct hierarchies to apply these transformation methods. Finally, ARX supports drawing a sample from the input dataset. Methods that can be used for this purpose include matching a dataset against another dataset, querying the dataset using an expressive query language and random sampling. This can be used to relate a dataset to an underlying population table or to reduce privacy risks. Random sampling is further used to introduce randomness into the differential privacy mechanism supported by the software (see next section).

2.3 | Compatibility of methods

ARX supports a wide range of privacy and utility models. In this section, we discuss their compatibility with the horizontal and vertical partitioning strategies integrated into the software. The use of horizontal partitioning requires that privacy models can be enforced independently on different subsets of the data and that utility can be estimated by calculating it independently for different subsets. The use of vertical partitioning requires utility to be estimated accordingly.

ARX implements a wide range of privacy models that address different threats, such as *membership disclosure*, *attribute disclosure*, and *identity disclosure*.[11] Moreover, the privacy models address different assumptions about the intent and background knowledge of adversaries, such as the *prosecutor model*, *journalist model*, and the *marketer model*.[67] *Syntactic models* enforce restrictions on the structure of data, *statistical models* estimate risks in relationship to a larger underlying population or the success probabilities of attacks while *semantic models* have more direct relationships to mathematical notions of privacy. An overview of the models supported by ARX is shown in Table 2. Many models are supported in different variants.

An overview of the compatibility of the privacy models supported by ARX with different transformation techniques is provided in Table 3. Most incompatibilities are due to the way in which sampling is used in the software to implement privacy models. The method for taking a sample of the dataset is used to implement differential privacy, to specify population tables and to implement the horizontal partitioning algorithm. Consequently, privacy models that use sampling can currently not be combined with local transformation models. This is one of the shortcomings of the current development stage of the software that we plan to address in future work (see Section 6). Moreover, we note that in some cases it is also not obvious whether the privacy guarantees specified by a model also hold when data are partitioned. We have formally proven this for most models, but not yet for population uniqueness. For this reason, local transformation is currently deactivated for this model in the software. The current version of the differential privacy algorithm implemented in ARX is not compatible with the horizontal or vertical partitioning methods, as carefully randomized partitioning schemes would be required to ensure that privacy is not violated.[44]

In ARX, many different data utility models can be used as optimization functions. As is shown in Table 4, the software supports *general-purpose models*, which can be utilized when it is unknown in advance how output data will be used, and *special-purpose* (or *workload-aware*) models which quantify the usefulness of data for specific applications.[11] Utility models typically estimate data utility by quantifying the amount of information loss, for example, by measuring differences or similarities between the input and the output dataset. Models can roughly be classified as measuring information loss on the *attribute-level*, *cell-level*, *record-level*, or *dataset-level*. Typical examples for changes on these levels are differences in the distributions of attribute values, reductions in the granularity of data, differences in the distinguishability of records, or changes to overall scores, such as the accuracy of prediction models trained on the data. Notably, its strong support of methods for building and evaluating prediction models makes ARX also one of the most comprehensive tools available for privacy-preserving data mining.

Table 5 outlines the compatibility of the utility models with the transformation techniques supported by ARX. Incompatibilities resulting from vertical partitioning arise when using microaggregration operators. During the anonymization process, utility is only estimated for affected cells based on generalization. Incompatibilities resulting from horizontal partitioning are due to the fact that the frequencies of values in the input and output dataset are only known for the partition that is currently being processed. We emphasize that all utility models supported by ARX can still be used with all transformation methods. The quantification of utility reported by the system may be slightly off, however.

TABLE 2 Overview of privacy models supported by ARX

Privacy model	Type	Disclosure model	Attacker model	Population table
δ-Presence[53]	Syntactic/statistical	Membership	Journalist	✓
k-Anonymity[54]	Syntactic/statistical	Identity	Prosecutor	—
Average risk[42]	Syntactic/statistical	Identity	Marketer	—
k-Map[54]	Syntactic/statistical	Identity	Journalist	✓
k-Map with frequency estimators[55,56]	Statistical	Identity	Journalist	—
Population uniqueness[57-60]	Statistical	Identity	Marketer	—
ℓ-Diversity[61,62]	Syntactic/statistical	Attribute	Prosecutor	—
t-Closeness[63]	Syntactic/statistical	Attribute	Prosecutor	—
δ-Disclosure privacy[64]	Syntactic/statistical	Attribute	Prosecutor	—
β-Likeness[65]	Syntactic/statistical	Attribute	Prosecutor	—
Game-theoretic model (prosecutor)[43,66]	Semantic	Identity	Prosecutor	—
Game-theoretic model (journalist)[43,66]	Semantic	Identity	Journalist	✓
(ϵ, δ)-Differential privacy[44]	Semantic	All	All	—

TABLE 3 Compatibility matrix of privacy models and transformation models in ARX

Privacy model	Multi-dimensional generalization	Full-domain generalization	Top- and bottom-coding	Categorization	Cell-level suppression	Attribute-level suppression	Record-level suppression	Random sampling	Sampling by query	Microaggregation (all functions)
δ-Presence	—	✓	✓	✓	—	✓	✓	—	—	✓
k-Anonymity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Average risk	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
k-Map	—	✓	✓	✓	—	✓	✓	—	—	✓
k-Map with frequency estimator	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Population uniqueness	(✓)	✓	✓	✓	(✓)	✓	✓	✓	✓	✓
ℓ-Diversity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
t-Closeness	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
δ-Disclosure privacy	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
β-Likeness	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Game-theoretic model (prosecutor)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Game-theoretic model (journalist)	—	✓	✓	✓	—	✓	✓	—	—	✓
(ϵ, δ)-Differential privacy	—	✓	✓	✓	—	✓	✓	—	—	—

Brackets indicate functionality that is currently deactivated.

TABLE 4 Overview of utility models supported by ARX

Utility model	Type	Scope	Optimization	Visual analysis
Missings	Generic	Cell	✓	✓
Granularity/loss[68]	Generic	Cell	✓	✓
Precision[22]	Generic	Cell	✓	✓
Nonuniform entropy[69,70]	Generic	Attribute	✓	✓
Average distinguishability[51]	Generic	Record	✓	✓
Discernibility[32,49]	Generic	Record	✓	✓
Ambiguity[29]	Generic	Record	✓	✓
Record-level entropy[66]	Generic	Record	✓	✓
Sum of squared errors	Generic	Record	—	✓
Publisher benefit[43]	Special purpose	Record	✓	✓
Classification accuracy[45,68,71]	Special purpose	Datasets	✓	✓

3 | EXPERIMENTAL DESIGN

3.1 | Tools and algorithms

In previous work, we have already shown that ARX outperforms prior algorithms in terms of scalability and/or data utility when implementing global data transformation schemes.[12,44,50] In this article, we show that this is also true for local generalization schemes enabled by the horizontal and vertical partitioning strategies described in Section 2.1. For this purpose, we compare our tool to related software. Specifically, we focus on the following transformation schemes:

- *Multi-dimensional generalization*: Solves an anonymization problem by generalization. Values are transformed by replacing them with values from the provided hierarchies. Identical records will also be transformed identically.[51]
- *Local generalization*: Solves an anonymization problem by local generalization. Generalization can be performed without hierarchies, for example, by creating sets of values or intervals and identical records can be transformed differently.[51]

In our evaluation, we focus on tools that implement highly automated anonymization processes, analogously to ARX. Moreover, the privacy models implemented by these tools interpret datasets as population data describing one individual per record. When calculating frequencies, missing values are treated as an own category that only matches other missing values. As a baseline for evaluating the performance of multi-dimensional generalization, we used the well-known Mondrian algorithm[51] as implemented by the open source UTD Anonymization Toolbox (version 2012).[33] Following a top-down partitioning approach, Mondrian starts off with the trivial partition which contains all records of the dataset and keeps partitioning until no further partitions can be formed without violating the privacy requirements specified. As a baseline regarding local generalization, we used the authors' implementation of the algorithm proposed by Sánchez et al[38] (details can be found in the supplementary material of the article[72]). This approach interprets categorical attributes as integer-valued, clusters records based on their centroids and then forms groups of indistinguishable records in each cluster by replacing values with corresponding intervals. We note that the competing algorithms have specifically been designed for the respective data transformation schemes implemented, while ARX supports all of them in an integrated manner using a single algorithm. When implementing local generalization with ARX we employ an aggregate function to generalize values within clusters in the output dataset. Finally, we note that in all experiments attributes were either generalized by replacing them with values from a generalization hierarchy or by replacing them with intervals. In the experiments with local generalization, all attributes were interpreted as numbers, as this is the approach implemented by the algorithm by Sánchez et al[38] We note that this comes without loss of generality, as the dynamic forming of intervals over numbers representing categories is equivalent to the forming of sets containing the values encoded by the numbers contained in the interval.

TABLE 5 Compatibility matrix of utility models and transformation models in ARX

Utility model	Multi-dimensional generalization	Full-domain generalization	Top- and bottom-coding	Categorization	Cell-level suppression	Attribute-level suppression	Record-level suppression	Random sampling	Sampling by query	Microaggregation (all functions)
Missings	✓	✓	✓	✓	✓	✓	✓	✓	✓	(✓)
Granularity/loss	✓	✓	✓	✓	✓	✓	✓	✓	✓	(✓)
Precision	✓	✓	✓	✓	✓	✓	✓	✓	✓	(✓)
Nonuniform entropy	(✓)	✓	✓	✓	(✓)	✓	✓	✓	✓	(✓)
Average distinguishability	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Discernibility	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ambiguity	(✓)	✓	✓	✓	(✓)	✓	✓	✓	✓	(✓)
Record-level entropy	✓	✓	✓	✓	✓	✓	✓	✓	✓	(✓)
Sum of squared errors	✓	✓	✓	✓	✓	✓	✓	✓	✓	(✓)
Publisher benefit	✓	✓	✓	✓	✓	✓	✓	✓	✓	(✓)
Classification accuracy	(✓)	✓	✓	✓	(✓)	✓	✓	✓	✓	(✓)

Brackets indicate that output data utility can only be approximated.

3.2 Datasets

We used six real-world datasets, most of which have already been utilized for evaluating previous work on data anonymization: (1) *US Census*, an excerpt from the 1994 census database, which serves as the de facto standard for evaluations of anonymization algorithms, (2) *Competition*, introduced in the KDD data mining competition in 1998, (3) *Crash Statistics*, NHTSA crash statistics from their Fatality Analysis Reporting System, (4) *Time Use Survey*, data from the American Time Use Survey, (5) *Health Interviews*, results from the Integrated Health Interview Series, and (6) *Community Survey*, responses to the American Community Survey, an ongoing survey conducted by the US Census bureau on demographic, social, and economic characteristics from randomly selected people living in the United States. The sizes of the datasets on disk range between 2.52 MB (US Census) and 107.56 MB (Health Interviews). To ensure compatibility with the algorithm by Sánchez et al and to simplify the distribution of data together with the source code used in the experiments, we performed dictionary encoding on all categorical attributes.[38] The datasets have different characteristics, which are listed in Table 6:

- *Dimensionality*, that is, the number of attributes. With 30 attributes the Community Survey dataset is of *high* dimensionality. All other datasets contain either eight or nine attributes and are of *medium* dimensionality.
- *Volume*, that is, the number of records. The datasets US Census, Competition, and Community Survey contain between 30 162 and 68 725 records and are of *low* volume. With a size of 100 937 and 539 253 records, respectively, the datasets Crash Statistics and Time Use Survey are of *medium* volume while Health Interviews is a *high* volume dataset comprising 1 193 504 records.
- *Identifiability*, which is based on the number of unique patterns of attribute values contained in the data. Each such combination has the potential to identify individuals in the dataset and thus the number of patterns can be used for risk estimation.[73] We have calculated the number of these so-called *minimal sample uniques* (MSUs) using the SUDA2 algorithm provided by sdcMicro, modified to print the number of MSUs identified. In addition to the overall number of MSUs per dataset we report the average number of MSUs per cell. The more MSUs the higher is the risk of re-identification and therefore identifiability. While Community Survey and Competition are of *high* and *medium* identifiability, respectively, all other datasets are of *low* identifiability.

For reference, further properties of the datasets are presented in Appendix C. As a rule of thumb, higher dimensionality, volume, or identifiability can be expected to increase execution times and decrease output data utility. We note that some of the evaluation datasets are samples from a larger population, which have been created using complex sampling designs. These aspects could be used to derive more exact risk estimates during data anonymization. The tools considered in our evaluation, however, only implement privacy models that make worst-case assumptions and they do not implement mechanisms for considering complex data structures. Hence, we did not include special variables, such as strata variables or sampling weights, into our evaluation datasets and assumed that all datasets describe one individual per record. We emphasize that this is a frequent assumption in many domains, for example, in medical research, which is also often made when comparing automated data anonymization procedures. Moreover, this approach allows for a fair comparison between the tools covered in this section. We will discuss its limitations in Section 6.

TABLE 6 Overview of the datasets and their complexity in terms of dimensionality, volume as well as identifiability

	Dimensionality		**Volume**		**Identifiability**		
Dataset	**Attributes**	**Complexity**	**Records**	**Complexity**	**MSUs**	**MSUs/cell**	**Complexity**
US Census	9	Medium	30 162	Low	62 809	0.23	Low
Competition	8	Medium	63 441	Low	791 475	1.56	Medium
Crash Statistics	8	Medium	100 937	Medium	175 271	0.22	Low
Time Use Survey	9	Medium	539 253	Medium	321 406	0.07	Low
Health Interviews	9	Medium	1 193 504	High	2 888 220	0.27	Low
Community Survey	30	High	68 725	Low	15 708 409	7.6	High

As a rule of thumb, higher degrees of complexity can be expected to increase execution times and decrease output data utility.
Abbreviation: MSU, minimal sample unique.

3.3 | Configuration and setup

When selecting privacy models to use in the evaluation, the individual methods supported by the tools and algorithms listed above must be considered. ARX supports all models presented in Table 2. The Mondrian algorithm from the UTD Anonymization Toolbox, however, only supports k-anonymity and the algorithm by Sánchez et al supports only k-anonymity and t-closeness. We therefore decided to present results for the k-anonymity privacy model, because it is the only model supported by all competitors. We are well aware of the weaknesses of k-anonymity and emphasize that ARX also supports multiple more recent models, as described in the previous sections.

Common parameterizations for k-anonymity used in the literature are $k = 2, 3, 5, 10$, which equal thresholds for prosecutor re-identification risk of 50%, 33%, 20% and 10%. We vary this parameter and the number of attributes that must be protected from linkage (the so-called *quasi-identifiers [QI]*) to study the effect of different risk thresholds and data dimensionality on output data utility as well as scalability. We note that increasing the number of quasi-identifiers is a simple way to significantly increase the number of anonymization problems studied and that it can also provide more detailed insights into the effect of data dimensionality on the algorithms' performance. When varying k we included all quasi-identifiers and when varying the number of quasi-identifiers we used $k = 5$. We evaluated the scalability of the different solutions by measuring elapsed real *execution times*. In order to obtain stable results, we calculated averages over multiple runs of each algorithm (the number of runs for each experiment was determined based on the stability of runtime measurements). For practical reasons, we introduced a hard time limit of 3600 seconds and runs that did not terminate within that time frame were cancelled.

To evaluate output data utility, we used a simple and intuitive general-purpose model, called *Granularity*, which measures the value-level precision of output data.[68] For reference, a formal definition is presented in Appendix D. All utility measurements have been normalized into a range of [0, 1], such that 100% represents an unmodified dataset, and 0% represents the a dataset from which all information has been removed. We note that general-purpose utility models have limitations regarding their ability to capture the usefulness of output data for specific application scenarios, for example, regression modeling. However, at the extreme points of general-purpose utility estimates, such models also provide a good indicator for the usefulness of data for specific applications. For example, a general-purpose utility of close to 100% indicates that almost no changes have been made to the data, which typically also corresponds with usefulness for performing concrete analyses. Analogously, a general-purpose utility of 50%, for example, indicates that significant changes have been made to the data, which typically also significantly impacts usefulness for specific applications.

The experiments were performed on a desktop machine equipped with a quad-core 3.2 GHz Intel Core i5 CPU running a 64-bit Windows NT kernel and a 32-bit JVM (1.8.0_202_x86). All tools tested leveraged only one of the CPU cores of the benchmark system. Our implementation of the benchmark and the datasets used are available online.[74]

4 | RESULTS OF EXPERIMENTS AND DISCUSSION

4.1 | Comparison with the UTD Anonymization Toolbox

Figure 6 shows the execution times measured when performing multidimensional generalization. We note that in some settings we were not able to process the datasets Crash Statistics, Health Interviews, and Community Survey with the implementation of the Mondrian algorithm from the UTD Anonymization Toolbox, since the application terminated with an error. In the figure, this is indicated by "x". Regarding the other setups, it can be seen that higher volume or identifiability resulted in higher execution time (Time Use Survey, Health Interviews). With ARX execution times increased with increasing privacy parameters, while with the UTD Anonymization Toolbox execution times decreased with increasing privacy protection. For processing the high-dimensional dataset, ARX needed not more than 1000 seconds, while all other datasets could be processed in not more than 100 seconds. The UTD Anonymization Toolbox needed significantly more time in all cases.

Figure 7 shows the data utility measured in the experiments. It can be seen that in all cases ARX returned output data to which almost no modifications had been made. The results show that data utility slightly decreased when the degree of privacy protection increased. When using the UTD Anonymization Toolbox, however, significant changes were made to input data, resulting in utility estimates as low as 60% in some cases. It can further be seen that with ARX output data utility decreased monotonically when the number of quasi-identifiers increased. This trend could generally also be observed for the UTD Anonymization Toolbox. Some instabilities could however be observed when processing the Time

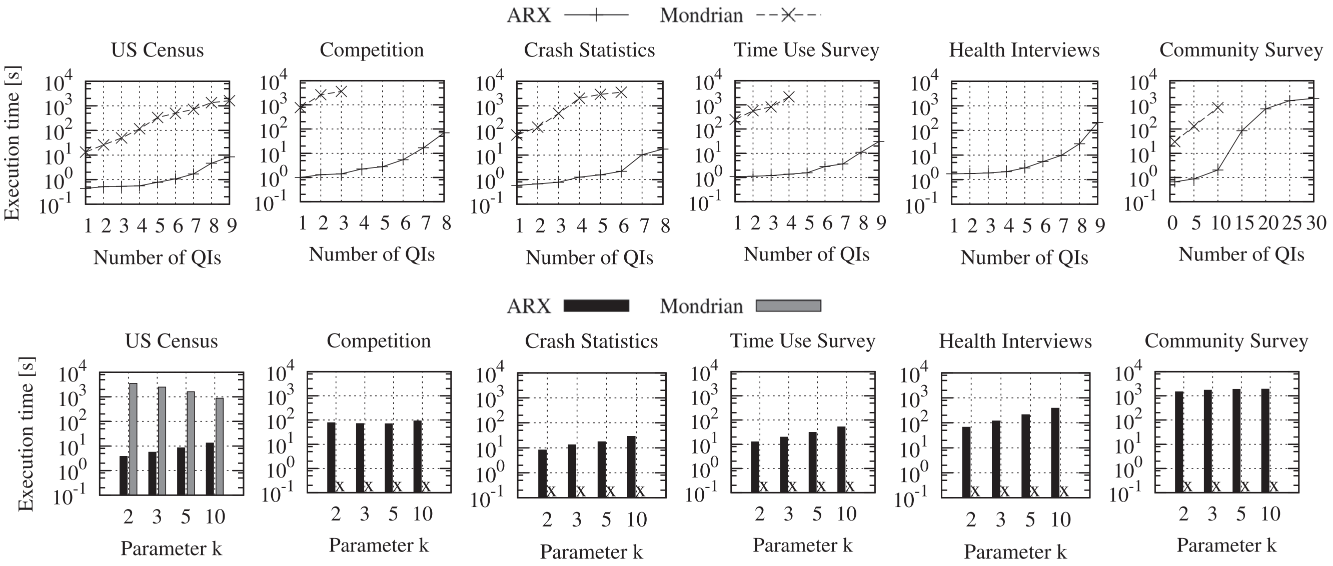


FIGURE 6 Comparison of the execution times of ARX and the UTD Anonymization Toolbox. Note: On *y*-axes, logarithmic scaling was used. In the bar charts, the symbol "x" indicates a missing data point due to the algorithm exceeding the time limit or terminating with an error

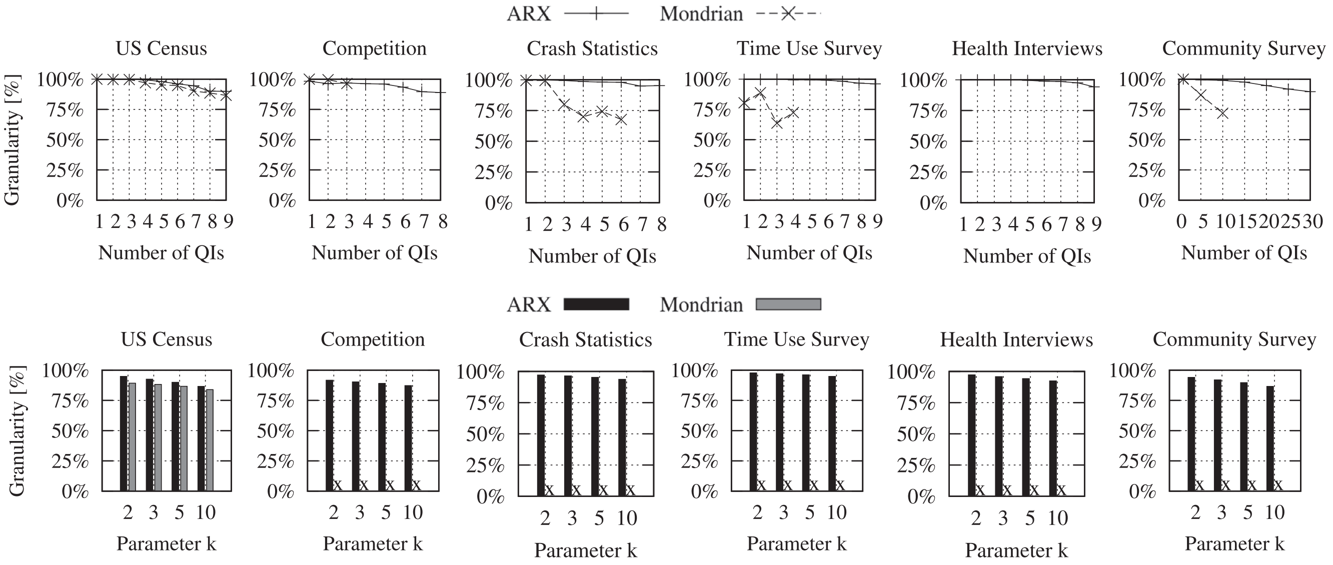


FIGURE 7 Comparison of the data utility obtained using ARX and the UTD Anonymization Toolbox. Note: On *y*-axes, logarithmic scaling was used. In the bar charts, the symbol "x" indicates a missing data point due to the algorithm exceeding the time limit or terminating with an error

Use Survey and Crash Statistics datasets. We note that the fact that ARX is able to significantly reduce the uniqueness of records in the Community Survey dataset with only about 10% reduction in data granularity implies that correlations exist between many of the attributes of the dataset. We conclude that, in our experiments with multidimensional generalization, the algorithm implemented by ARX exhibited significantly higher scalability than the Mondrian algorithm implemented by the UTD Anonymization Toolbox and at the same time provided higher degrees of output data utility.

4.2 | Comparison with the algorithm by Sánchez et al

Figure 8 shows the execution times measured when comparing ARX to the local generalization algorithm by Sánchez et al. It can be seen that ARX performed comparable to this algorithm in low-dimensional settings, while ARX performed

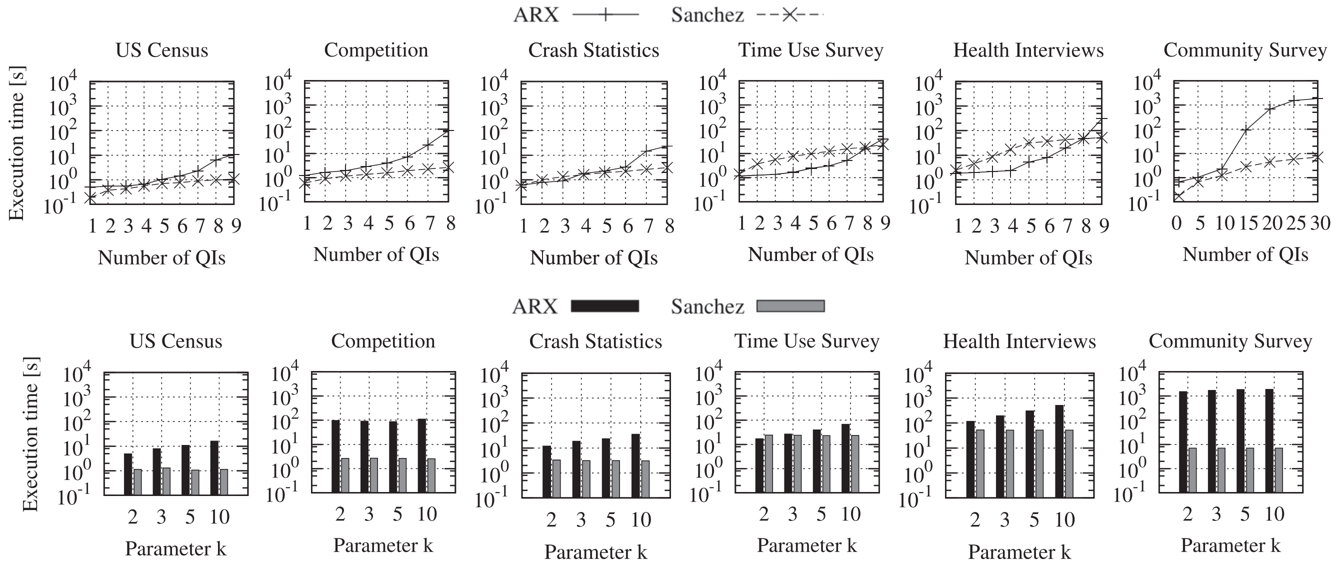


FIGURE 8 Comparison of the execution times of ARX and the algorithm by Sánchez et al. Note: On *y*-axes, logarithmic scaling was used

worse in high-dimensional settings. It can further be seen that ARX was less scalable when processing the dataset with high identifiability. The results also show that ARX outperformed the algorithm by Sánchez et al when processing the Time Use Survey dataset, which has the second highest volume of the datasets considered, but very low identifiability. This can be explained by the fact that the optimizations implemented into ARX are particularly effective when identifiability is low[31,50] and that the runtime complexity of the approach by Sánchez et al is dominated by sorting the dataset. This also implies that the performance of the algorithm by Sánchez et al mostly depends on the number of records contained in a dataset, which is also reflected by our results.

Figure 9 shows the data utility measured in the experiments. It can be seen that in all cases ARX returned output datasets to which almost no modifications had been made. When using the approach by Sánchez et al, however, significant changes were made to input data, again resulting in utility estimates as low as 60% in some cases. This is remarkable, as the transformation method implemented by ARX is less flexible, as it always guarantees that identical records in input data are transformed to identical records in the output dataset. Again, data utility decreased monotonically when risk

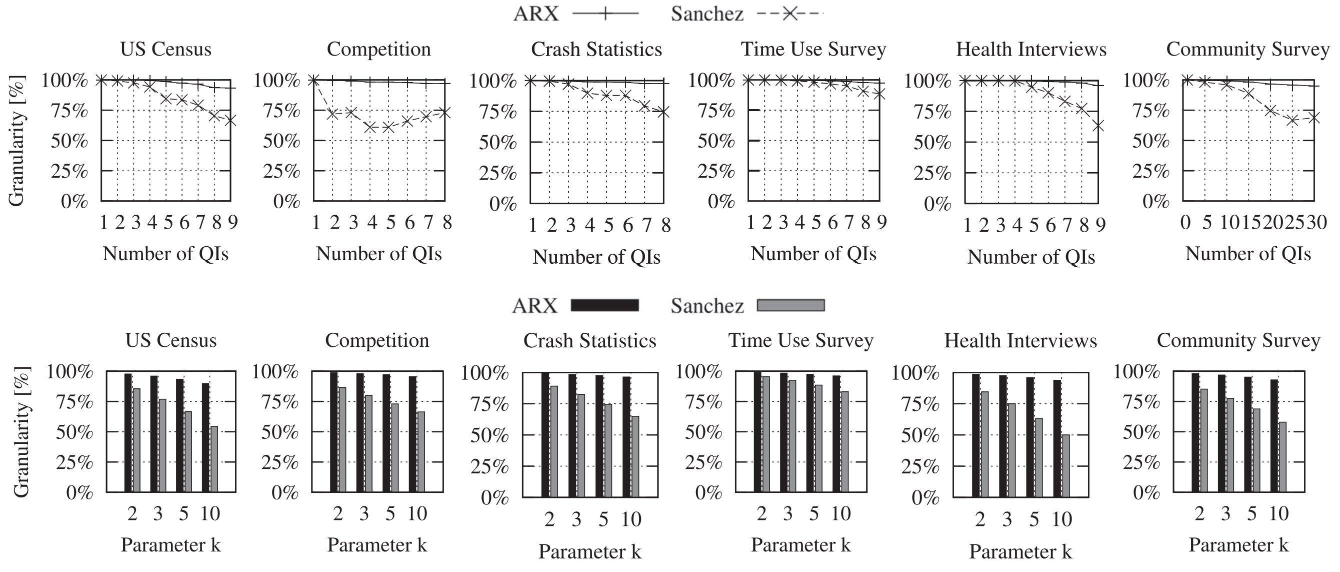


FIGURE 9 Comparison of the data utility obtained using ARX and the algorithm by Sánchez et al

thresholds increased but the effect was much stronger when using the algorithm by Sánchez et al. We conclude that, in our experiments with local generalization, the approach by Sánchez et al exhibited higher scalability but the algorithm implemented by ARX provided higher degrees of output data utility.

5 | SUMMARY AND PRACTICAL EXPERIENCES

In this article, we have presented an overview of the current development state of the ARX Data Anonymization Tool. We have described recent extensions to the software that enable users to utilize a wide variety of data transformation methods that were previously only supported by specific tools or algorithms. We have presented the results of an extensive experimental evaluation which has shown that ARX often outperforms related software. The development of methods that make ARX so flexible was not only a major methodological challenge, but it also contributed significantly to the success of the software. To illustrate this, we briefly present some examples of official policies and guidelines, research projects, and data publishing activities that have made use of the software.

On the level of guidelines, ARX has for example been mentioned by the European Medicines Agency as a solution for implementing quantitative risk assessments when implementing Policy 0070[75] on the sharing of data from clinical trials.[76] Moreover, ARX has been listed in a guideline by the European Union Agency for Network and Information Security (ENISA) on methods for implementing privacy and data protection by design principles.[77] Another guideline mentioning ARX has been released by the UK Anonymization Network (UKAN), which is an organization promoting and advising on best practices in data anonymization.[78] The document has also been adapted by the Office of the Australian Information Commissioner.[79] ARX has also been covered in a comprehensive analysis of anonymization tools released by the Directorate for Research, Studies, Evaluation and Statistics of the central administration of the French Ministry of Social Affairs and Health.[80] It has further been mentioned in a report on requirements and implementation options for anonymization services by the Finnish Ministry of Transport and Communications,[81] in a guide to data anonymization by the Personal Data Protection Commission of Singapore,[82] a security standard released by the Polish Ministry of Digitalization,[83] a report on data anonymization by the Dutch Ministry of Justice and Security[84] as well as a report by the Korean Ministry of Science and ICT.[85] These examples show the importance of open source anonymization tools for supervisory authorities.

On the level of scientific data management, various institutions have included ARX into software collections. Examples include the Finnish Social Science Data Archive,[86] EPFL,[87] the University of Guelph,[88] the University of Munich,[89] and the University of Kassel.[90] The graphical frontend of ARX is also frequently used in training courses. For example, the Korea Internet & Security Agency (KISA) and the TMF e.V., the umbrella organization for networked medical research in Germany, offer regular training programs.[91,92] ARX has further been covered in many handbooks on the topic.[40,93,94] Recently ARX has also been integrated into the big data processing framework KNIME,[95] and one of ARX's core algorithms has been selected to form the backbone of SAP HANA Data Anonymization.[96]

ARX has also been used in several research projects, mostly through its application programming interface. One important area is research on privacy-preserving big data analytics platforms. For example, Costa et al described a platform for big data management in the telecommunication sector that offers privacy-enhancing features through ARX.[97] Kim et al proposed a distributed analytics platform based on ensemble learning for healthcare data. They used data anonymized with ARX as a baseline in experimental comparisons.[98] A second line of research using ARX focuses on trust and access control. An interesting example is the article by Armando et al, which describes a risk-aware access control framework for information disclosure. The presented prototype includes a risk mitigation module which uses adaptive anonymization operations implemented on top of ARX.[99] Another example is the work by Jiang et al, in which game-theoretic methods have been used to develop a credibility model in cooperative networks and ARX has been included in the evaluation.[100] The development of new data anonymization methods is another area in which ARX is frequently utilized. An interesting example is the work by Stammler et al, who have used ARX to implement and evaluate an enhanced variant of the ℓ-diversity privacy model which uses an asymptotically unbiased estimator for the Shannon entropy.[62] Li et al have proposed and implemented a graph-based framework for privacy-preserving data publishing, which they evaluated by comparing the output of their framework with the output of ARX.[101] Moreover, Xu et al proposed a contract-based approach to handle the trade-off between privacy and utility, which has been implemented on top of ARX.[102] Finally, Park et al have developed a data synthesis mechanism based on Generative Adversarial Networks and they used ARX as a baseline technology in their evaluation.[103]

ARX has also been used to anonymize datasets for public and private dissamination. However, since official guidelines unfortunately do not usually provide specific instructions on how data needs to be anonymized, only little information is publicly available on practical applications. One example is the work by Kuzilek et al describing the Open University Learning Analytics Dataset, which is a representative subset of student data collected at the Open University. The data were anonymized using ARX in a process that has been certified by the Open Data Institute.[104] As another example, Ursin et al have used ARX to assess and manage the re-identification risk of a large dataset from the Norwegian Cervical Cancer Screening Program.[105]

6 | LIMITATIONS AND CHALLENGES AHEAD

ARX's flexibility and a relatively intuitive and easy-to-use interface are key factors that contributed to the software's success. However, we emphasize that the methods implemented by the software are complex from a mathematical and statistical perspective and, as a consequence, anonymization in real-world settings can usually only be carried out by experts. For example, risks models must be selected according to the context, and risks must then be reduced precisely to an extent that ensures that the data are reliably protected. In addition, one must be aware of the intended use of a dataset to ensure that the anonymized data remain useful. Moreover, there are several limitations that we plan to address in future work.

First, ARX does currently not support many methods provided by data anonymization tools from the statistics community, such as sdcMicro. Important examples include methods for considering the effect of complex sampling designs on re-identification risks when anonymizing data or different means of calculating the frequency of records for risk estimation. The main reason why we have not yet implemented such techniques is that they are not frequently used in the area of health data privacy, which is our primary application domain. However, we plan to extend the software in this direction in future work. Another area of future work is to compare ARX to other algorithms using transformation methods not studied in this article. Important examples include cell suppression and methods for aggregating continuous variables in such a way that they remain continuous and keep their scale of measure (eg, replacing them by the mean within clusters).

Second, while ARX is much more scalable than many other solutions in the field, it can currently only be used to anonymize medium-sized datasets with up to a few million rows and up to 50 quasi-identifying variables. Nowadays, data controllers often need to deal with gigabytes and terabytes of data, with in some cases hundreds of attributes that need to be protected. One example is large sparse datasets used for creating machine learning models.[106] Due to its high degree of automatization, ARX is well suited for implementing anonymization operators that can then be distributed amongst a large number of nodes to enable or speed up the processing of very large datasets. However, integrating appropriate strategies for distributing data and processing the results obtained from different nodes is challenging. This is particularly true for ARX, where parallelization strategies must be implemented carefully to not impact the flexibility of the software.

Another important area of future development is to improve ARX's abilities to process high-dimensional data along two axes. First, we plan to improve the scalability of finding solutions to anonymization problems with a high number of quasi-identifiers by implementing an alternative to the algorithm currently used by the software. The genetic algorithm proposed by Wan et al for anonymizing genetic data is an interesting candidate[66] but integrating it is challenging due to the different context in which it was proposed. Second, we plan to improve the utility of output data in high-dimensional settings by implementing methods to better handle complex inter-attribute relationships.[107] One possible solution to this problem is to treat the data as transactional, that is, set-valued, and to employ specific privacy models, such as k^m-anonymity,[108] which is implemented by Anamnesia[36] and SECRETA.[37]

Another related area with significant challenges ahead is to improve the compatibility of the privacy models implemented with local transformation methods. In this context, we plan to redesign our sampling subsystem to ensure that also models that rely on sampling can be used when applying local data transformation. Moreover, for some models, for example, those that use statistical models to estimate population uniqueness, it is not yet clear whether their privacy guarantees hold in the local transformation context. We plan to formally analyze this and to develop variants that can be used with local transformations if needed. These steps are also needed to guarantee privacy-preservation in the distributed settings outlined above. Finally, our differential privacy algorithm[44] needs to be extended with differentially private procedures which incorporate the horizontal and vertical partitioning methods.

We further plan to include more methods from the area of statistical disclosure control and further less formal transformation methods into ARX. An important example is the SUDA2 algorithm,[73] which can be used to implement various types of risk analysis and anonymization and which is frequently used in the statistics community. Furthermore, we plan to include data masking techniques (eg, for random data generation and shuffling) into the software to enable users to combine formal methods of data anonymization with a wide range of such basic transformation operations.

Finally, we are working on many features to make the software even more reliable and usable in practical applications. For example, we have recently integrated the data anonymization operations provided by ARX into the ETL environment

Pentaho Data Integration,[109] and we are working to integrate them into further environments, such as Talend Open Studio.[110] A significant challenge in this process is to not negatively impact the flexibility of the software.

ACKNOWLEDGEMENTS

The authors would like to thank the various contributors to ARX. They are particularly grateful to the contributions of Florian Kohlmayer, Fabian Stahnke, and Martin Waltl, who have contributed to the development and evaluation of the horizontal and vertical partitioning methods described in this article. Moreover, we would like to thank the anonymous reviewers for their comments and suggestions which have greatly improved this manuscript.

ORCID

Fabian Prasser https://orcid.org/0000-0003-3172-3095
Johanna Eicher https://orcid.org/0000-0003-4871-0282
Raffael Bild https://orcid.org/0000-0002-7398-5598

REFERENCES

1. Article 29 Data Protection Working Party. Opinion 05/2014 on anonymisation techniques; 2014. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.
2. Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans Knowl Data Eng*. 2005;17(6):734-749.
3. Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: from big data to big impact. *MIS Q*. 2012;36(4):1165-1188.
4. US Department of Health and Human Services Office for Civil Rights. Standards for privacy of individually identifiable health information: final rule. *Fed Reg*. 2002;67(157):53181.
5. Council of the European Union, European Parliament. Regulation (EU) 2016/679 of the European parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46. *Off J Eur Union*. 2016;59(L119):1-88.
6. Standardization Administration of China. GB/T 35273-2017 information technology – personal information security specification; 2018.
7. Desai T, Ritchie F, Welpton R. *Five Safes: Designing Data Access for Research*. Bristol: University of the West of England; 2016.
8. Cramer R, Damgård IB, Nielsen JB. *Secure Multiparty Computation*. Cambridge: Cambridge University Press; 2015.
9. Domingo-Ferrer J, Torra V. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Min Knowl Disc*. 2005;11(2):195-212.
10. Xia W, Heatherly R, Ding X, Li J, Malin BA. R-U policy frontiers for health data de-identification. *J Am Med Inform Ass*. 2015;22(5):1029-1041.
11. Fung BCM, Wang K, Fu AW-C, Yu PS. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Boca Raton, FL: CRC Press; 2010.
12. Prasser F, Bild R, Eicher J, Spengler H, Kohlmayer F, Kuhn KA. Lightning: utility-driven anonymization of high-dimensional data. *Trans Data Priv*. 2016;9(2):161-185.
13. Prasser F, Kohlmayer F, Spengler H, Kuhn KA. A scalable and pragmatic method for the safe sharing of high-quality health data. *IEEE J Biomed Health Inform*. 2018;22(2):611-622.
14. Leoni D. Non-interactive differential privacy: a survey. Paper presented at: Proceedings of the 1st International Workshop on Open Data; 2012:40-52.
15. El Emam K, Jonker E, Arbuckle L, Malin BA. A systematic review of re-identification attacks on health data. *PLoS One*. 2011;6(12):e28071.
16. Duncan GT, Elliot M, Salazar-González J-J. *Statistical Confidentiality: Principles and Practice*. New York, NY: Springer; 2011.
17. Narayanan Arvind, Shmatikov Vitaly. Robust de-anonymization of large sparse datasets. *Symposium on Security and Privacy*. Piscataway, NJ: IEEE; 2008;111–125.
18. Sweeney L. Computational disclosure control - a primer on data privacy protection (PhD thesis). Massachusetts Institute of Technology; 2001.
19. Hundepool A, Domingo-Ferrer J, Franconi L, et al. *Statistical Disclosure Control*. Hoboken, NJ: John Wiley & Sons; 2012.
20. McSherry FD Privacy integrated queries: an extensible platform for privacy-preserving data analysis. Paper presented at: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data; 2009:19-30.
21. Roy I, Setty STV, Kilzer A, Shmatikov V, Witchel E. Airavat: security and privacy for MapReduce. *NSDI*. 2010;10:297-312.
22. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *Int J Uncert Fuzz Knowl-Based Syst*. 2002;10(05):571-588.
23. Sweeney L. Datafly: a system for providing anonymity in medical data. *Database Security XI*. Boston, MA: Springer; 1998:356-381.
24. Babu KS, Reddy N, Kumar N, Elliot M, Jena SK. Achieving k-anonymity using improved greedy heuristics for very large relational databases. *Trans Data Priv*. 2013;6(1):1-17.
25. Soria-Comas J, Domingo-Ferrer J, Sánchez D, Martinez S. t-closeness through microaggregation: strict privacy with enhanced utility preservation. *IEEE Trans Knowl Data Eng*. 2015;27(11):3098-3110.

26. Byun JW, Kamra A, Bertino E, Li N. Efficient k-anonymization using clustering techniques. Paper presented at: Proceedings of the International Conference on Database Systems for Advanced Applications; 2007:188-200.
27. Gionis A, Mazza A, Tassa T. k-Anonymization revisited. Paper presented at: Proceedings of the 24th International Conference on Data Engineering; 2008:744-753.
28. Goldberger J, Tassa T. Efficient anonymizations with enhanced utility. Paper presented at: Proceedings of the International Conference on Data Mining; 2009:106-113.
29. Nergiz ME, Clifton C. Thoughts on k-anonymization. Paper presented at: Proceedings of the 22nd International Conference on Data Engineering; 2006:96.
30. Zakerzadeh H, Aggarwal CC, Barker K. Managing dimensionality in data privacy anonymization. *Knowl Inf Syst*. 2016;49(1):341-373.
31. Prasser F, Kohlmayer F, Kuhn KA. Efficient and effective pruning strategies for health data de-identification. *BMC Med Inform Decis Mak*. 2016;16(1):49.
32. Bayardo RJ, Agrawal R. Data privacy through optimal k-anonymization. Paper presented at: Proceedings of the 21st International Conference on Data Engineering; 2005:217-228.
33. UT Dallas Data Security and Privacy Lab. UTD anonymization toolbox; 2012. http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php.
34. Cornell Database Group. Cornell anonymization toolkit; 2014. https://sourceforge.net/projects/anony-toolkit/.
35. Dai C, Ghinita G, Bertino E, Byun J-W, Ninghui L. TIAMAT: a tool for interactive analysis of microdata anonymization techniques. *Proc VLDB Endow*. 2009;2(2):1618-1621.
36. OpenAIRE. Anamnesia; 2019. https://amnesia.openaire.eu/index.html.
37. Poulis Giorgos, Gkoulalas-Divanis Aris, Loukides Grigorios, Skiadopoulos Spiros, Tryfonopoulos C. SECRETA: a system for evaluating and comparing relational and transaction anonymization algorithms. Paper presented at: Proceeding of the 17th International Conference on Extending Database Technology; 2014:620-623.
38. Sánchez D, Martínez S, Domingo-Ferrer J. Comment on "Unique in the shopping mall: on the reidentifiability of credit card metadata". *Science*. 2016;351(6279):1274-1274.
39. Fung Benjamin C M. Selected publications; 2019. http://dmas.lab.mcgill.ca/fung/publicationsBySelection.htm.
40. Templ M. *Statistical Disclosure Control for Microdata: Methods and Applications in R*. Cham, Switzerland: Springer; 2017.
41. Hundepool A, Willenborg L. ARGUS: software packages for statistical disclosure control. In: Payne R, Green P, eds. *COMPSTAT*. Physica, Heidelberg; 1996;341-345.
42. Prasser F, Kohlmayer F, Kuhn KA. The importance of context: risk-based de-identification of biomedical data. *Methods Inf Med*. 2016;55(4):347-355.
43. Prasser F, Gaupp J, Wan Z, et al. An open source tool for game theoretic health data de-identification. Paper presented at: Proceedings of the AMIA Annual Symposium; 2017:1430-1439.
44. Bild R, Kuhn KA, Prasser F. SafePub: a truthful data anonymization algorithm with strong privacy guarantees. *Proc Priv Enhanc Technol*. 2018;2018(1):67-87.
45. Prasser F, Eicher J, Bild R, Spengler H, Kuhn KA. A tool for optimizing de-identified health data for use in statistical classification. Paper presented at: Proceedings of the 30th International Symposium on Computer-Based Medical Systems; 2017:169-174.
46. Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA. ARX - A comprehensive tool for anonymizing biomedical data. Paper presented at: Proceedings of the AMIA Annual Symposium; 2014:984-993.
47. Prasser F, Kohlmayer F. Putting statistical disclosure control into practice: the ARX data anonymization tool. *Medical Data Privacy Handbook*. Cham: Springer; 2015:111-148.
48. Le Fevre K, DeWitt DJ, Ramakrishnan R. Incognito: efficient full-domain k-anonymity. Paper presented at: Proceedings of the International Conference on Management of Data*;* 2005:49-60.
49. El Emam K, Dankar FK, Issa R, et al. A globally optimal k-anonymity method for the de-identification of health data. *J Am Med Inform Ass*. 2009;16(5):670-682.
50. Kohlmayer F, Prasser F, Eckert C, Kemper A, Kuhn KA. Flash: efficient, stable and optimal k-anonymity. Paper presented at: Proceedings of the International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing; 2012:708-717.
51. Le Fevre Kristen, De Witt David J, Ramakrishnan Raghu. Mondrian multidimensional k-anonymity. Proceedings of the 22nd International Conference on Data Engineering. 2006;:25–25.
52. Ohno-Machado L, Vinterbo S, Dreiseitl S. Effects of data anonymization by cell suppression on descriptive statistics and predictive modeling performance. *J Am Med Inform Ass*. 2002;9(Suppl 6):S115-S119.
53. Nergiz ME, Atzori M, Clifton C. Hiding the presence of individuals from shared databases. Paper presented at: Proceedings of the International Conference on Management of Data; 2007:665-676.
54. Sweeney L. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 2002;10(05):557-570.
55. El Emam K, Dankar FK. Protecting privacy using k-anonymity. *J Am Med Inform Assoc*. 2008;15(5):627-637.
56. Pannekoek J. Statistical methods for some simple disclosure limitation rules. *Statistica Neerlandica*. 1999;53(1):55-67.
57. Chen G, Keller-McNulty S. Estimation of identification disclosure risk in microdata. *J Off Stat*. 1998;14(1):79.
58. Hoshino N. Applying Pitman's sampling formula to microdata disclosure risk assessment. *J Off Stat*. 2001;17(4):499-520.
59. Zayatz Laura Voshell. Estimation of the percent of unique population elements on a microdata file using the sample. Statistical Research Division Report Number: Census/SRD/RR-91/08; 1991.

60. Dankar F, El Emam K, Neisa A, Roffey T. Estimating the re-identification risk of clinical data sets. *BMC Med Inform Decis Mak*. 2012;12(1):66.
61. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: privacy beyond k-anonymity. Paper presented at: Proceedings of the 22nd International Conference on Data Engineering; 2006:24.
62. Stammler S, Katzenbeisser S, Hamacher K. Correcting finite sampling issues in entropy l-diversity. Paper presented at: Proceedings of the International Conference on Privacy in Statistical Databases; 2016:135-146.
63. Li N, Li T, Venkatasubramanian S. t-Closeness: privacy beyond k-anonymity and l-diversity. Paper presented at: Proceedings of the 23rd International Conference on Data Engineering; 2007:106-115.
64. Brickell J, Shmatikov V. The cost of privacy: destruction of data-mining utility in anonymized data publishing. Paper presented at: Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining; 2008:70-78.
65. Cao J, Karras P. Publishing microdata with a robust privacy guarantee. *Proc VLDB Endow*. 2012;5(11):1388-1399.
66. Zhiyu W, Yevgeniy V, Weiyi X, et al. A game theoretic framework for analyzing re-identification risk. *PLoS One*. 2015;10(3):e0120592.
67. El Emam K, Arbuckle L. *Anonymizing Health Data: Case Studies and Methods to Get You Started*. 1st ed. Sebastopol, CA: O'Reilly Media, Inc.; 2013.
68. Iyengar VS. Transforming data to satisfy privacy constraints. Paper presented at: Proceedings of the International Conference on Knowledge Discovery and Data Mining; 2002:279-288.
69. Gionis A, Tassa T. k-anonymization with minimal loss of information. Paper presented at: Proceedings of the European Symposium on Algorithms; 2007:439-450.
70. Prasser F, Bild R, Kuhn KA. A generic method for assessing the quality of de-identified health data. Paper presented at: Proceedings of the Medical Informatics Europe (MIE2016 @ HEC2016); 2016:312-316.
71. LeFevre K, DeWitt DJ, Ramakrishnan R. Workload-aware anonymization techniques for large-scale datasets. *ACM Trans Database Syst*. 2008;33(3):1-47.
72. Sánchez D, Martínez S, Domingo-Ferrer J. Supplementary materials for "How to avoid reidentification with proper anonymization" – comment on "Unique in the shopping mall: on the reidentifiability of credit card metadata". arXiv:1511.05957v22015.
73. Manning AM, Haglin DJ, Keane JA. A recursive search algorithm for statistical disclosure assessment. *Data Min Knowl Disc*. 2008;16(2):165-196.
74. A benchmark of different transformation models supported by ARX; 2019. https://github.com/arx-deidentifier/transformation-benchmark.
75. European Medicines Agency. EMA/240810/2013 - European Medicines Agency policy on publication of clinical data for medicinal products for human use; 2014. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf.
76. European Medicines Agency. EMA/90915/2016 – external guidance on the implementation of the European medicines agency policy on the publication of clinical data for medicinal products for human use; 2018. https://www.ema.europa.eu/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-3.pdf.
77. European Union Agency for Network and Information Security. Privacy and data protection by design; 2015. https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design.
78. Elliot M, Mackey E, O'Hara K, Tudor C. *The anonymisation decision-making framework*. Manchester: UKAN; 2016.
79. Office of the Australian Information Commissioner. The de-identification decision-making framework; 2017. https://www.oaic.gov.au/privacy/guidance-and-advice/de-identification-decision-making-framework/.
80. Ministère des Solidarités et de la Santé. Données de santé: Anonymat et risque de ré-identification; 2015. https://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/publications/les-dossiers-de-la-drees/dossiers-solidarite-et-sante/article/donnees-de-sante-anonymat-et-risque-de-re-identification.
81. Bäck Asta, Keränen Janne. Anonymisointipalvelut. Tarve ja toteutusvaihtoehdot Liikenne- ja viestintäministeriö; 2017. https://julkaisut.valtioneuvosto.fi/handle/10024/79579.
82. Personal Data Protection Commission of Singapore. Guide to basic data anonymisation techniques; 2018. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-_v1-(250118).pdf.
83. Polish Ministry of Digitalization. Open data - Security standard; 2018. https://dane.gov.pl/media/ckeditor/2018/11/06/security-standard_2018.odt.
84. Dutch Ministry of Justice and Security. On statistical disclosure control technologies; 2018. https://www.wodc.nl/binaries/Cahier2018-20_2889_Fulltext_tcm28-362210.pdf.
85. Ministry of Science and ICT. A research on de-identification technique for personal identifiable information; 2016. https://www.fsd.tuni.fi/aineistonhallinta/en/anonymisation-and-identifiers.html.
86. Finnish Social Science Data Archive. Data management guidelines: anonymisation and personal data; 2018. https://www.fsd.tuni.fi/aineistonhallinta/en/anonymisation-and-identifiers.html.
87. Research Data Library Team. RDM Walkthrough Guide. École polytechnique fédérale de Lausanne (EPFL) Bibliothèque. URL: https://www.epfl.ch/campus/library/wp-content/uploads/2019/09/RDM_Walkthrough_Guide_20190930.pdf.
88. University of Guelph. Clean and prepare your data; 2018. https://guides.lib.uoguelph.ca/CleanAndPrepareData/5.
89. LMU Munich. Conduct your study; 2019. https://www.osc.uni-muenchen.de/toolbox/resources_for_researchers/conduct_your_study/index.html.
90. University of Kassel. Management of research data; 2019. https://www.uni-kassel.de/themen/forschungsdatenmanagement/service-hilfe/faq.html.

91. Korea Internet & Security Agency. KISA promotes training on identification of personal information. https://www.kisa.or.kr/notice/press_View.jsp?mode=view&p_No=8&b_No=8&d_No=1570.
92. TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung. ANONTrain: Praktische Anwendung von Anonymisierungswerkzeugen. http://www.tmf-ev.de/Desktopmodules/Bring2Mind/DMX/Download.aspx?EntryId=28213&PortalId=0.
93. Domingo-Ferrer J, Sánchez D, Hajian S. Database privacy. *Privacy in a Digital, Networked World*. Basel, Switzerland: Springer; 2015.
94. Torra V. *Data privacy: Foundations, new developments and the big data challenge*. Basel, Switzerland: Springer; 2017.
95. Data Anonymization in KNIME. A redfield privacy extension walkthrough; 2019. https://www.knime.com/blog/data-anonymization-in-knime-a-redfield-privacy-extension-walkthrough.
96. Stephan K, Jens H, Johann-Christoph F. SAP HANA goes private: from privacy research to privacy aware enterprise analytics. *Proc VLDB Endow*. 2019;12(12):1998-2009.
97. Costa C, Chatzimilioudis G, Zeinalipour-Yazti D, Mokbel MF. Efficient exploration of telco big data with compression and decaying. Paper presented at: Proceedings of the 33rd International Conference on Data Engineering; 2017:1332-1343.
98. Kim J, Ha H, Chun B-G, Yoon S, Cha SK. Collaborative analytics for data silos. Paper presented at: Proceedings of the 32nd International Conference on Data Engineering; 2016:743-754.
99. Armando A, Bezzi M, Metoui N, Sabetta A. Risk-based privacy-aware information disclosure. *Int J Secur Softw Eng*. 2015;6(2):70-89.
100. Jiang C, Kuang L, Han Z, Ren Y, Hanzo L. Information credibility modeling in cooperative networks: equilibrium and mechanism design. *IEEE J Select Areas Commun*. 2017;35(2):432-448.
101. Li X-Y, Zhang C, Jung T, Qian J, Chen L. Graph-based privacy-preserving data publication. Paper presented at 35th International Conference on Computer Communications; 2016:1-9.
102. Xu L, Jiang C, Chen Y, Ren Y, Liu KJR. Privacy or utility in data collection? a contract theoretic approach. *IEEE J Select Topics Signal Process*. 2015;9(7):1256-1269.
103. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. *Proc VLDB Endow*. 2018;11(10):1071-1083.
104. Kuzilek J, Hlosta M, Zdrahal Z. Open university learning analytics dataset. *Scientific Data*. 2017;4:170171.
105. Ursin G, Sen S, Mottu J-M, Nygård M. Protecting privacy in large datasets: first we assess the risk; then we fuzzy the data. *Cancer Epidem Prevent Biomar*. 2017;26(8):1219-1224.
106. Domingos PM. A few useful things to know about machine learning. *Commun ACM*. 2012;55(10):78-87.
107. Aggarwal CC. On k-anonymity and the curse of dimensionality. Paper presented at: Proceedings of the 31st International Conference on Very Large Data Bases; 2005:901-909.
108. Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of set-valued data. *Proc VLDB Endow*. 2008;1(1):115-125.
109. Prasser F, Spengler H, Bild R, Eicher J, Kuhn KA. Privacy-enhancing ETL-processes for biomedical data. *Int J Med Inform*. 2019;126:72-81.
110. Bowen J. *Getting Started with Talend Open Studio for Data Integration*. Birmingham: Packt Publishing Ltd; 2012.

How to cite this article: Prasser F, Eicher J, Spengler H, Bild R, Kuhn KA. Flexible data anonymization using ARX—Current status and challenges ahead. *Softw Pract Exper*. 2020;50:1277–1304. https://doi.org/10.1002/spe.2812

APPENDIX A PSEUDOCODE OF THE ALGORITHM

The core of the flexible transformation process described in this article is a routine which performs full-domain attribute generalization followed by record suppression and value aggregation (the latter is also called vertical partitioning). This process is sketched in Figure A1. The suppression limit `s` is used to specify that not more than `s` records may be suppressed. The process of optimal full-domain generalization followed by record suppression and aggregation is encapsulated in the call to the method `generalizeAndAggregate`. This method is not described in further detail, as the underlying algorithms Flash and Lightning have been covered in previous publications.[12,50] The only difference to the original algorithms is that the effect of record suppression is ignored when calculating the utility of the output produced by the available generalization schemes.

Figure A2 illustrates how the the method `transformRecords` is being applied to subsets of the records from the input dataset to implement the horizontal partitioning strategy. The pseudocode is formulated iteratively rather than recursively for ease of understanding.

In line 7, full-domain generalization is performed on the dataset `d`, resulting in the dataset `t`. `t` may contain records, which either have been transformed (ie, values generalized or aggregated) or which have been suppressed. The original versions of suppressed records are then extracted from `t` via the method `extractSuppressionCandidates` in line 8.

```
Dataset transformRecords(Dataset d, PrivacyParameter p, Integer s)
{
   Arx arx = new Arx();
   arx.addPrivacyModel(new PrivacyModel(p));
   arx.setSuppressionLimit(s);
   return arx.generalizeAndAggregate(d);
}
```

FIGURE A1 Pseudocode illustrating the method `transformRecords` [Colour figure can be viewed at wileyonlinelibrary.com]

```
1  Dataset transform(Dataset d, PrivacyParameter p, Integer partitions)
2  {
3     Dataset result = {};
4     for (Integer remaining = partitions; remaining > 0; remaining--)
5     {
6        Integer s = |d| - |d| / remaining;
7        Dataset t = transformRecords(d, p, s);
8        d = extractSuppressionCandidates(t);
9        if (|d| == |t|) { // If all records have been suppressed
10          result = union(result, t);
11          return result;
12       }
13       Dataset a = extractTransformedRecords(t);
14       result = result ⊠ a;
15    }
16 }
```

FIGURE A2 Pseudocode illustrating the anonymization method [Colour figure can be viewed at wileyonlinelibrary.com]

These records are then processed in the next iteration if the termination condition (line 9) is not met. In line 13, the method `extractTransformedRecords` returns all records which have been subject to attribute generalization or aggregation. These are then added to the intermediate result (line 14). The parameter `partitions` (also called *p* in Section refsec:advanced) determines the maximal number of iterations. In each iteration, the suppression limit used when calling `transformRecords` is calculated appropriately in line 6 to guarantee that the condition in line 9 is satisfied within at most `partitions` iterations. The choice of `partitions` balances execution times against data quality.

APPENDIX B EXAMPLE ILLUSTRATING THE APPROACH

In this section, we provide an example illustrating an application of our algorithm. In this process, we use the example dataset from Figure 4, which we have extended with two additional attributes *height* and *income*. Domain generalization

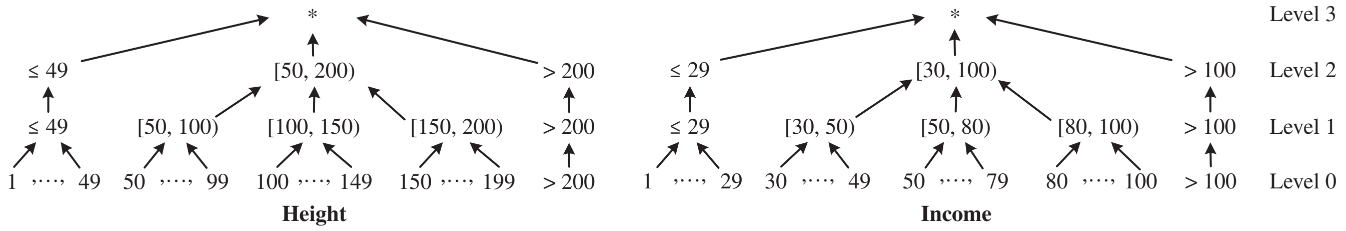
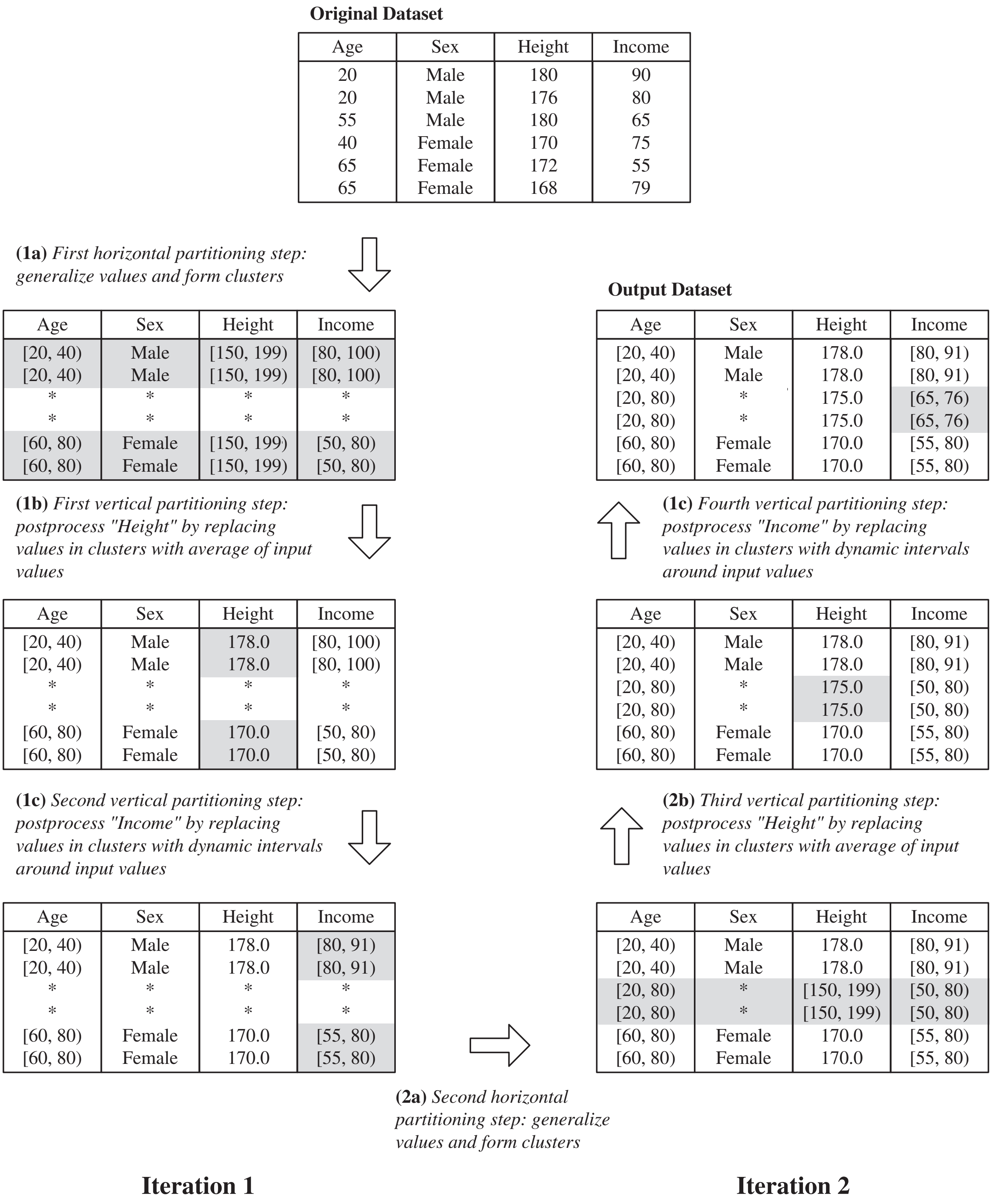


FIGURE B1 Domain generalization hierarchies for the additional attributes. The hierarchy to the left specifies possible generalizations of values of the attribute height and the hierarchy to the right specifies possible generalizations of values of the attribute income

hierarchies for the attributes *age* and *sex* have been provided in Figure 3. Hierarchies for the additional attributes are presented in Figure B1 below.

Figure B2 shows the original dataset as well as all steps executed to generate a 2-anonymous output dataset by applying local generalization to the attributes *age* and *sex* as well as aggregating *height* by replacing values with the arithmetic mean and aggregating *income* by generating dynamic intervals around values in each cluster. The algorithm terminates after two iterations, where each iteration consists of three steps. Cells transformed in each step are highlighted in grey.



Original Dataset

Age	Sex	Height	Income
20	Male	180	90
20	Male	176	80
55	Male	180	65
40	Female	170	75
65	Female	172	55
65	Female	168	79

(1a) *First horizontal partitioning step: generalize values and form clusters*

Age	Sex	Height	Income
[20, 40)	Male	[150, 199)	[80, 100)
[20, 40)	Male	[150, 199)	[80, 100)
*	*	*	*
*	*	*	*
[60, 80)	Female	[150, 199)	[50, 80)
[60, 80)	Female	[150, 199)	[50, 80)

(1b) *First vertical partitioning step: postprocess "Height" by replacing values in clusters with average of input values*

Age	Sex	Height	Income
[20, 40)	Male	178.0	[80, 100)
[20, 40)	Male	178.0	[80, 100)
*	*	*	*
*	*	*	*
[60, 80)	Female	170.0	[50, 80)
[60, 80)	Female	170.0	[50, 80)

(1c) *Second vertical partitioning step: postprocess "Income" by replacing values in clusters with dynamic intervals around input values*

Age	Sex	Height	Income
[20, 40)	Male	178.0	[80, 91)
[20, 40)	Male	178.0	[80, 91)
*	*	*	*
*	*	*	*
[60, 80)	Female	170.0	[55, 80)
[60, 80)	Female	170.0	[55, 80)

(2a) *Second horizontal partitioning step: generalize values and form clusters*

Age	Sex	Height	Income
[20, 40)	Male	178.0	[80, 91)
[20, 40)	Male	178.0	[80, 91)
[20, 80)	*	[150, 199)	[50, 80)
[20, 80)	*	[150, 199)	[50, 80)
[60, 80)	Female	170.0	[55, 80)
[60, 80)	Female	170.0	[55, 80)

(2b) *Third vertical partitioning step: postprocess "Height" by replacing values in clusters with average of input values*

Age	Sex	Height	Income
[20, 40)	Male	178.0	[80, 91)
[20, 40)	Male	178.0	[80, 91)
[20, 80)	*	175.0	[50, 80)
[20, 80)	*	175.0	[50, 80)
[60, 80)	Female	170.0	[55, 80)
[60, 80)	Female	170.0	[55, 80)

(1c) *Fourth vertical partitioning step: postprocess "Income" by replacing values in clusters with dynamic intervals around input values*

Output Dataset

Age	Sex	Height	Income
[20, 40)	Male	178.0	[80, 91)
[20, 40)	Male	178.0	[80, 91)
[20, 80)	*	175.0	[65, 76)
[20, 80)	*	175.0	[65, 76)
[60, 80)	Female	170.0	[55, 80)
[60, 80)	Female	170.0	[55, 80)

Iteration 1

Iteration 2

FIGURE B2 Example illustrating the partitioning strategies. Cells transformed in each step are highlighted in grey

- In step (1a), the first horizontal partitioning step, the dataset is generalized and clusters are formed. To this end, a generalization scheme is applied to the original dataset resulting in three clusters each containing two records. The second cluster contains two suppressed records. These will be transformed in the next iteration. The attributes *age* and *sex* are transformed using the generalization hierarchies.
- In step (1b), the first vertical partitioning step, the attribute height is aggregated by replacing the values in each cluster with the average of the associated values from the input dataset.
- Finally, in the last step of the first iteration, (1c), which constitutes the second vertical partitioning step, the attribute *income* is aggregated by replacing the values in each cluster with dynamic intervals around the associated input values.

In the second iteration, the same process is repeated in steps (2a), (2b), and (2c) for the two records suppressed in the first iteration, resulting in the final output dataset.

APPENDIX C SPECIFICATION OF THE DATASETS USED IN THE EXPERIMENTS

In this appendix, we present more details about the datasets used in the experiments. We note that we used the datasets to compare the performance (in terms of scalability and output data utility) of different anonymization algorithms to each other and not to perform case studies using a specific anonymization algorithm. The properties of the datasets which are most important for this comparison (ie, volume, dimensionality, uniqueness of data) are listed in Table 6. For reference, we list further details about the attributes of the datasets in this section. We note that in practice the selection of quasi-identifiers needs to be performed in a context-specific manner considering additional safeguards such as access restrictions (see Section 6). Analogously to many other studies using the same or similar datasets, we therefore simply selected a set of privacy-relevant attributes for each dataset to perform the comparison. As can be seen in the following paragraphs, the selected attributes included demographics (eg, age, marital status, sex), social parameters (eg, education, insurance coverage), financial data (eg, income), and health parameters (eg, weight, health problems). Finally, we note that all datasets are also available in our online repository.[74]

Table C1 presents a list of the attributes of the "US Census" dataset, which comprises eight categorical attributes and one numeric attribute. The heights of the generalization hierarchies used for anonymization varied between two and five. The dataset contains an excerpt from the 1994 US census database from which records containing "null" values have been removed. We note that this dataset is a de facto standard dataset for comparing anonymization algorithms and that we have removed records containing "null" values only to replicate the setup most commonly used, not because the algorithms studied are not able to handle missing data (see Section 3.1). Further information is available online: http://archive.ics.uci.edu/ml/datasets/adult.

Table C2 presents a list of the attributes of the "Competition" dataset. As can be seen, the dataset comprises two categorical and six numeric attributes. The heights of the generalization hierarchies used for anonymization varied between two and six. The dataset originates from the 1998 KDD data mining competition. Further information is available online: http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html.

Table C3 presents a list of the attributes of the "Crash statistics" dataset, which comprises seven categorical attributes and one numeric attribute. The heights of the generalization hierarchies used for anonymization varied between two and six. We note that the attributes "ideathmon" and "ideathday" are categorical, because the dataset contains special categories for missing values ("not applicable" and "unknown"). The dataset originates from the Fatality Analysis Reporting System (FARS) of the US National Highway Traffic Safety Administration (NHTSA) and can be accessed here: ftp://ftp.nhtsa.dot.gov/FARS/.

Table C4 presents a list of the attributes of the "Time Use Survey" dataset, which comprises eight categorical attributes and one numeric attribute. The heights of the generalization hierarchies used varied between two and six. The dataset originates from the American Time Use Survey. Further information is available online: http://atusdata.org/index.shtml.

Table C5 presents a list of the attributes of the "Health Interviews" dataset. As can be seen, the dataset comprises five categorical and four numeric attributes. The heights of the generalization hierarchies used for anonymization varied between two and six. The dataset originates from the US Integrated Health Interview Series. Further information is available online: https://nhis.ipums.org/nhis/.

Table C6 presents a list of the attributes of the "Community Survey" dataset, which comprises 27 categorical and three numeric attributes. The heights of the generalization hierarchies used for anonymization varied between two and five.

TABLE C1 Specification of the "US Census" dataset

Attribute	Data type	Distinct values	Hierarchy height
sex	Categorical	2	2
age	Numeric	72	5
race	Categorical	5	2
marital-status	Categorical	7	3
education	Categorical	16	4
native-country	Categorical	41	3
workclass	Categorical	7	3
occupation	Categorical	14	3
salary-class	Categorical	2	2

The table presents a list of the attributes contained in the dataset, which consists of 30 162 records.

TABLE C2 Specification of the "Competition" dataset

Attribute	Data type	Distinct values	Hierarchy height
ZIP	Numeric	13 294	6
AGE	Numeric	94	5
GENDER	Categorical	6	2
INCOME	Numeric	7	3
STATE	Categorical	53	2
RAMNTALL	Numeric	814	5
NGIFTALL	Numeric	81	5
MINRAMNT	Numeric	58	5

The table presents a list of the attributes contained in the dataset, which consists of 63 441 records.

TABLE C3 Specification of the "Crash Statistics" dataset

Attribute	Data type	Distinct values	Hierarchy height
iage	Numeric	99	6
irace	Categorical	20	3
ideathmon	Categorical	14	4
ideathday	Categorical	33	4
isex	Categorical	3	2
ihispanic	Categorical	10	3
istatenum	Categorical	51	4
iinjury	Categorical	8	3

The table presents a list of the attributes contained in the dataset, which consists of 100 937 records.

TABLE C4 Specification of the "Time Use Survey" dataset

Attribute	Data type	Distinct values	Hierarchy height
Region	Categorical	4	3
Age	Numeric	83	6
Sex	Categorical	3	2
Race	Categorical	23	3
Marital status	Categorical	7	3
Citizenship status	Categorical	6	3
Birthplace	Categorical	155	3
Highest level of school completed	Categorical	18	4
Labor force status	Categorical	6	3

The table presents a list of the attributes contained in the dataset, which consists of 539 253 records.

TABLE C5 Specification of the "Health Interviews" dataset

Attribute	Data type	Distinct values	Hierarchy height
YEAR	Numeric	13	6
QUARTER	Numeric	4	3
REGION	Categorical	4	3
PERNUM	Numeric	25	4
AGE	Numeric	86	5
MARSTAT	Categorical	10	3
SEX	Categorical	2	2
RACEA	Categorical	16	2
EDUC	Categorical	26	2

The table presents a list of the attributes contained in the dataset, which consists of 1 193 504 records.

TABLE C6 Specification of the "Community Survey" dataset

Attribute	Data type	Distinct values	Hierarchy height
Insurance purchased	Categorical	2	2
Workclass	Categorical	10	3
Divorced	Categorical	3	2
Income	Numeric	464	5
Sex	Categorical	2	2
Mobility	Categorical	4	2
Military service	Categorical	5	2
Self-care	Categorical	3	2
Grade level	Categorical	17	3
Married	Categorical	3	2
Education	Categorical	25	4
Widowed	Categorical	3	2
Cognitive	Categorical	3	2
Insurance Medicaid	Categorical	2	2
Ambulatory	Categorical	3	2
Living with grandchildren	Categorical	3	2
Age	Numeric	93	4
Insurance employer	Categorical	2	2
Citizenship	Categorical	5	3
Indian Health Service	Categorical	2	2
Independent living	Categorical	3	2
Weight	Numeric	561	5
Insurance Medicare	Categorical	2	2
Hearing	Categorical	2	2
Marital status	Categorical	5	3
Vision	Categorical	2	2
Insurance Veteran's Association	Categorical	2	2
Relationship	Categorical	18	3
Insurance Tricare	Categorical	2	2
Childbirth	Categorical	3	2

The table presents a list of the attributes contained in the dataset, which consists of 68 725 records.

The dataset contains the data collected in the state of Massachusetts during the year 2013 as responses to the American Community Survey (ACS), an ongoing survey conducted by the US Census Bureau on demographic, social and economic characteristics from randomly selected people living in the US. Further information is available online: https://www.census.gov/programs-surveys/acs/.

APPENDIX D DEFINITION OF THE UTILITY MODEL USED IN THE EXPERIMENTS

In this appendix, we present a formal definition of the utility model used in the experiments. We denote the number of records in the dataset with n and the number of attributes in the dataset with m. The “granularity” model is a general-purpose utility measure based on the “loss” model proposed by Iyengar.[68] It is defined as:

$$1 - \frac{1}{m} \sum_{1 \leq x \leq m} \text{loss}(x), \tag{D1}$$

where $\text{loss}(x) \in [0, 1]$ returns the information loss for attribute x.

The information loss for an attribute x, denoted by $\text{loss}(x) \in [0, 1]$, is defined as the average information loss over all values of this attribute in the dataset:

$$\text{loss}(x) = \frac{1}{n} \sum_{1 \leq y \leq n} \text{loss}(x, y). \tag{D2}$$

The information loss per value, denoted by $\text{loss}(x, y) \in [0, 1]$, is calculated depending on the type of the attribute x and the transformation applied to the attribute:

1. For categorical and numeric attributes transformed using an associated generalization hierarchy, information loss per cell is defined as:

$$\text{loss}(x, y) = \frac{\text{leafs}(x, \text{value}(x, y)) - 1}{\text{leafs}(x, \text{root}(x)) - 1}, \tag{D3}$$

   where $\text{value}(x, y)$ returns the value of attribute x in record y, $\text{root}(x)$ returns the value of the root node of the generalization hierarchy for attribute x and $\text{leafs}(x, v)$ returns the number of leaf nodes rooted at the value v in the hierarchy of attribute x.
2. For numeric attributes which have been transformed into intervals (either by using a generalization hierarchy in which inner nodes represent intervals or by dynamic aggregation into intervals), information loss per cell is defined as:

$$\text{loss}(x, y) = \frac{|\text{upper}(\text{value}(x, y)) - \text{lower}(\text{value}(x, y))|}{|\max(x) - \min(x)|}, \tag{D4}$$

   where $\text{value}(x, y)$ returns the value of attribute x in record y, $\text{lower}(v)$ returns the lower bound of the interval described by value v, $\text{upper}(v)$ returns the upper bound of the interval described by value v, $\min(x)$ returns the smallest value of attribute x in the input dataset and $\max(x)$ returns the largest value of attribute x in the input dataset.
3. For values which have been suppressed, information loss is defined as:

$$\text{loss}(x, y) = 1, \text{if value}(x, y) \text{ is suppressed}, \tag{D5}$$

   which equals the information loss measured for attribute values which have been completely generalized or transformed into an intervals covering the complete domain of a numeric attribute.
4. For all other values, information loss is defined as:

$$\text{loss}(x, y) = 0, \text{in all other cases}, \tag{D6}$$

which implies that the model is not able to capture changes to data utility caused by other types of transformation, for example, by aggregating numeric values by replacing them with their mean.

We note that the model has been implemented in a manner that takes care of a wide range of edge cases. For example, it is made sure that no division by zero occurs should the domain of a variable consist of only one value and it is considered whether upper or lower bounds of intervals are inclusive or exclusive should the domain of a variable consist of integer values only. In summary, the model returns values in the range $[0, 1]$, where the original dataset has a utility of 100% and a transformed dataset in which all attribute values have been removed (either by generalization, suppression or by replacing them with intervals covering the complete domain of the attribute) has a utility of 0%.

Finally, we note that the fact that this model is not able to capture changes to data utility caused by aggregation operators other than the forming of dynamic intervals (eg, operators which replace values with their mean) is not relevant for the experiments presented in this article. The reason is that we only used generalization, suppression and replacement by dynamic intervals as other transformation operators are not supported by the tools to which we compared our software. ARX does, however, support further utility models such as the sum of squared errors, which can be used to analyze the impact of further types of aggregation (see Section 2.3).