Technische Universität München
TUM School of Computation, Information and Technology

Practical Approaches to the Truthful Anonymization of Microdata

Diplom-Informatiker
Raffael Manuel Bild

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Daniel Rückert

Prüfer der Dissertation:

1. Prof. Dr. Martin Boeker
2. Prof. Dr. Oliver Kohlbacher

Die Dissertation wurde am 31.01.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 20.11.2024 angenommen.

Für meine Eltern

Abstract

The amount of sensitive personal data collected and processed in modern times is ever-increasing. On the one hand, leveraging these developments, for example, using technologies from the fields of big data analytics and artificial intelligence, bears tremendous potential in many areas such as economics and research. On the other hand, such endeavors pose potential risks to the privacy of individuals. Therefore, privacy protection measures have to be applied when personal data is being processed.

A typical technical approach to privacy protection is data anonymization. It basically works by transforming personal data in a way that reduces risks to the privacy of individuals. This inevitably leads to a loss or distortion of information. Hence, data anonymization algorithms are required for selecting transformations that result in a reasonable trade-off between privacy risks and data quality.

This thesis addresses open questions from this area of research with a particular focus on anonymization methods that are truthful, which means that they preserve the semantic consistency of data. This is desirable in many application areas, such as the medical domain. The main contributions of this thesis are as follows:

Firstly, it is shown that thruthful data anonymization which satisfies differential privacy, a particularly strong privacy protection measure, is feasible in practice. An according flexible anonymization algorithm is proposed that is based on a relationship between truthful transformation techniques and differential privacy which has previously only been studied from a theoretical perspective. Evaluations and comparisons with prior work show that the proposed solution is scalable and provides data quality that can compete with, and sometimes even outperform, state-of-the-art solutions, even though they are not truthful and they are tailored to specific application scenarios.

Secondly, it is shown how data containing numeric attributes can be anonymized in a sufficiently scalable manner. To this end, various optimizations are proposed for implementing a well-known, computationally expensive approach for this purpose. They range from the mathematical level to the implementation level. Experimental evaluations show that these optimizations significantly reduce execution times in practice.

Finally, the effects of rounding errors during computations using decimal numbers on the privacy guarantees provided by implementations of various anonymization methods are analyzed. A reliable computing framework to mitigate the resulting negative impacts on the degree of protection is proposed. Extensive evaluations show that implementing data anonymization which is safe with respect to rounding errors is feasible and that it can be achieved with negligible impacts on scalability and data quality.

To make the proposed results available for applications in practice, they have been integrated into the open-source data anonymization tool ARX that has been used in multiple research projects, enabled several data publishing activities and has been mentioned in various official guidelines and policies.

Zusammenfassung

Die Menge an sensiblen personenbezogenen Daten, die modernen Zeiten gesammelt werden, nimmt stetig zu. Einerseits birgt die Verarbeitung dieser Daten, beispielsweise unter Verwendung von Technologien aus den Bereichen der Big Data Analyse und künstlichen Intelligenz, ein enormes Potenzial in vielen Bereichen wie Wirtschaft und Forschung. Andererseits bringen solche Bemühungen potenzielle Risiken für die Privatsphäre von Einzelpersonen mit sich. Daher sind Datenschutzmaßnahmen essenziell, wann immer personenbezogene Daten verarbeitet werden.

Eine zentrale technische Datenschutzmaßnahme ist die Datenanonymisierung. Dabei werden personenbezogene Daten auf eine Weise transformiert, welche Risiken für die Privatsphäre reduziert. Dies führt zwangsläufig auch zu einem Verlust oder einer Verrauschung von Informationen. Daher sind Anonymisierungsalgorithmen erforderlich, um die Auswahl von Transformationen zu unterstützen, welche zu einem angemessenen Kompromiss zwischen Datenschutzrisiken und Datenqualität führen.

Diese Arbeit behandelt offene Fragen aus diesem Forschungsbereich, wobei der Schwerpunkt auf Anonymisierungsmethoden liegt, die wahrheitserhaltend sind, was heißt, dass sie die semantische Konsistenz der Daten erhalten. Dies ist in vielen Anwendungsbereichen wünschenswert, wie beispielsweise im Bereich der medizinischen Forschung. Die wichtigsten Beiträge dieser Arbeit sind wie folgt:

Erstens wird gezeigt, dass wahrheitserhaltende Datenanonymisierung, die eine besonders starke Datenschutzgarantie namens Differential Privacy gewährleistet, praktikabel ist. Dazu wird ein entsprechender flexibler Anonymisierungsalgorithmus vorgeschlagen, der auf einer Beziehung zwischen wahrheitserhaltend Transformationsmethoden und Differential Privacy basiert, die zuvor nur aus theoretischer Sicht untersucht wurde. Evaluationen und Vergleiche mit früheren Arbeiten zeigen, dass der vorgeschlagene Algorithmus skalierbar ist und einen Grad an Datenqualität ermöglicht, der mit modernen, existierenden Lösungen mithalten oder sie manchmal sogar übertreffen kann, auch wenn diese Lösungen nicht wahrheitserhaltend und auf spezielle Anwendungsszenarien zugeschnitten sind.

Zweitens wird gezeigt, wie Daten mit numerischen Attributen in ausreichend skalierbarer Weise anonymisiert werden können. Hierzu werden verschiedene Optimierungen vorgeschlagen, um eine bekannte, besonders rechenintensive Methode für den Schutz numerischer Attribute effizient umzusetzen. Diese Optimierungen reichen von der mathematischen Ebene bis zur Implementierungsebene. Experimentelle Untersuchungen zeigen, dass die vorgeschlagenen Optimierungen in der Praxis zu einer erheblichen Reduktion von Ausführungszeiten führen.

Schließlich werden die Auswirkungen von Rundungsfehlern bei Berechnungen mit Dezimalzahlen auf die Datenschutzgarantien, die von Implementierungen verschiedener Anonymisierungsmethoden geboten werden, analysiert. Es wird ein zuverlässiges Berechnungsframework vorgeschlagen, um negative Auswirkungen auf das Schutzniveau zu vermeiden. Umfangreiche Evaluierungen zeigen, dass die Implementierung von Datenanonymisierung, die Sicherheit hinsichtlich Rundungsfehlern bietet, machbar ist und mit vernachlässigbaren Auswirkungen auf Skalierbarkeit und Datenqualität umgesetzt werden kann.

Um die vorgeschlagenen Ergebnisse für praktische Anwendungen verfügbar zu machen, wurden sie in das Open Source Datenanonymisierungstool ARX integriert, das in diversen Forschungsprojekten eingesetzt wurde, mehrere Datenveröffentlichungsaktivitäten ermöglicht hat und in verschiedenen offiziellen Richtlinien erwähnt wird.

Danksagung

Als Erstes möchte ich mich herzlich bei meinem Betreuer Prof. Klaus A. Kuhn und meinem Mentor Prof. Fabian Prasser dafür bedanken, dass sie diese Dissertation ermöglicht und unterstützt haben. Ich habe dabei sehr viel über das wissenschaftliche Arbeiten gelernt.

Mein besonderer Dank gilt meinen Eltern Christel und Ingo sowie meinem Bruder Michael für die liebevolle, kontinuierliche Unterstützung und Motivation, nicht nur in Bezug auf diese Doktorarbeit, sondern in allen Belangen.

Schließlich möchte ich mich bei meinen aktuellen und ehemaligen Kolleginnen und Kollegen am Institut für KI und Informatik in der Medizin für die wunderbare Zusammenarbeit und schönen Zeiten auch abseits der Arbeit bedanken. Insbesondere danke ich Prof. Martin Boeker, Johanna Eicher, Cornelia Fütterer, Florian Kohlmayer, Ingrid Martin, Raphael Scheible und Helmut Spengler für anregende Diskussionen, hilfreiche Hinweise und Verbesserungsvorschläge sowie wertvolle Unterstützung und Motivation während der Erstellung dieser Arbeit.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction and Outline

1.1 Introduction

In the digital age, ever-growing amounts of sensitive personal data are being collected and processed. This data covers almost all areas of our daily lives, including our online behavior and financial transactions as well as our locations and medical conditions over time. On the one hand, leveraging these developments using modern technologies, for example from the fields of artificial intelligence and big data analytics, bears tremendous potential in many areas such as economics, research and health care. Example applications include product recommender systems and learning health systems in which "every clinical encounter contributes to research and research is being applied in realtime to clinical care" [DPDA+16]. On the other hand, such endeavors pose potential risks to the privacy of individuals, which might result in severe consequences such as discrimination or redlining. Therefore, regulations and laws such as the US Health Insurance Portability and Accountability Act (HIPAA) [oHfCR02], the European General Data Protection Regulation (GDPR) [Cou16] or the Chinese Personal Information Security Specification [Sta18] mandate the application of a wide range of safeguards when processing personal data, ranging from the organizational to the technical level.

A typical legal basis for processing personal data on the organizational level is the consent of the data subjects. This basically means that personal data can be used for secondary purposes, such as marketing or research, if individuals authorize or consent to this use of their data. However, this approach has the disadvantage that it often requires valuable resources and can be associated with significant costs to provide individuals with sufficient information to obtain truly informed consent, to document and manage those consents, and finally to enforce them in the context of specific data use scenarios. Consent-based use is particularly challenging for retrospective data, as this can involve the cost of having to subsequently contact potentially large groups of individuals, some of whom may have moved or died in the meantime. In addition, individuals who give consent may have different characteristics than individuals who do not give consent, which can lead to biased datasets [EEA13, EJMA13].

A prominent privacy protection measure on the technical level is the encryption of personal data [KL20]. This technique is based on converting the original data, known as plaintext, into a non-disclosive form, known as ciphertext. The ciphertext should not reveal any information about the original plaintext to a potential observer, and only authorized parties should be able to decrypt the ciphertext back to the plaintext. This is typically performed using a key that is known only to the authorized party and within a secure environment. While data encryption can protect the confidentiality of data, it also has some inherent drawbacks. For example, it requires a secure management of keys for decrypting ciphertexts and restricts the group of potential data recipients to selected authorized parties. Moreover, these parties have to be trustworthy, and even if this can be assured (for example through additional contractual measures), data encryption does not protect from accidental privacy violations. Examples include IT security breaches or inadvertent disclosure by the authorized party. This drawback also applies to other technical measures for protecting the confidentiality of data, such as authentication and authorization [Bis03]. Cryptographic solutions such as encryption are often computationally expensive, especially when employed in a manner that allows for confidential computations, for example in the context of secure multiparty computation [CDN15] or homomorphic encryption [Gen09]. Moreover, many encryption methods rely on certain assumptions about limitations of the computational power of potential adversaries that can become unrealistic as technology evolves, and it has been argued that encryption is in general unsuitable when data is to be shared frequently and broadly [GM17].

A different technical approach to privacy protection is data anonymization [FWFP10], which is also termed de-identification or statistical disclosure control in certain areas. It essentially works by permanently transforming personal data in a way that reduces the risks to the privacy of individuals. In contrast to data encryption, data anonymization does not require key management and allows data to be shared with large groups of potentially untrusted parties, but it does not allow access to the original data.

The first step in a data anonymization process is typically the "masking" of directly identifying attributes, such as names and insurance numbers, in order to make them inaccessible to potential adversaries. To this end, a variety of techniques have been proposed. They include the suppression, i.e. removal, of whole attributes, randomization, i.e. the replacement of directly identifying data with randomly generated values, and pseudonymization. The latter basically means that directly identifying attributes are separated from the remaining attributes and stored in a secure environment, typically within an independent organizational unit, while associations to the dataset containing the non-directly identifying attributes are represented by pseudonyms (i.e. non-speaking identifiers). This approach is characterized by the fact that the association between individuals and their data can be restored by authorized persons under controlled conditions if required. This can, for example, be necessary if additional findings have been collected as part of a study in the medical domain and the patient has requested to be contacted again in this case [EEA13].

It is, however, well known that the masking of direct identifiers alone is not sufficient to protect the privacy of individuals. For example, Latanya Sweeney has shown that some attributes that are not directly identifying by themselves can still be used for the re-identification of an individual when combined or linked to other information [Swe02b]. Such attributes are termed key variables [WDW96] or quasi-identifiers [LDR06] in the literature. Typical examples include ZIP code, gender and age. Sweeney was able to use a combination of those to identify the record belonging to William Weld, who was the governor of Massachusetts at that time, within a presumably anonymized dataset by linking it to a public voter list [Swe02b]. As another example, Arvind Narayanan and Vitaly Shmatikov successfully identified records of known users in a dataset containing movie ratings of 500,000 Netflix users that was supposed to be anonymized by linking it with the Internet Movie Database [NS08]. As a third example, the public release of presumably anonymized search data from more than 600,000 users by AOL has led to the re-identification of individuals when it was linked to phone-book listings by reporters of the New York Times. As a consequence of this incident, a class action lawsuit was filed against AOL, with millions of US dollars going to the class members and lawyers [EEA13].

All these examples constitute successful re-identification attacks. Re-identification, however, is not the only privacy threat: Sensitive information, such as diagnoses or income, can also be disclosed without identifying the specific record that belongs to a person [Lam93]. It can, for example, occur by merely learning that the record of a person must belong to a subset of records that share common features, such as a common diagnosis code, to deduce that this diagnosis must also apply to the individual [MKGV07, NAC07]. This threat is known as attribute disclosure. Finally, even the knowledge that an individual's data is included in a dataset at all can pose a potential privacy risk in itself, for example, in the case of datasets obtained from criminal records or debtor lists. This threat is termed membership disclosure in the literature [NAC07].

As these examples illustrate, anonymizing data in such a manner that privacy risks are reduced to an acceptable level is challenging. It requires transformations of the whole dataset that go far beyond the masking of directly identifying attributes. These transformations inevitably result in a loss or distortion of information. Hence, data anonymization methods have to balance reductions of privacy risks against reductions of data quality so that the anonymized data can still be useful. Determining transformations that result in a reasonable trade-off between these conflicting objectives is a non-trivial optimization problem. It typically requires tool support and often involves repeated evaluations of data that has been transformed in different ways with respect to privacy risks and data quality [EEDI^{+}09]. Moreover, it strongly depends on the intended use of the data. Central problem areas in this field of research are:

- **Assuring privacy protection.** This is obviously the main requirement for data anonymization methods. This area includes the specification of formal models for assessing and quantifying the degree of privacy protection with respect to privacy threats such as re-identification, attribute and membership disclosure that are relevant in given application scenarios. Guaranteeing that a sufficient degree of privacy protection is provided often also requires careful analyses of the design and implementation of data anonymization methods to avoid accidental violations of the expected degree of privacy protection provided.

- **Offering scalability.** To be feasible in practice, solutions are required to be scalable in the sense that they can process sufficient amounts of data with sufficient performance in terms of processing time and memory consumption. This is particularly challenging in the context of data anonymization where complex optimization problems have to be solved and methods are often applied repeatedly. This area includes the design of efficient algorithms as well as highly optimized implementations.

- **Facilitating flexibility.** Finally, it is desirable that methods for anonymizing data are flexible in the sense that they can be tailored to a variety of different application scenarios. This area includes the design of generic methods that can be parameterized with a variety of different models for assessing data quality and privacy risks and that ideally allow for flexible combinations of these models. The specification of according formal quality models that approximate the utility of data with respect to different application scenarios is an active area of research in its own right [EKP17, EBS^{+}20].

This publication-based doctoral thesis addresses challenges from all three problem areas. It proposes solutions that have been presented in three full papers of which the author of this thesis is the first author and which have been published in international, peer-reviewed journals and conference proceedings.

1.2 Outline

This thesis is structured as follows:

- Chapter 1 introduced the topic, the scope and the problem areas addressed.
- Chapter 2 describes methodological background about anonymization methods as well as computations using decimal numbers which are investigated in the context of data anonymization in this work.
- Chapter 3 formulates the research questions this thesis aims to answer.
- Chapter 4 provides an overview of the contributions of the work presented and relates them to the research questions presented in Chapter 3.
- Chapter 5 discusses the results presented, relates them to prior work and points out possible directions for further research.
- Chapter 6 concludes this dissertation.
- Appendix A contains the full texts of the papers included in this thesis.
- Appendix B provides an overview of all further publications to which the author has contributed during the time of work on this thesis.

CHAPTER 2

Methodological Background

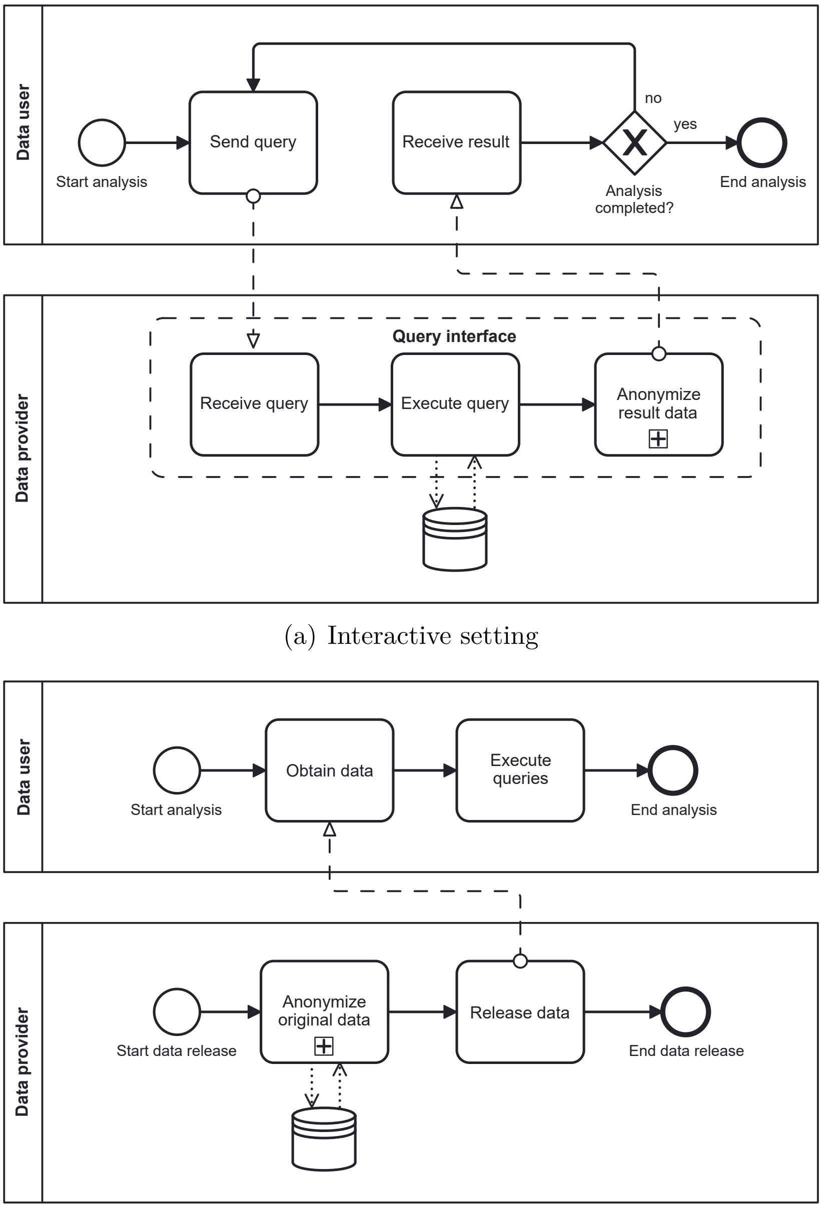
This chapter introduces relevant formalisms, definitions and methodological concepts from the field of data anonymization. It presents transformation methods and data anonymization algorithms for selecting a suitable concrete transformation strategy for a given dataset in general. Then, it describes different models for measuring the quality of data and, in particular depth, various privacy models and differential privacy, which are of particular relevance to this thesis. Moreover, this chapter introduces background information from the field of computations using decimal numbers, which is applied in the context of data anonymization in subsequent chapters.

2.1 Data Anonymization

In the literature, a vast amount of methods for data anonymization have been proposed. They can be structured according to various criteria, ranging from the supported usage scenarios over characteristics of the transformations performed to the categories of data that can be processed.

Regarding the supported usage scenarios, it can be distinguished between data anonymization methods for interactive settings and methods for non-interactive settings [Dwo06]. Figure 2.1 shows process models in BPMN notation [Whi04] that illustrate both.

As can be seen in Figure 2.1(a), in the interactive setting, an interface is provided for the data user that accepts queries, executes them on the original data and anonymizes the results before returning them to the data user. Data anonymization methods for the interactive setting generate safe output data, typically with a low degree of granularity, such as aggregated data. These approaches are usually tailored to a specific set of restricted queries and they often support only a limited amount of queries per user, as the degree of sensitive information disclosed with each query may add up [DR13]. An inherent drawback of the interactive setting is that the maintenance and monitoring of



(a) Interactive setting

(b) Non-interactive setting

Figure 2.1: Different settings in which data anonymization methods can be applied.

the query interface as well as the management and surveillance of its users and their queries generates costs and requires resources.

In contrast, in the non-interactive setting sketched in Figure 2.1(b), the data provider anonymizes the original data and releases the resulting dataset. Such datasets typically contain microdata, i.e. data on the level of individual persons, of high granularity. Since the data is made available to (possibly multitudes of) data users for processing, data anonymization methods for the non-interactive setting can be seen as producing safe input data for analyses. From the perspective of users, access to microdata has several advantages which stem from its inherent flexibility: The user can execute arbitrary queries on the data, both in terms of their quantity and variety. For example, access to microdata allows an analyst to generate an arbitrary amount of statistics in any manner desired, including individual-level multivariate analyses [DESG11]. Compared to interactive approaches, non-interactive approaches therefore allow for a more flexible use of data and do not require the operation of a query interface. They have hence been recommended for applications in various field, including the medical domain [DE12, EEA13]. However, in contrast to interactive approaches, typically the immediate sovereignty over fine-granular data is lost. Moreover, due to its fine granularity, microdata is often much more prone to privacy attacks than aggregated data and hence particularly challenging to anonymize.

Methods for data anonymization can also be categorized based on the kind of transformations they perform as being truthful or non-truthful (also known as perturbative) [DEE13]. Perturbative methods permit transformations which distort data in ways which may change semantics, for example, by changing the age of an individual from 42 to 48 via the addition of random noise. Such methods typically preserve the domain of attributes, which is desirable for certain kinds of analyses [DFT05]. In contrast, thruthful methods guarantee that semantics are preserved and that merely the information content of data is reduced. For example, the age 42 of an individual could be generalized to the interval $[40-60[$. Truthfulness can be desirable in many areas [BA05]. Examples include governmental as well as industrial applications [PGDL$^+$14] and the medical domain, in which semantic inconsistencies introduced by perturbation (such as dosages or combinations of drugs which would be harmful) have led to problems in introducing non-truthful approaches [DEE13].

Moreover, it can be distinguished between methods for anonymizing longitudinal data and methods for anonymizing cross-sectional data [RMGM08, PLGDS13]. The former kind of data includes series of values associated with events that have been documented over extended periods of time, such as locations or commercial transactions of individuals. Thereby, the focus is typically on a small and specific set of attributes of interest. The latter kind of data consists of values of different attributes that have been documented at a specific point in time. Such data typically comprises a wider range of attributes, such as age, gender, ZIP code and income, and hence, it typically provides a broader view on different properties of each individual. It is well-known that the anonymization of longitudinal data requires different techniques than the anonymization of cross-sectional data [EEA13, Agg05, TMK08].

As illustrated in Figure 2.2, the focus of this thesis is on non-interactive, truthful methods for the anonymization of cross-sectional microdata. As discussed, this type of data anonymization has desirable properties in many application areas, in particular within the medical domain.

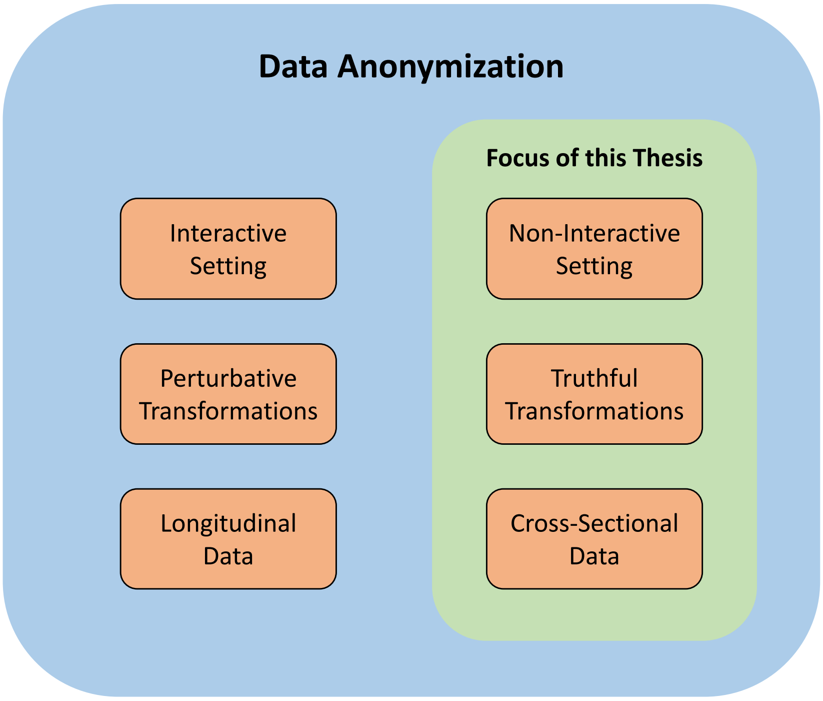


Figure 2.2: Methodological focus of this thesis within the field of data anonymization.

2.1.1 Datasets

In this work, a dataset containing cross-sectional microdata is without loss of generality regarded as a single table in which each column corresponds to one attribute and each row to the data contributed by one individual. Unless noted otherwise, it is assumed that no attribute is directly identifying by itself, assuming that such attributes have been masked (cf. Chapter 1) and therefore need not be considered further. As it is common in literature, each row is referred to as a record and datasets are regarded as multisets of records. For an arbitrary dataset D with m attributes, the domains of attribute 1 to m are formally denoted by $\Omega_1, ..., \Omega_m$ so that D is a multiset $D \subseteq \Omega_1 \times ... \times \Omega_m$ and every record $r = (r_1, ..., r_m) \in D$ is an ordered sequence $r \in \Omega_1 \times ... \times \Omega_m$. For a given sequence of m attributes, the universe of all datasets $D \subseteq \Omega_1 \times ... \times \Omega_m$ is denoted with $\mathcal{D}$.

2.1.2 Transformation Methods

Methods for data anonymization commonly transform data in a manner that reduces the uniqueness of (combinations of) values. Typically, the goal is to decrease the distinguishability of records in a dataset.

Relevant techniques used for this purpose can be categorized along different axes. Some methods are data-dependent, which means that not only the domains of the attributes are considered, but also the concrete values contained in the dataset at hand. Otherwise, they are called data-independent [LQS11].

Some methods transform a dataset using global recoding, which means that identical values of an attribute are always transformed into the same value. In contrast, local recoding allows identical values to be transformed into different values [DESG11]. In the following, three of the most common transformation methods are introduced.

Microaggregation means that the records of a dataset are grouped into clusters and within each cluster, values are made indistinguishable by replacing them with a representative, such as the mean or the median [DFMS02]. This method is inherently data-dependent, perturbative, and it performs local recoding [PKK16b].

Using *suppression*, values are removed or replaced with a semantic-free placeholder. It can be performed on the level of individual cells, records or attributes. This method is truthful and in general, it performs local recoding. It can be performed in a data-dependent way, typically by suppressing (combinations of) values which appear infrequently in a given dataset. But it can also be performed data-independently, for example, in the form of random sampling. In this process, each record is independently sampled, i.e. retained, with a fixed sampling probability β or otherwise suppressed.

Generalization reduces uniqueness in a truthful manner. It can be used as a global or as a local recoding method. A typical way to perform generalization is to replace every value of an attribute with a more general but semantically consistent value on a fixed level of a so-called generalization hierarchy. Figure 2.3 shows exemplary generalization hierarchies for the attributes "Age" and "Gender". As can be seen, the values on level 0 of a hierarchy correspond to the original values in the dataset, i.e. they form the domain Ω_i of the attribute. The set of generalized values on levels greater than 0 is denoted by Λ_i. For a given value $r'_i \in \Omega_i \cup \Lambda_i$, each value on level 0 which is an element of the subtree rooted at r'_i is called a leaf node of r'_i. For example, the leaf nodes of "$[40, 80[$" in Figure 2.3 are "40", ..., "79".

With full-domain generalization, all values of an attribute are generalized to the same level. In contrast, using subtree generalization, each value $r_i \in \Omega_i$ of an attribute can be generalized to a fixed value $r'_i \in \Omega_i \cup \Lambda_i$ in such a way that different values may be generalized to different levels [FWFP10]. For example, The values "0", ..., "39" and all values of at least "80" of the attribute "Age" may be left unchanged, while the leaf nodes "40", ..., "79" of the subtree rooted at the node "$[40, 80[$" shown in Figure 2.3 may be replaced with this interval. Both full-domain and subtree generalization result in global recoding [PK15] and they are single-dimensional [LDR06], which means that each attribute is transformed individually and independently.

Multi-dimensional generalization [LDR06] is more general because it transforms possibly several attributes in groups, considering combinations of their values. It permits

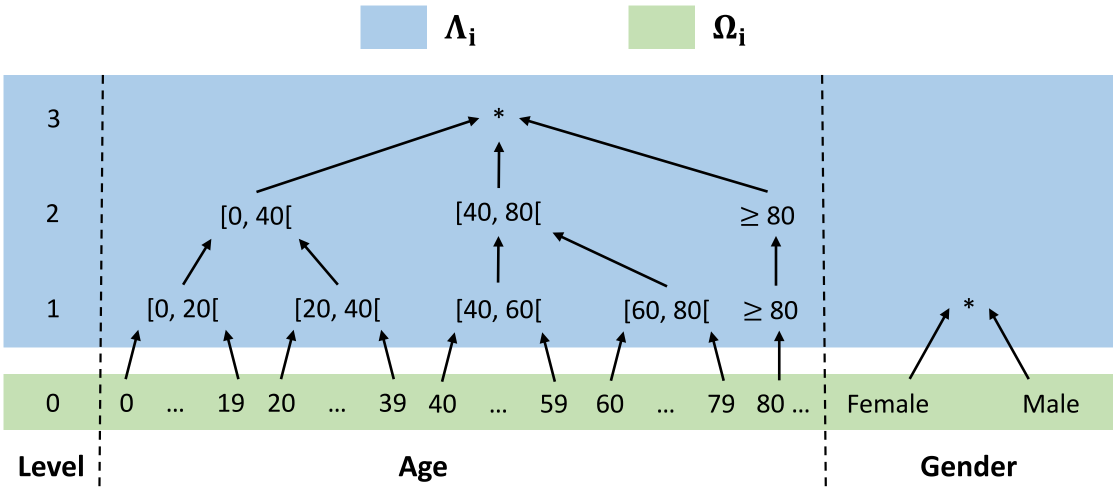


Figure 2.3: Example generalization hierarchies for the attributes “Age” and “Gender”.

that the same value of a given attribute may be generalized differently in different records, depending on the values of other attributes, which results in local recoding. For example, the age “45” may be generalized to “[40,60[” in records having the gender “Female”, while it may be generalized to “[40,80[” in records having the gender “Male”.

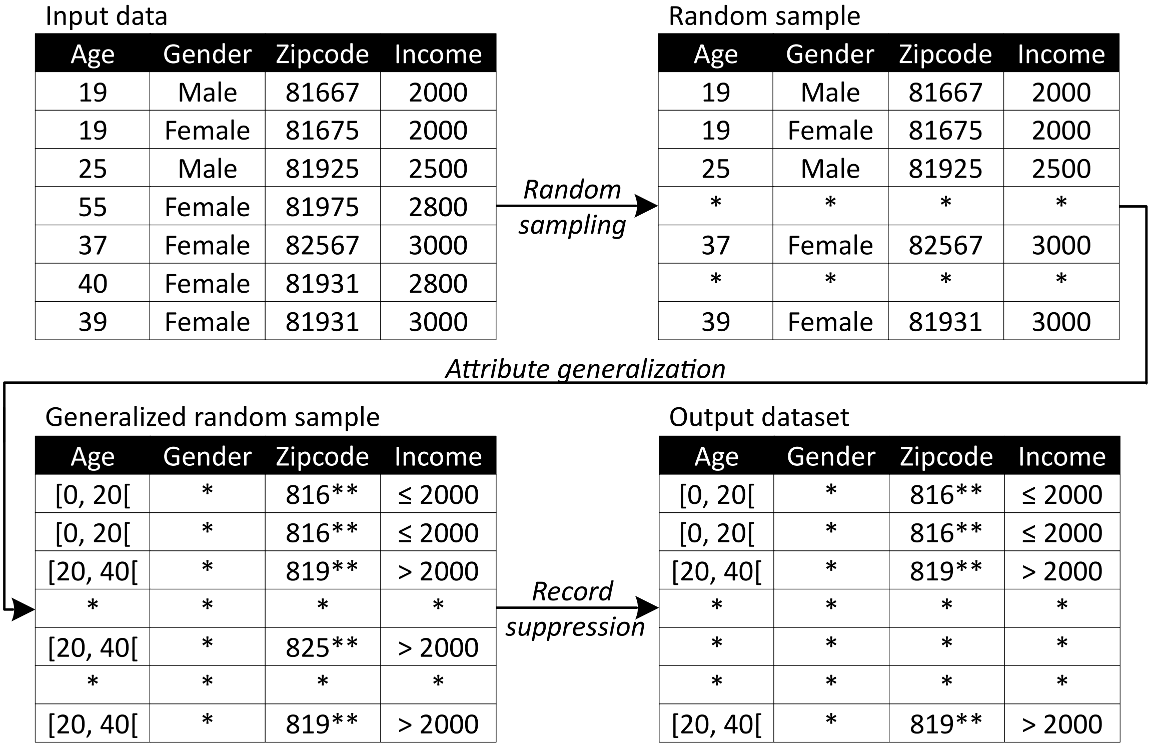
A generalization scheme is defined as a function

$$g : \Omega_1 \times ... \times \Omega_m \rightarrow (\Omega_1 \cup \Lambda_1) \times ... \times (\Omega_m \cup \Lambda_m)$$

which maps a record r to a (possibly) generalized record $g(r)$ so that the value of every attribute is either kept as-is or replaced with a value contained in the associated hierarchy. Generalization schemes can be used for formalizing transformations which perform full-domain, subtree or multi-dimensional generalization. A generalization scheme g is data-independent if it has been defined without considering the properties of the concrete dataset at hand, i.e. if it has been defined using a strategy which yields the same results for all possible datasets $D \subseteq \Omega_1 \times ... \times \Omega_m$.

Figure 2.4 shows how a dataset is transformed using random sampling followed by (full-domain) attribute generalization and record suppression. After the random sampling step, every value of the attributes “Age” and “Gender” is generalized to level one of the corresponding generalization hierarchy shown in Figure 2.3 while also the values of the other attributes are generalized in a consistent manner. Finally, the only remaining unique record is suppressed.

In this work, the suppression of a record is indicated by replacing it with the placeholder $* = (*, ..., *)$. Without loss of generality, it is assumed that a generalization hierarchy is provided for each attribute of a given dataset $D \subseteq \Omega_1 \times ... \times \Omega_m$. Since



Input data

Age	Gender	Zipcode	Income
19	Male	81667	2000
19	Female	81675	2000
25	Male	81925	2500
55	Female	81975	2800
37	Female	82567	3000
40	Female	81931	2800
39	Female	81931	3000

Random sample

Age	Gender	Zipcode	Income
19	Male	81667	2000
19	Female	81675	2000
25	Male	81925	2500
*	*	*	*
37	Female	82567	3000
*	*	*	*
39	Female	81931	3000

Generalized random sample

Age	Gender	Zipcode	Income
[0, 20[	*	816**	≤ 2000
[0, 20[	*	816**	≤ 2000
[20, 40[	*	819**	> 2000
*	*	*	*
[20, 40[	*	825**	> 2000
*	*	*	*
[20, 40[	*	819**	> 2000

Output dataset

Age	Gender	Zipcode	Income
[0, 20[	*	816**	≤ 2000
[0, 20[	*	816**	≤ 2000
[20, 40[	*	819**	> 2000
*	*	*	*
*	*	*	*
*	*	*	*
[20, 40[	*	819**	> 2000

Figure 2.4: Example showing how a dataset is transformed using random sampling, generalization and suppression.

generalizing a value to the highest level effectively suppresses the value, the root values of all generalization hierarchies are also denoted with $*$.

2.1.3 Data Anonymization Algorithms

As described in Chapter 1, data anonymization is a non-trivial task which goes far beyond the masking of directly identifying attributes. It requires processes for the (semi) automated selection of combinations of transformations that result in a suitable balance between privacy protection and data quality. This challenge has been addressed by different fields of science.

Statistics and research agencies have been publishing summaries and microdata since decades. Thereby, the data typically describes a sample from some underlying population so that analyses performed on the sample can be used to make inferences about the whole population. Sampling – be it performed implicitly during the acquisition of data or in the form of explicit random sampling as described in Section 2.1.2 – already provides a certain degree of privacy protection [WDW96, EEA13]. Prior to its release, the sampled data is then usually further transformed using methods of statistical disclosure control which summarize or perturb the data [Off]. To this end, a "principles-based" approach using soft rules-of-thumb, which are subject to review and change by experts, rather than hard rules are typically used [RE15]. This a posteriori approach to privacy protection focuses on data quality while attempting to protect privacy as much as possible [DESG11].

Researchers from the field of computer science started to address the problem of data anonymization by suggesting formal models for measuring privacy risks that typically involve a strict threshold on the level of risk which is deemed acceptable. Within the constraints imposed by such privacy models, data is then transformed in such a way that data quality is preserved as far as possible. From an abstract point of view, this a priori approach to privacy protection simplifies the problem of achieving a reasonable balance between the conflicting goals of privacy protection and data quality to an optimization problem with a single objective function. It favors strict formal degrees of privacy protection over data quality.

Regardless of the approach: To be feasible in practice, data anonymization requires support by automation. A posteriori data anonymization processes typically leave more room for manual selections of transformations and for evaluations of data than a priori processes. To support a priori data anonymization, typically fully automated algorithms are employed. The development of such algorithms that are sufficiently scalable and preferably allow for flexible parameterizations to tailor their outputs to different application scenarios is an active area of research. From an abstract point of view, most such algorithms implement a process as sketched in Figure 2.5. For a given input dataset, they search the space of applicable (combinations of) data transformations or refine a given transformation and assess the resulting privacy risks and quality of data using privacy and quality models, respectively. Upon their termination, they return the solution that provides the highest degree of data quality among all transformations assessed that satisfy risk thresholds which have been specified for the privacy model beforehand.

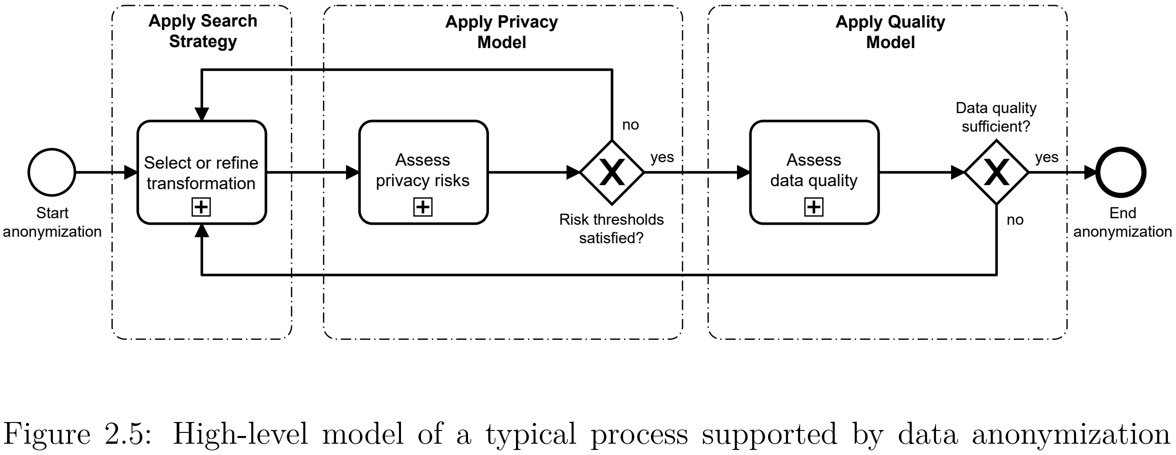
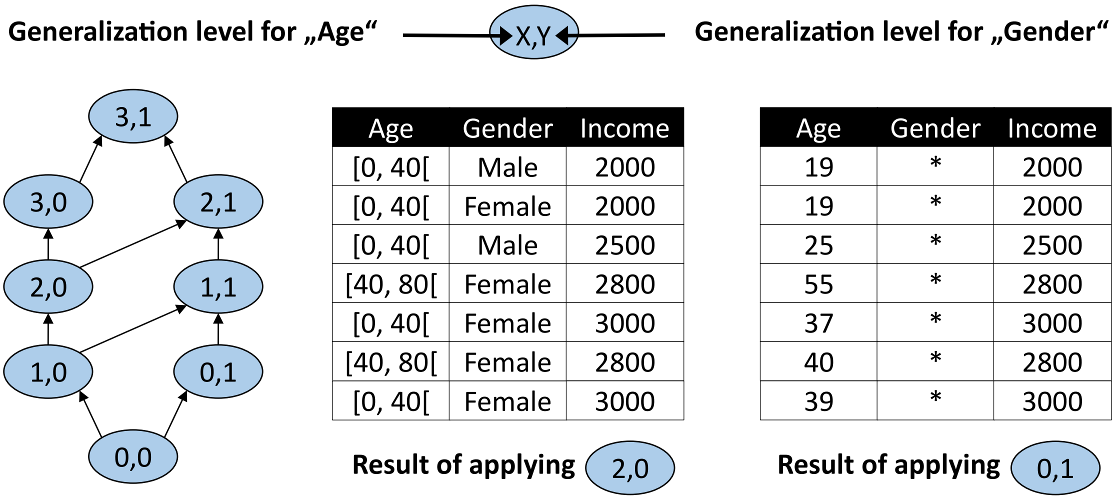


Figure 2.5: High-level model of a typical process supported by data anonymization algorithms.

The structure of the search spaces used by anonymization algorithms strongly depends on the kind of transformations they support. Traditional truthful data anonymization algorithms are based on search spaces that consist of generalization schemes (cf. Section 2.1.2). They transform data by applying the according generalization strategies followed by the suppression of all records that do not comply with the risk threshold specified by the privacy model [PK15, EEDI+09].

When using this approach with full-domain generalization, the resulting search space can be modeled as a graph consisting of one node for each combination of generalization levels of the generalization hierarchy associated with each attribute. These search spaces have the algebraic structure of a lattice [DP02] and are hence usually referred to as generalization lattices [LDR05]. Figure 2.6 shows the generalization lattice induced by the hierarchies shown in Figure 2.3 as well as transformed versions of the example input dataset from Figure 2.4 that demonstrate the generalization strategies corresponding to two nodes. An arrow indicates that a node is a direct successor of a more specialized node in the sense that it can be derived from its predecessor by incrementing the generalization level of exactly one attribute.



Age	Gender	Income
[0, 40[	Male	2000
[0, 40[	Female	2000
[0, 40[	Male	2500
[40, 80[	Female	2800
[0, 40[	Female	3000
[40, 80[	Female	2800
[0, 40[	Female	3000

Age	Gender	Income
19	*	2000
19	*	2000
25	*	2500
55	*	2800
37	*	3000
40	*	2800
39	*	3000

Figure 2.6: Example of a generalization lattice and the transformations associated with two different nodes.

Obviously, the higher the number of transformations an anonymization algorithm supports, the better is the data quality that can be theoretically achieved. However, anonymization algorithms which perform complex combinations of transformations typically cannot yield a globally optimal solution because the resulting search spaces are far too large. Already in the relatively simple case of generalization lattices, the number of possible transformations grows exponentially with the number of attributes, i.e. the dimensionality of data [PBE+16]. Consequently, a multitude of heuristic search strategies [Swe97, BRK+13, NAC07, WYC04, LDR06], pruning methods [PKK16a] and clustering algorithms [SCDFSM15, BKBL07, GMT08, GT09, NC07] have been proposed for selecting possibly good transformations within a reasonable timeframe.

Most of these algorithms are special-purpose algorithms in the sense that they are designed for specific kinds of privacy models and quality models. This means that they effectively focus only on specific privacy risks and data use scenarios. Examples of more flexible data anonymization algorithms that can tailor their outputs to various different application scenarios include Flash [KPE+12], Lightning [PBE+16] as well as the generic top-down search algorithm and the genetic search algorithm presented in [MBDP21]. Flash can be used to determine an optimal solution when processing datasets of low to medium dimensionality, while the other algorithms utilize heuristic search approaches that can also be used for anonymizing data of higher dimensionality. In their current stage of development, all of these algorithms can be flexibly parameterized to use a wide range of different combinations of privacy models and quality models as well as various transformation methods [PES+20].

2.1.4 Privacy Models

Privacy models used in the context of data anonymization algorithms as described in the previous section typically measure privacy risks based on syntactical properties of transformed datasets. They are hence also known as syntactic privacy models [CT13].

k-Anonymity [Swe02b] is the oldest and most well-known such model. It was proposed by Latanya Sweeney to protect from re-identification attacks as described in Chapter 1. For a given risk threshold k, it basically requires that each record is indistinguishable from at least $k-1$ other records. Formally, it can be defined as follows:

Definition 1 (*k-Anonymity [Swe02b]*)

A dataset $D \subseteq (\Omega_1 \cup \Lambda_1) \times \ldots \times (\Omega_m \cup \Lambda_m)$ satisfies k-anonymity if each record $r' \in D$ cannot be distinguished from at least $k-1$ other records in D. Thereby, two records are considered indistinguishable if they share the same value for each quasi-identifier.

An alternative definition is based on the following concept of an equivalence class:

Definition 2 (*Equivalence Class [LDR06]*)

For a given dataset $D \subseteq (\Omega_1 \cup \Lambda_1) \times \ldots \times (\Omega_m \cup \Lambda_m)$ and record $r' \in D$, an equivalence class *$E = E_{r'} \subseteq D$ is defined as the multiset of all records in D that have the same combination of quasi-identifier values as r'.*

Equivalence classes according to this definition are also equivalence classes in the mathematical sense. An equivalence class E is said to satisfy k-anonymity if $|E| \geq k$ holds. With these notions, a dataset D satisfies k-anonymity according to Definition 1 if every equivalence class $E \subseteq D$ satisfies k-anonymity.

In a k-anonymous dataset, the re-identification risk of all records is restricted to not more than $1/k$ [NT15]. The reason is that, intuitively speaking, every individual hides in a group with a size of at least k. The dataset shown in Figure 2.4 satisfies 2-anonymity (for every subset of attributes considered as quasi-identifiers) after the suppression of the record with the unique (generalized) ZIP code value. The transformed example dataset shown to the left in Figure 2.6 satisfies 2-anonymity as well with respect to the quasi-identifiers “Age” and “Gender”.

As already mentioned in Chapter 1, re-identification is not the only privacy threat, and k-anonymity does not protect from membership or attribute disclosure. For example, if an attacker has the background knowledge that data about a woman aged 55 is contained in the transformed dataset shown to the left in Figure 2.6, he can still infer the income of this person. In order to protect from such attacks, Machanavajjhala et al. have proposed a further privacy model termed ℓ-diversity [MKGV07]. It is defined as follows:

Definition 3 (*ℓ-diversity [MKGV07]*)

An equivalence class E is ℓ-diverse if it contains at least ℓ “well-represented” values for a sensitive attribute i. A dataset $D \subseteq (\Omega_1 \cup \Lambda_1) \times ... \times (\Omega_m \cup \Lambda_m)$ is ℓ-diverse if all its equivalence classes are ℓ-diverse.

In the article [MKGV07], Machanavajjhala et al. present different interpretations of the notion of “well-represented” used in Definition 3 that result in different variants of ℓ-diversity:

1. Distinct-ℓ-diversity essentially guarantees that each equivalence class contains at least ℓ different sensitive attribute values.

2. Recursive-ℓ-diversity basically assures that the most frequently occurring sensitive attribute values in an equivalence class are not too frequent and that the rarely occurring values are not too rare.

3. Entropy-ℓ-diversity is based on the information-theoretic notion of (Shannon) entropy [Sha01] and defines an equivalence class E to be ℓ-diverse if it satisfies

$$-\sum_{i=1}^{n} p_i \ln(p_i) \geq \ln(\ell)$$

   where $(p_1, ..., p_n)$ is the distribution of sensitive attribute values in E.

These three variants of ℓ-diversity are presented in ascending order of the degree of privacy protection provided. It is worth pointing out that a ℓ-diverse dataset always satisfies ℓ-anonymity. The reason is that every equivalence class has to contain at least ℓ distinct sensitive attribute values, which implies that every equivalence class has to contain at least ℓ records [Koh16].

While ℓ-diversity protects from privacy threats that go beyond re-identification, at least two further attacks exists which may result in sensitive attribute disclosure and against which ℓ-diversity does not protect [LLV07]. The first one is termed skewness attack and basically exploits uneven distributions of sensitive attribute values within an equivalence class. The second one is known as similarity attack and essentially exploits sensitive attribute values which are syntactically different, but semantically closely related.

To protect against sensitive attribute disclosure in a way which also mitigates such attacks, Li et al. have proposed the privacy model t-closeness [LLV07]. Informally, it guarantees that the distribution of sensitive attribute values within each equivalence class is close to the corresponding distribution within the whole dataset. In order to measure the distance between two distributions, the authors propose to use the Earth Mover's Distance (EMD) [RTG00] which essentially quantifies the minimal amount of work needed to transform one distribution into another by means of moving distribution mass. Formally, t-closeness is defined as follows:

Definition 4 (*t-closeness [LLV07]*)

An equivalence class E is t-close if the EMD of sensitive attribute values to the distribution of sensitive attribute values in the whole dataset is at most t. A dataset $D \subseteq (\Omega_1 \cup \Lambda_1) \times ... \times (\Omega_m \cup \Lambda_m)$ is t-close if all its equivalence classes are t-close.

The article [LLV07] shows different ways to calculate the EMD for attributes with different scales of measure. For categorical attributes, the authors propose two variants: Firstly, a simple one, which assumes an equal distance, namely one, between any two categorical values. Secondly, a more complex one, which uses a generalization hierarchy to calculate the distance between two values. Finally, for a totally ordered (and in particular a numeric) sensitive attribute, the authors propose to calculate the EMD of sensitive values $P(E) = (p_1, ..., p_n)$ within a given equivalence class E and the corresponding distribution $Q = (q_1, ..., q_n)$ within the whole dataset according to the formula

$$EMD\,[P(E), Q] = \frac{1}{n-1} \sum_{i=1}^{n} \left| \sum_{j=1}^{i} (p_j - q_j) \right| . \tag{2.1}$$

This variant of t-closeness is known as ordered-distance t-closeness.

The privacy models which have been described in this section are established models for protecting from various different privacy threats which may lead to re-identification and sensitive attribute disclosure. They are relevant in different usage scenarios and they are the central syntactic privacy models addressed in this thesis. However, it is worth pointing out that a multitude of further syntactic privacy models have been proposed in the literature:

For example, the model β-likeness [CK12] aims to protect against sensitive attribute disclosure in a manner that provides a more intuitive notion of protection than t-closeness. However, unlike t-closeness, β-likeness can only be used for protecting categorical attributes.

Furthermore, privacy models based on concepts from the field of game theory have been proposed which take the costs and benefits of potential attacks into account, based on the assumption that a potential attacker only attempts re-identification if the potential gains outweigh the costs of an attack [WVX+15, PGW+17]. They have been successfully applied in the context of genomic data sharing [WVX+17].

The privacy model m-invariance [XT07] has been proposed to protect against privacy threats in the context of the repeated re-publication of a dataset which is being updated over time.

δ-Presence [NAC07] is a well-known privacy model that protects against membership disclosure. It requires that the background knowledge of an adversary is modeled as a dataset that is a superset of the dataset to be anonymized. It then essentially enforces bounds on the probability that records from the larger dataset are contained in the smaller dataset. It is worth pointing out that protection against membership disclosure according to δ-presence can also reduce the risk of re-identification and attribute disclosure.

All privacy models introduced so far are suitable for the protection of cross-sectional data. In contrast, k^m-anonymity [TMK08] has been proposed specifically for the protection of longitudinal data which is typically high-dimensional. This privacy model is comparable to k-anonymity with the difference that it protects against adversaries who having knowledge about at most m quasi-identifying attributes.

Many more examples besides these few selected ones can be found in the survey [WE18] by Wagner et al. that covers more than 80 privacy models.

2.1.5 Quality Models

Measuring the quality of transformed data in the context of data anonymization is challenging because the usefulness of data strongly depends on the use case. Hence, various models have been proposed and discussed in the literature that can be used to estimate data quality with respect to different scenarios.

When it is unknown in advance how the output data will be used or when the variety of intended uses is broad, general-purpose quality models can be employed. They typically estimate data quality by quantifying the amount of information loss, for example based on comparisons of the input data with transformed data [FWFP10].

Such models can be roughly classified as measuring information loss on the level of individual attributes, cells or records. Typical examples of changes on these levels are differences in the distributions of attribute values, reductions in the granularity of individual values or changes in the distinguishability of records, respectively.

A commonly used attribute-level general-purpose model is

- (Non-Uniform) Entropy, which essentially measures the amount of information that can be obtained about values of attributes in the input dataset by observing values of the corresponding attributes in the transformed dataset [DW99].

Typical examples of cell-level general-purpose models are

- Precision, which quantifies information based on the generalization levels of transformed values [Swe02a], and
- Loss, which assesses the extent to which transformed values cover the domains of their respective attributes [Iye02].

Finally, examples of common record-level general-purpose models include

- Group Size, which measures information loss based on the average size of equivalence classes [LDR06], and
- Discernibility, which essentially penalizes records based on the size of their respective equivalence class [BA05].

Special-purpose (also known as workload-aware) quality models approximate the quality of data with respect to specific applications [LLV07,LDR08]. A typical example for such an application from the fields of machine learning and statistics is statistical classification. The task thereby is to predict the values of a predefined set of class attributes from combinations of values of a given set of feature attributes. This is implemented with supervised learning in which a classifier, i.e. a prediction model, is created using a training set that contains correctly assigned feature and class attribute values [WE05].

Specific quality models have been developed for optimizing data for the purpose of statistical classification. A well-known example is the classification metric proposed by Iyengar which measures the suitability of data as a training set for statistical classifiers [Iye02]. Essentially, it works by penalizing records which contain infrequent combinations of feature and class attribute values.

2.1.6 Differential Privacy

Traditional syntactic privacy models are typically based on assumptions about an attacker's background knowledge and protect from specific privacy threats. As can be concluded from the examples provided in Section 2.1.4, this has led to the development of more and more such models over time which aim to provide protection in scenarios that were previously not adequately addressed. This is an ongoing process as new attack scenarios are continuously being discovered and it has been argued that the degree of privacy protection provided by models focusing on syntactical properties of data only is inherently limited [Dwo08, DFT08].

Cynthia Dwork has proposed a different approach to privacy preserving data processing termed differential privacy [Dwo06, DMNS06] that aims to overcome these limitations and to provide a rigorous definition of privacy protection. Unlike syntactical privacy models, differential privacy is not a property of a dataset, but a property of a data processing method. It does not make assumptions about the background knowledge of attackers or about specific privacy threats. This also means that effectively, the need to identify potential quasi-identifiers and sensitive attributes is obviated. Essentially, differential privacy guarantees that the probability of any possible output of a probabilistic algorithm (termed mechanism in the literature) does not change significantly if the data of one person is added to or removed from the input dataset. This allows a person to plausibly deny that his or her data was even included in a dataset that has been processed. It effectively means that an individual can safely contribute data because this will not significantly influence any possible consequences the output of a differentially private mechanism may have for that individual, be they adverse or not. However, as a consequence of the strong degree of privacy protection provided, differentially private mechanisms often have to make compromises in terms of flexibility or data quality. They are usually special-purpose algorithms for specific applications or support only limited sets of queries and they are typically perturbative (see [YJ14, VSBH13, DR13, FJ15, JLE14, JYC15] for examples).

Formally, ϵ-differential privacy can be defined with respect to datasets D_1 and D_2 satisfying $|D_1 \oplus D_2| = 1$, which means that D_1 can be obtained from D_2 by either inserting or removing one record, as follows:

Definition 5 (*ϵ-differential privacy [BCD+07]*)

A mechanism $\mathcal{M}$ provides ϵ-differential privacy if for all datasets D_1 and D_2 with $|D_1 \oplus D_2| = 1$ and all measurable $S \subseteq Range(\mathcal{M})$, the following holds:

$$P[\mathcal{M}(D_1) \in S] \leq \exp(\epsilon) \times P[\mathcal{M}(D_2) \in S].$$

The parameter ϵ, which is commonly termed the privacy budget, is typically a small positive number and determines the degree of privacy protection provided. Smaller values of ϵ result in tighter bounds on the permitted changes of output probabilities resulting from the insertion or removal of one record and hence in stronger degrees of protection.

A well-known way for achieving ϵ-differential privacy is termed randomized response [DR13]. This method can be used to perform privacy-preserving surveys. It can be illustrated by a person flipping a coin in secret and answering "yes" if it comes up heads while giving the true answer otherwise.

Another approach which is frequently used when output values are real numbers is the Laplacian mechanism [DMNS06]. It basically works by perturbing correct outputs by adding random noise drawn from an appropriately scaled Laplace distribution. This mechanism is commonly applied by query interfaces in interactive data usage scenarios (cf. Section 2.1) that support the execution of statistical queries.

Another common method known as the exponential mechanism [MT07] can be used for achieving ϵ-differential privacy when the output range $\mathcal{R}$ is an arbitrary measure space. It basically works by ranking all potential outputs $r \in \mathcal{R}$ for a given input dataset $D \in \mathcal{D}$ using a so-called score function $s : \mathcal{D} \times \mathcal{R} \to \mathbb{R}$ which assigns a real-valued score to each pair $(D, r) \in \mathcal{D} \times \mathcal{R}$. It then randomly selects and returns an output $r \in \mathcal{R}$ according to a specific probability distribution which depends on the privacy budget ϵ and on the so-called sensitivity Δs. The latter denotes the biggest possible difference in the output of s which can result from the insertion or removal of one record. More precisely, the sensitivity Δs of a score function $s : \mathcal{D} \times \mathcal{R} \to \mathbb{R}$ can be defined as

$$\Delta s \; = \; \sup_{r \in \mathcal{R}} \; \sup_{D_1, D_2 \in \mathcal{D} \colon \, |D_1 \oplus D_2| = 1} \; |s(D_1, r) - s(D_2, r)| \, .$$

With this notion, the exponential mechanism can be formally defined as follows:

Definition 6 (*Exponential mechanism [MT07]*)

Let $\mathcal{R}$ be an arbitrary set on which a measure μ is defined. For every function $s : \mathcal{D} \times \mathcal{R} \to \mathbb{R}$ with a finite sensitivity Δs, the exponential mechanism $\mathcal{E}^{\epsilon}(D, s)$ chooses and outputs an element $r \in \mathcal{R}$ with a probability proportional to $\exp\left(\frac{s(D,r)\epsilon}{2\Delta s}\right) \times \mu(r)$.

As can be seen from Definition 6, the exponential mechanism assigns higher probabilities to potential outputs with higher scores. At the same time, higher sensitivities result in lower differences between the probabilities assigned to different potential outputs. Consequently, score functions should be designed in such a way that higher scores are assigned to outputs which are more desirable while the sensitivity should be possibly small.

In order to compose more complex algorithms out of building blocks for achieving ϵ-differential privacy such as the ones described above, a variety of composition theorems can be used. One example is the following theorem which states that sequences of ϵ-differentially private computations also provide ϵ-differential privacy:

Theorem 1 (*Sequential composition [McS09]*)

For $i = 1, ..., n$, let the mechanism $\mathcal{M}_i$ provide ϵ_i-differential privacy. Then the sequence $\mathcal{M}_1^{r_1}(D), ..., \mathcal{M}_n^{r_n}(D)$, where $\mathcal{M}_i^{r_i}$ denotes mechnism $\mathcal{M}_i$ supplied with the outcomes of $\mathcal{M}_1, ..., \mathcal{M}_{i-1}$, satisfies $(\sum_{i=1}^{n} \epsilon_i)$-differential privacy.

It is known that truthful microdata cannot be released in a manner which satisfies strict ϵ-differential privacy [Dwo11]. Hence, unless noted otherwise, this thesis focuses on a relaxation known as (ϵ, δ)-differential privacy which tolerates that the bound $\exp(\epsilon)$ in Definition 5 may be exceeded with a small probability δ. This relaxation can be formally defined as follows:

Definition 7 (*(ϵ, δ)-differential privacy [DEE13]*)

A mechanism $\mathcal{M}$ provides (ϵ, δ)-differential privacy if for all datasets D_1 and D_2 with $|D_1 \oplus D_2| = 1$ and all measurable $S \subseteq Range(\mathcal{M})$,

$$P[\mathcal{M}(D_1) \in S] \leq \exp(\epsilon) \times P[\mathcal{M}(D_2) \in S]$$

is satisfied with a probability of at least $1 - \delta$.

It should be noted that some articles such as [DKM$^+$06] use a slightly different relaxation of ϵ-differential privacy which is more relaxed, i.e. provides a lower degree of privacy protection, than Definition 7. The interested reader can find further details about these variants of differential privacy and their relation for example in the articles [KS08, GMW$^+$12].

2.2 Computations Using Real Numbers

To be usable in practice, data anonymization methods need to be integrated into reliable and sufficiently scalable software tools. Such programs have to perform computations using formulas as presented above which often involve arithmetic using real numbers. In principle, such computations cannot be performed exactly by computers as they have finite resources. Consequently, the set of numbers which can be represented exactly by computers, the so-called machine numbers, is finite and there exists an infinite amount of unrepresentable real numbers between any pair of machine numbers. Arithmetic using such numbers can in principle merely approximate exact mathematical results and inevitably introduces approximation errors which strongly depend on the way the machine numbers are defined. These errors have been extensively studied in the field of numerical analysis. Central notions in this context are

1. the absolute error $e_{abs} = |x - \tilde{x}|$ and

2. the relative error $e_{rel} = e_{abs}/\,|x| = |x - \tilde{x}|\,/\,|x|$ for $x \neq 0$

where x is a real number and $\tilde{x}$ the machine number used to approximate it. Usually the absolute error is less relevant in practice than the relative error as the latter one measures approximation error in relation to the order of magnitude of the exact real value [Opf93].

2.2.1 Floating-Point Arithmetic

By far the mostly used representation of machine numbers is in the form of floating-point numbers. They have been designed to keep the relative approximation error as low as possible and they are almost always implemented according to the IEEE standard 754 [IEE08]. This standard specifies various formats that define representations of numeric values and special values that do not represent real numbers. Each numeric value is represented using three integers: A sign s, a significand c, and an exponent q. The mathematical value of such a number is

$$(-1)^s \times c \times b^q$$

where the base b can be either 2 or 10. Special values include $+\infty$, $-\infty$, and NaN, which stands for Not a Number and represents unrepresentable or undefined values, for example the result of zero divided by zero. Each format determines the base and has a fixed precision, i.e. number of digits of the significand, as well as a limited range of possible exponents.

Moreover, the IEEE 754 standard specifies that basic arithmetic operations using floating-point numbers have to be done as if it was possible to perform the corresponding operation with infinite precision and then to round the result to the nearest representable number. More formally, if $\tilde{x}$ and $\tilde{y}$ are floating-point numbers, $\oplus$ denotes a floating-point operation and $+$ the associated infinite precision mathematical operation, the standard guarantees that the following holds:

$$\tilde{x} \oplus \tilde{y} = round(\tilde{x} + \tilde{y}).$$

The precise semantics of the rounding operations denoted here with *round* are defined by rounding modes. They include "round to nearest" which means that exact results are rounded to the nearest floating-point value. Other rounding modes are “round towards zero”, “round towards $+\infty$” (also known as “ceiling”) and “round towards $-\infty$” (also known as “floor”).

Computations using floating-point numbers are efficient and supported by designated hardware implementations termed Floating-Point Units (FPUs). However, during sequences of floating-point calculations, these rounding errors inevitably add up and can lead to outputs which deviate significantly from the mathematically exact results [MBDD+09]. One approach for solving such issues is the careful static analysis of floating-point errors and their possible propagations. However, it is well-known that formally arguing about floating-point arithmetic is difficult and error-prone [MBDD+09, Mon08, Mir12].

2.2.2 Reliable Computing Techniques

To avoid the issue of floating-point error propagation, various reliable computing technologies can be used which are suitable for computations on different subsets of the real numbers and which have different advantages and disadvantages.

One obvious approach is to use software implementations of machine numbers in the form of arbitrary precision decimals, such as floating-point numbers consisting of significands and exponents with arbitrary lengths. However, this can only be used for calculations which yield results having a finite decimal expansion. Other computations, for example one divided by three, cannot be performed with this method in such a way that mathematically exact results are produced.

Another approach is exact arithmetic using rational numbers represented by fractions in the form of pairs of arbitrarily long integer enumerators and denominators. Rational numbers are known to have the algebraic structure of a field and hence, such machine numbers can be used to perform arbitrary combinations of addition, subtraction, multiplication and division. However, they cannot be used to implement all calculations involving elementary functions. For example, calculating roots, logarithms or the exponential function using rational arguments can yield irrational numbers. Moreover, computations using such numbers can become very slow when large enumerators or denominators are involved and when common denominators have to be calculated in the context of operations such as additions or subtractions.

Finally, interval arithmetic [Daw11], which allows to dynamically compute strict upper and lower bounds on the results of mathematical operations, can be used for reliable calculations on the whole set of real numbers. The basic idea is not to operate on (approximations of) real numbers, but rather on closed intervals within which the respective exact real numbers are guaranteed to lie. Functions operating on such intervals yield intervals which are guaranteed to include the exact result for any possible combination of real values contained in the input intervals. For example, addition on intervals can be performed by calculating

$$[\tilde{x_1}, \tilde{x_2}] + [\tilde{y_1}, \tilde{y_2}] = [\mathit{floor}\,(\tilde{x_1} + \tilde{y_1})\,, \mathit{ceiling}\,(\tilde{x_2} + \tilde{y_2})].$$

Using interval arithmetic, complex combinations of floating-point operations can be implemented on the whole set of real numbers while bounds on (both absolute and relative) errors of results can be computed dynamically during the execution of a program. Computations can be performed efficiently by exploiting optimized implementations of operations which modify the endpoints of intervals, for example using FPUs. Moreover, because of its conceptual simplicity, interval arithmetic can be implemented in a manner that is easy to understand, and hence, to validate. However, in contrast to approaches such as arithmetic using fractions as described above, interval arithmetic cannot be used to obtain exact results.

CHAPTER 3

Problem Statement

This chapter introduces the research questions this thesis aims to answer. They cover aspects from all problem areas discussed in Chapter 1, namely privacy protection, scalability and flexibility. The research questions are labeled **Q.1** to **Q.3** and they are addressed by the publications **Ref.1** to **Ref.3**, respectively, that are presented in the next chapter.

3.1 Differentially Private Data Anonymization

As discussed in Section 2.1.6, differential privacy provides a strong degree of privacy protection and does not rely on assumptions about an attacker's background knowledge. Consequently, it has received considerable attention in the field of data anonymization. However, as a consequence of the strong protection, methods from this field often have to make compromises in terms of flexibility or data quality. Traditional differentially private mechanisms are hence usually special purpose algorithms that provide good data quality while they have restricted output domains which can effectively support only specific usage scenarios. Examples include algorithms from the field of genetic research that output significant single-nucleotide polymorphisms for genome-wide association studies [YJ14] and methods from the field of machine learning, for example, for constructing statistical classifiers [VSBH13].

While differential privacy was initially proposed for the interactive scenario, in recent years, a growing number of differentially private mechanisms for non-interactive microdata release have been developed. They focus on strong privacy guarantees while striving to preserve the usefulness of their outputs for a possibly wide range of applications. However, methods from this area usually either generate synthetic datasets which mimic characteristics of the original dataset, or they produce perturbed versions of the original dataset by adding random noise. Current methods have been criticized for being difficult to explain to non-experts, for example to ethics committees, and for their perturbative nature [DEE13]. The development of truthful methods for differentially private microdata release is ongoing research that has been conducted primarily from a theoretical perspective up to now. The first research question addresses this problem:

Q.1 Can truthful microdata release that satisfies differential privacy be implemented in practice in a manner that is scalable and flexible while still providing a sufficient degree of data quality?

3.2 Efficient Protection of Numeric Attributes

Protecting data from sensitive attribute disclosure is challenging, in particular when sensitive attributes are numeric. As described in Section 2.1.4, t-closeness [LLV07] is a state-of-the-art privacy model that can be used for this purpose. It essentially guarantees that the distributions of sensitive attribute values within all equivalence classes do not deviate too much from the corresponding distribution of sensitive information in the entire dataset. The model has been specified in different variants that can be applied to attributes with different scales of measure. In particular, ordered-distance t-closeness is one of the few privacy models that have been proposed for the protection of sensitive numeric attributes. However, implementing it in practice in a sufficiently scalable and efficient manner is difficult. The reason is that directly evaluating the model as defined in Section 2.1.4 is particularly computationally expensive. It requires the computation of a double sum which is associated with a high time complexity of $O(n^2)$ where n is the numbers of records. This is particularly problematic in the context of data anonymization algorithms as described in Section 2.1.3 that evaluate privacy models repeatedly on data that has been transformed in different ways in order to determine a possibly good transformation strategy. The second research question tackles this challenge:

Q.2 How can numeric attributes be protected from sensitive attribute disclosure in a sufficiently scalable manner?

3.3 Reliable Data Anonymization

Implementing data anonymization in practice often requires reflecting the mathematical definitions of privacy models in software. This holds in particular for flexible data anonymization algorithms as described in Section 2.1.3 which directly compute the definitions of (possibly combinations of) privacy models on transformed data to assure that privacy protection constraints are satisfied. These definitions often involve calculations using real numbers which, as discussed in Section 2.2, are frequently approximated using floating-point numbers with a limited range and precision by computers.

It is well-known that calculations using such numbers inevitably introduce rounding errors which can add up during sequences of computations, leading to results that can differ significantly from the mathematically exact results [Mon08, Wil94]. Moreover, the exact behavior of floating-point operations can vary from implementation to implementation. It does not only depend on the particular design of computations in software, for example on the order in which arithmetic operations are executed, but it can also depend on the hardware, compiler configurations, libraries used, runtime environments etc., even when programming frameworks such as Java which strive for platform-independence are used [Mon08]. Floating-point errors are particularly problematic in the field of data anonymization where numerical inaccuracies can lead to privacy violations. For example, it has been shown that straight-forward floating-point implementations of the Laplacian mechanism (cf. Section 2.1.6) can be exploited in such a way that the entire content of a (presumably protected) database can be extracted [Mir12]. However, while general effects of floating-point arithmetic on the accuracy of computations have been extensively studied (see for example [Hig02]), there is still a lack of research about their impacts on the privacy guarantees provided by methods for data anonymization, in particular in the context of flexible data anonymization algorithms. The third research question aims to fill this gap:

Q.3 What are the effects of floating-point errors on the privacy guarantees provided by implementations of various data anonymization methods in the context of flexible data anonymization algorithms?

CHAPTER 4

Overview of Contributions

The main contributions of this cumulative thesis comprise results of three full papers of which the author of this dissertation is the first author and which have been published in international, peer-reviewed journals and conference proceedings. They investigate the research questions **Q.1**, **Q.2** and **Q.3** formulated in the previous chapter and they are accordingly referred to as contribution **Ref.1**, **Ref.2** and **Ref.3**. These publications contribute to all of the three problem areas privacy protection, scalability and flexibility in the context of data anonymization introduced in Chapter 1. Figure 4.1 illustrates which problem areas are mainly addressed by which contribution.

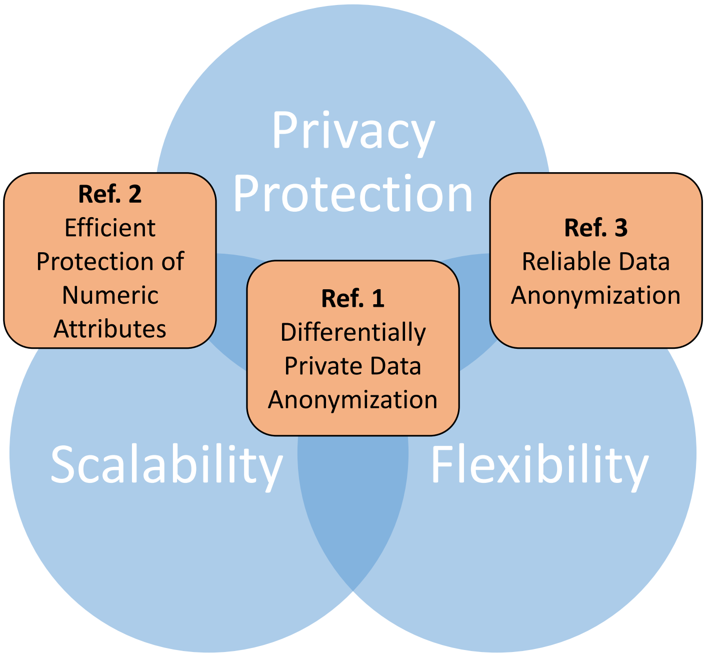

Privacy Protection
Ref. 2 Efficient Protection of Numeric Attributes
Ref. 3 Reliable Data Anonymization
Ref. 1 Differentially Private Data Anonymization
Scalability
Flexibility

Figure 4.1: Overview of the problem areas mainly addressed by the publications included in this thesis.

All contributions contain theoretical and conceptual results as well as practical evaluations using several real-world datasets, including: (1) *US Census*, an excerpt of records from the 1994 US Census database which is often used for evaluating anonymization algorithms, (2) *PVA Fundraising*, a dataset containing information about a fund raising appeal of the organization Paralyzed Veterans of America, (3) *Crash Statistics*, a dataset covering traffic accidents, (4) *Time Use Survey*, a dataset consisting of responses to a survey on individual time use in the US and (5) *Health Interviews*, a dataset containing records from a survey on the health of the US population [PES^{+}20]. The datasets have increasing volumes, ranging from about $30,000$ to more than a million records. All contain personal information such as gender or age and sensitive information such as income-related data. Table 4.1 provides an overview of these datasets and their properties.

Name	**Number of Attributes**	**Number of Records**	**Sensitive Attribute**
US Census	9	30,162	Salary Class
PVA Fundraising	8	63,441	Income
Crash Statistics	8	100,937	Injury Status
Time Use Survey	9	539,253	Labor Status
Health Interviews	8	1,193,504	Education

Table 4.1: Overview of the datasets used in the experimental evaluations.

In order to make the proposed results available for applications in practice, they have been integrated into the open-source data anonymization tool ARX [PES^{+}20, EBS^{+}20, PK15] which has been used in several research projects [XJC^{+}15, KHC^{+}16, CCZYM17], enabled multiple data publishing activities [KHZ17, JBD^{+}21, USMN17] and has been mentioned in various official guidelines and policies [Age, oS, MEO16].

In the following, each of the included publications is summarized, with a focus on the findings with respect to the research question investigated, the methods used and the results obtained. Moreover, the individual contributions of the author of this thesis are highlighted. Further information about each publication, including the respective full text, can be found in Appendix A. Other publications to which the author has contributed, but which are not included in this dissertation, are listed in Appendix B.

4.1 Differentially Private Data Anonymization

The article **Ref.1** shows that thruthful data anonymization which satisfies differential privacy is feasible. It does so by proposing a practical data anonymization algorithm based on a relationship between k-anonymity and differential privacy that has previously only been studied from a theoretical perspective (see [LQS11, LQS12]). These results have been further developed into a scalable and flexible algorithm that can be configured to optimize the quality of data for various different application scenarios while provably satisfying (ϵ, δ)-differential privacy. To develop this algorithm, multiple challenges had to be overcome:

Firstly, the articles [LQS11, LQS12] do not propose solutions for computing tight values of the parameters used for transforming data that guarantee a user-specified degree of privacy protection, which is a natural requirement for a data anonymization method. For that, proofs had to be extended and completed, including the development of a strategy for computing the maximum of a non-monotonic infinite sequence.

Secondly, the approach was initially limited to data-independent generalization [LQS11] which may severely impact data quality [DEE13]. While it is known that a differentially private algorithm for selecting a data-dependent generalization scheme can be used [LQS12], no concrete such method was described. To this end, **Ref.1** proposes a flexible search strategy and proves that it satisfies differential privacy. This method is based on repeated applications of the exponential mechanism composed using the sequential composition theorem. It can be parameterized with arbitrary score functions for quantifying data quality.

Thirdly, due to the above-mentioned previously open problems, no methods for tailoring the approach to specific applications and no practical evaluations could be performed yet. To this end, **Ref.1** presents score functions for optimizing output data with respect to different well-known quality models. They include Non-Uniform Entropy, Precision, Loss, Group Size, Discernibility as well as a special-purpose model tailored towards statistical classification. Analytical and experimental evaluations and comparisons with prior work show that the proposed solution is scalable and provides data quality which can compete with, and sometimes even outperform, state-of-the-art solutions, even though they are perturbative and restricted to specific applications.

Individual Contributions of Thesis Author: The thesis author has significantly contributed to the development and conceptual design of the research project. Moreover, the author has contributed to the gathering, collection, acquisition or provision of data, software or sources. Further, the author has significantly contributed to the analysis and evaluation or interpretation of data, sources and conclusions drawn from them. Finally, the author has contributed to the drafting of the manuscript.

4.2 Efficient Protection of Numeric Attributes

The publication **Ref.2** investigates how numeric attributes can be protected from sensitive attribute disclosure in a scalable manner. For this, it proposes various optimizations for efficiently evaluating ordered-distance t-closeness, which is one of the few privacy models for this purpose, and which is particularly computationally expensive. More precisely, algorithmic representations of the model based on directly calculating the EMD using summation (cf. Formula 2.1) for each equivalence class require an amount of computations in the order of $O(n^2)$, where n is the numbers of records.

The first optimization addresses the calculation of the distribution of values in each equivalence class required for calculating the EMD. The standard approach for this is to determine the frequency of each distinct value using hash tables. Evaluations of existing hash table implementations have shown that they perform more complex calculations than required for the given use case, even when they are already optimized towards high performance. For this reason, **Ref.2** proposes a custom-developed, efficient hash table using Fibonacci hashing based on the golden ration [Knu98].

The second optimization is based on the observation that values which are absent in a given equivalence class can lead to summands within the sum for calculating the EMD that depend only on the input dataset and can hence be pre-calculated in an initialization step (these summands have the form $|-q_1... - q_k|$ in Formula 2.1). Whenever the EMD is evaluated for a specific equivalence class, an according partial sum of this form can be retrieved. If this partial sum already exceeds a threshold depending on the privacy parameter t, then the dataset cannot satisfy t-closeness and hence the computation can already be stopped.

The third optimization can be used when the mentioned pre-calculated partial sum does not exceed the threshold. Then, this partial sum is used as a starting point of the summation process and the summation can be completed by adding the remaining summands only.

Experimental evaluations show that the proposed optimizations can significantly improve execution times, by a factor of up to two. They are particularly relevant in the context of flexible anonymization algorithms that evaluate privacy models repeatedly on data transformed in different ways to optimize data quality.

Individual Contributions of Thesis Author: The thesis author has significantly contributed to the development and conceptual design of the research project. Moreover, the author has contributed to the gathering, collection, acquisition or provision of data, software or sources. Further, the author has significantly contributed to the analysis and evaluation or interpretation of data, sources and conclusions drawn from them. Finally, the author has contributed to the drafting of the manuscript.

4.3 Reliable Data Anonymization

The paper **Ref.3** studies the effects of floating-point rounding errors on the privacy guarantees provided by various truthful data anonymization methods. It addresses straight-forward implementations of privacy models including k-anonymity, ℓ-diversity and t-closeness and the exponential mechanism. Such direct implementations are particularly important in the context of flexible anonymization algorithms that can be parameterized with (possibly combinations of) various different privacy models.

The analyses performed are based on conservative methods from the field of numerical mathematics such as forward error analyses [Hig02] to derive upper bounds for the additive exceedance of the privacy parameters of the methods considered. In most cases, the negative impacts have been found to be negligible in realistic settings. For example, the exceedance of the privacy parameter ϵ of a straight-forward floating-point implementation of the exponential mechanism is in the order of 10^{-10} or less in practical settings. However, an analysis of the algorithm proposed in **Ref.1** shows that the relatively complex computations required for calculating the parameters used for the transformation of data can lead to significant exceedances of the privacy parameter δ. It is shown that using a straight-forward floating-point implementation, relative errors of δ of up to 28% can occur in certain practically relevant settings. This happens when (floating-point approximations of) irrational values of ϵ such as $\ln(2)$ or $\ln(3)$, which are common in the literature [DEE13], are used.

To solve such issues, a flexible, reliable computing framework based on techniques described in Section 2.2.2, including exact fractional arithmetic and interval arithmetic, is proposed. It provides reliable implementations of basic arithmetic operators, elementary functions such as roots, logarithms and the exponential function as well as common comparison operators. This framework can be used to implement calculations using decimal numbers in the context of data anonymization algorithms in such a way that the actual degree of privacy protection is at least as strong as specified by the user. Due to a lack of suitable existing solutions, it has been implemented from scratch.

Extensive evaluations show that implementing flexible data anonymization which is reliable with respect to floating-point errors is feasible and that it can be achieved with negligible impacts on scalability and data quality in realistic settings.

Individual Contributions of Thesis Author: The thesis author has significantly contributed to the development and conceptual design of the research project. Moreover, the author has contributed to the gathering, collection, acquisition or provision of data, software or sources. Further, the author has significantly contributed to the analysis and evaluation or interpretation of data, sources and conclusions drawn from them. Finally, the author has contributed to the drafting of the manuscript.

CHAPTER 5

Discussion

This chapter discusses the main results presented in this thesis, relates them to prior work and highlights possible directions for future research.

5.1 Differentially Private Data Anonymization

The article **Ref.1** shows that truthful data anonymization which satisfies differential privacy is indeed feasible. To this end, it presents a differentially private mechanism for the release of truthful microdata that can be flexibly parameterized to optimize the quality of output data for different applications.

While differential privacy provides a strong degree of privacy protection, the model has been criticized for being abstract, difficult to parameterize and to explain to non-experts [DEE13]. However, due to the relatively simple, truthful modifications of data employed, the mechanism presented in **Ref.1** offers an intuitive interpretation of the privacy protection provided: With a given probability which can be derived from the privacy parameter ϵ, the data of an individual will not be contained in the output dataset at all and even if it is included, it is generalized in such a way that it cannot be distinguished from the data of at least $k-1$ other individuals, where k can be derived from the privacy parameters ϵ and δ.

Analytical and experimental evaluations have shown that the method is practical in terms of scalability and data quality and that privacy budgets in the order of $\epsilon = 1$, which are common in the literature, are a recommendable parameterization.

In the current stage of development, the mechanism employs random sampling followed by k-anonymization via full-domain generalization and record suppression (cf. Section 2.1.2), just as shown in Figure 2.4. These are relatively simple transformation techniques that often remove a significant amount of information. However, the proposed method removes information in a controlled manner that preserves frequent combinations of attribute values and hence, it preserves a significant degree of utility. In particular, experimental comparisons with prior work in **Ref.1** have shown that statistical classifiers trained on output data can compete with, and sometimes even outperform, perturbative state-of-the-art differentially private algorithms for statistical classification. It has to be noted, though, that the degree of protection provided is slightly lower compared to perturbative methods which usually satisfy strict ϵ-differential privacy rather than (ϵ, δ)-differential privacy. However, as discussed in Section 2.1.6, mechanisms which output truthful microdata cannot possibly satisfy strict ϵ-differential privacy.

The article **Ref.1** also contains an experimental comparison with the approach presented in [FEB14, FEB12] which is the most closely related as it is also a truthful, (ϵ, δ)-differentially private microdata release algorithm that employs random sampling and generalization. In these experiments, almost all information has been removed from the input datasets, which renders this competitor impractical at its current stage of development.

The work presented in **Ref.1** could be extended in several directions through future research. One interesting line of future work is the development of further score functions which optimize output data for further applications, for example, for learning tasks such as regression. To this end, the development of score functions tailored to the special-purpose quality models presented in [LDR08] appears to be promising. Moreover, it would be worthwhile to investigate how more flexible transformation techniques than full-domain generalization, such as subtree generalization, multi-dimensional generalization or cell suppression (cf. Section 2.1.2), could be integrated into the approach to further increase the quality of data provided. Along this line of future improvements, it would also be interesting to research how implicit random sampling performed during the acquisition of data could be considered in order to reduce the amount of explicit random sampling performed by the mechanism.

Finally, it is well-known that methods based on k-anonymization are suited for protecting data from low to medium dimensionality, but cannot retain sufficient data quality when processing high-dimensional data [Agg05]. For this reason, it would be interesting to investigate approaches for vertically partitioning high-dimensional datasets and processing the resulting subsets of attributes individually. This would also facilitate parallel processing, which, in turn, can reduce execution times. Other possible solutions to tackle the problem of high dimensionality could be the development of alternative differentially private search strategies. To this end, it appears promising to develop differentially private variants of the heuristic algorithms mentioned in Section 2.1.3. In this context, it also seems worthwhile to investigate how other privacy models such as k^m-anonymity, that have been proposed specifically for the protection of longitudinal data which is typically high-dimensional [TMK08], could be integrated into the mechanism.

5.2 Efficient Protection of Numeric Attributes

The paper **Ref.2** investigates how numeric attributes can be protected from sensitive attribute disclosure in an efficient and scalable manner. To this end, it presents different optimizations which can be used for evaluating the privacy model ordered-distance t-closeness that range from the implementation level to the mathematical level. They include the usage of a custom hash map and the employment of pre-calculated partial sums to prevent redundant computations when a transformed dataset satisfies t-closeness while also enabling an early termination of computations when it does not.

Experimental evaluations of the optimizations presented in **Ref.2** show that they can significantly reduce execution times in practice, by a factor of up to two. Each optimization had a positive effect in all experiments while the impact varied between different setups. In particular, the higher the number of distinct sensitive attribute values was, the better were the speedups that could be achieved. At the same time, higher numbers of distinct sensitive attribute values result in higher execution times, which implies that the proposed optimizations are the most effective when they are needed the most.

All optimizations aim to speed up the evaluation of ordered-distance t-closeness by directly computing Formula 2.1. This is required in order to integrate ordered-distance t-closeness into a flexible data anonymization algorithm as described in Section 2.1.3 that can be parameterized with (combinations of) different privacy and quality models that are directly evaluated on the transformed data. Privacy models are evaluated repeatedly in this setting on data that is transformed in different ways and therefore, their efficient evaluation is of particular importance.

Methods that have been proposed for enforcing t-closeness are usually special-purpose approaches specifically designed for this privacy model or they require the computation of the Earth Mover's Distance without providing further information on how to do this in a scalable way. (see [CKKT11,SCDF13,LLV07,LLV09] for prominent examples). To the best knowledge of the author of this thesis, **Ref.2** is the first work that specifically addresses the challenge of efficiently calculating the Earth Mover's Distance.

The article [DFSC15] studies a relationship between t-closeness and differential privacy from a theoretical perspective. It proposes a stochastic variant of t-closeness and shows that it can imply differential privacy under reasonable assumptions. It would be interesting to investigate if this privacy model could be implemented in a scalable manner using optimizations such as the ones described in **Ref.2**. Moreover, while the evaluation of t-closeness is particularly computationally expensive, it would be interesting to investigate possible optimizations for the implementations of further privacy models in future research.

5.3 Reliable Data Anonymization

The article **Ref.3** investigates the effects of floating-point rounding errors on the degree of privacy protection provided by implementations of a variety of methods for data anonymization, ranging from syntactical data anonymization to mechanisms which satisfy differential privacy. The results of the numerical analyses presented show that in most cases, violations of the expected degree of protection which may result from floating-point errors are negligible in realistic settings. Analyses of the method presented in **Ref.1**, however, show that the required computations of the parameters for the transformation of data can lead to significant exceedances of the expected privacy threshold δ due to floating-point errors.

To solve such problems, a flexible framework based on reliable computing techniques as described in Section 2.2 is proposed. It provides reliable implementation of various basic mathematical operations and functions. This framework is used to implement the method presented in **Ref.1** in such a way that the actual degree of privacy protection provided is at least as strong as requested by the user. Experimental evaluations of the resulting implementation show that the impacts of the reliable computing framework on data quality and execution times are minor when realistic parameters are used.

The results presented in **Ref.3** have implications for the results published in the articles **Ref.1** and **Ref.2** that are worth discussing. Firstly, the evaluations of the differentially private mechanism presented in **Ref.1** have been performed with an initial floating-point implementation. However, the results and conclusions presented can be transferred to the current implementation of the mechanism which uses the proposed reliable computing framework. The reason is that, as mentioned, the effects of this framework on data quality and on execution times have been found to be irrelevant when realistic parameters are being used. Moreover, the experimental evaluations and comparisons with prior work presented in **Ref.1** did not use the irrational values of ϵ which have led to the mentioned exceedances of δ in the initial floating-point implementation.

Secondly, the numerical analysis of t-closeness presented in **Ref.3** has focused on errors occurring in unoptimized implementations that directly evaluate Formula 2.1 using floating-point arithmetic and found them to be negligible in realistic settings. These results can be transferred to the optimized implementation of t-closeness presented in **Ref.2** because the optimizations employed can at most reduce the amount of floating-point operations performed. Consequently, the impacts of floating-point errors on the degree of privacy protection provided by the optimized implementation can at most improve compared to the unoptimized implementations addressed by **Ref.3**.

There exists a vast amount of general literature about floating-point arithmetic and the study of the effects of numerical approximation has formed an own line of research termed numerical analysis [Gol91, MBDD$^+$09, Mon08, Hig02, Gau97].

An example of an article which considered floating-point arithmetic specifically in the context of data anonymization is the paper [SKH16]. It proposes an implementation of ℓ-diversity using Grassberger's correction to increase the quality of data. The authors argue that the resulting formula for evaluating the model is exact to double precision.

The publication [Mir12] describes a vulnerability present in many implementations of the Laplacian mechanism. It results from the fact that common floating-point implementations actually sample from approximations of the Laplacian distribution which deviate from the mathematical model. These approximations have missing values and values which appear too frequently, which breaks the requirements of differential privacy. The article shows that this vulnerability can be exploited to extract the entire content of a database. It proposes a mitigation strategy based on rounding performed after the addition of noise.

The effects of floating-point arithmetic on the privacy guarantees provided by other approaches to data anonymization, however, have remained largely uncovered in the literature. The article **Ref.3** contributes towards filling this gap.

Natural directions for future work include numerical analyses of implementations of further data anonymization methods and the development of reliable implementations using the proposed framework when significant violations of the expected degree of privacy protection are found.

CHAPTER 6

Conclusion

This dissertation has presented several scientific contributions to the field of truthful microdata anonymization, addressing challenges that range from the degree of privacy protection provided and the scalability of solutions to the flexibility of methods. The main contributions are (1) a flexible, truthful anonymization algorithm that satisfies differential privacy, (2) multiple optimizations for protecting numeric attributes in an efficient manner, and (3) an analysis of the negative impacts of floating-point rounding errors on the degree of protection provided by implementations of several data anonymization methods as well as a reliable computing framework to mitigate those adverse impacts whenever required.

A primary motivation of the work presented was to support making data anonymization readily applicable in practice, with a particular focus on applications in the medical domain, where the truthfulness of methods is of particular importance. However, it is worth pointing out that the proposed results are universally applicable and not restricted to a single application domain. To challenge their practicability, all solutions presented have been extensively evaluated using several datasets of different sizes. To make the proposed results available to end-users, they have been integrated into the open-source data anonymization tool ARX, where they can be flexibly combined with many other anonymization methods already implemented and hopefully many more to come in the future.

The integration of increasing numbers of anonymization methods with different advantages and disadvantages that can be flexibly combined in a way that is easy to use is ongoing work. The overarching goal is to support appropriate anonymization in a possibly broad range of application scenarios to achieve adequate privacy protection, typically using a combination of several measures ranging from the technical to the organizational level. This is of increasing importance in times when the amount of personal data being collected and the demand to process it is continuously growing.

CHAPTER 7

Bibliography

[Age] European Medicines Agency. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/clinical-data-publication/support-industry/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data. [Online; accessed 26-Oct-2023].

[Agg05] Charu C Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of 31st International Conference on Very Large Data Bases*, pages 901–909, 2005.

[BA05] Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *International Conference on Data Engineering*, pages 217–228, 2005.

[BCD+07] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 273–282, 2007.

[Bis03] Matt Bishop. Computer security: Art and science. *Westford, MA: Addison Wesley Professional*, 2003.

[BKBL07] Ji-Won Byun, Ashish Kamra, Elisa Bertino, and Ninghui Li. Efficient k-anonymization using clustering techniques. In *International Conference on Database Systems for Advanced Applications*, pages 188–200, 2007.

[BRK$^+$13] Korra Sathya Babu, Nithin Reddy, Nitesh Kumar, Mark Elliot, and Sanjay Kumar Jena. Achieving k-anonymity using improved greedy heuristics for very large relational databases. *Transactions on Data Privacy*, 6(1):1–17, 2013.

[CCZYM17] Constantinos Costa, Georgios Chatzimilioudis, Demetrios Zeinalipour-Yazti, and Mohamed F Mokbel. Efficient exploration of telco big data with compression and decaying. In *IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 1332–1343, 2017.

[CDN15] Ronald Cramer, Ivan Bjerre Damgård, and Jesper Buus Nielsen. *Secure multiparty computation*. Cambridge University Press, 2015.

[CK12] Jianneng Cao and Panagiotis Karras. Publishing microdata with a robust privacy guarantee. *Proceedings of the VLDB Endowment*, 5(11):1388–1399, 2012.

[CKKT11] Jianneng Cao, Panagiotis Karras, Panos Kalnis, and Kian-Lee Tan. SABRE: A sensitive attribute bucketization and redistribution framework for t-closeness. *The VLDB Journal*, 20(1):59–81, 2011.

[Cou16] 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. *Official Journal of the European Union*, 59(L119):1–88, 2016.

[CT13] Chris Clifton and Tamir Tassa. On syntactic anonymity and differential privacy. In *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 88–93, 2013.

[Daw11] Hend Dawood. *Theories of interval arithmetic: Mathematical foundations and applications*. Lambert Academic Publishing, 2011.

[DE12] Fida K Dankar and Khaled El Emam. The application of differential privacy to health data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 158–166, 2012.

[DEE13] Fida K Dankar and Khaled El Emam. Practicing differential privacy in health care: A review. *Transactions on Data Privacy*, 6(1):35–67, 2013.

[DESG11] George T Duncan, Mark Elliot, and Juan-José Salazar-González. *Statistical confidentiality: Principles and practice.* Springer, 2011.

[DFMS02] Josep Domingo-Ferrer and Josep Maria Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.

[DFSC15] Josep Domingo-Ferrer and Jordi Soria-Comas. From t-closeness to differential privacy and vice versa in data anonymization. *Knowledge-Based Systems*, 74:151–158, 2015.

[DFT05] Josep Domingo-Ferrer and Vicenç Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.

[DFT08] Josep Domingo-Ferrer and Vicenç Torra. A critique of k-anonymity and some of its enhancements. In *Third International Conference on Availability, Reliability and Security*, pages 990–993, 2008.

[DKM$^+$06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503, 2006.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284, 2006.

[DP02] Brian A Davey and Hilary A Priestley. *Introduction to lattices and order.* Cambridge University Press, 2002.

[DPDA$^+$16] Stephanie OM Dyke, Anthony A Philippakis, Jordi Rambla De Argila, Dina N Paltoo, Erin S Luetkemeier, Bartha M Knoppers, Anthony J Brookes, J Dylan Spalding, Mark Thompson, Marco Roos, et al. Consent codes: Upholding standard data use conditions. *PLoS Genetics*, 12(1):e1005772, 2016.

[DR13] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2013.

[DW99] Ton DeWaal and Leon Willenborg. Information loss through global recoding and local suppression. *Netherlands Official Statistics*, 14:17–20, 1999.

[Dwo06] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming*, pages 1–12. Springer, 2006.

[Dwo08] Cynthia Dwork. An ad omnia approach to defining and achieving private data analysis. In *International Conference on Privacy, Security, and Trust in KDD*, pages 1–13, 2008.

[Dwo11] Cynthia Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011.

[EBS$^+$20] Johanna Eicher, Raffael Bild, Helmut Spengler, Klaus A Kuhn, and Fabian Prasser. A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models. *BMC Medical Informatics and Decision Making*, 20(1):1–14, 2020.

[EEA13] Khaled El Emam and Luk Arbuckle. *Anonymizing health data: Case studies and methods to get you started.* O'Reilly Media, 2013.

[EEDI$^+$09] Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, et al. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5):670–682, 2009.

[EJMA13] Khaled El Emam, Elizabeth Jonker, Ester Moher, and Luk Arbuckle. A review of evidence on consent bias in research. *The American Journal of Bioethics*, 13(4):42–44, 2013.

[EKP17] Johanna Eicher, Klaus A Kuhn, and Fabian Prasser. An experimental comparison of quality models for health data de-identification. *Studies in Health Technology and Informatics*, 245:704–708, 2017.

[FEB12] Mohamed R Fouad, Khaled Elbassioni, and Elisa Bertino. Towards a differentially private data anonymization. CERIAS Tech Report 2012-1, Center for Education and Research, Purdue University, 2012.

[FEB14] Mohamed R Fouad, Khaled Elbassioni, and Elisa Bertino. A supermodularity-based differential privacy preserving algorithm for data anonymization. *IEEE Transactions on Knowledge and Data Engineering*, 26(7):1591–1601, 2014.

[FJ15] Liyue Fan and Hongxia Jin. A practical framework for privacy-preserving data analytics. In *International Conference on World Wide Web*, pages 311–321, 2015.

[FWFP10] Benjamin CM Fung, Ke Wang, Ada Wai-Chee Fu, and S Yu Philip. *Introduction to privacy-preserving data publishing: Concepts and techniques.* Chapman & Hall/CRC Data Mining and Knowledge Discovery, 2010.

[Gau97] Walter Gautschi. *Numerical analysis.* Springer Science & Business Media, 1997.

[Gen09] Craig Gentry. *A fully homomorphic encryption scheme.* PhD thesis, Stanford University, 2009.

[GHLP12] Johannes Gehrke, Michael Hay, Edward Lui, and Rafael Pass. Crowd-blending privacy. In *Annual Cryptology Conference*, pages 479–496, 2012.

[GM17] Puneet Goswami and Suman Madan. Privacy preserving data publishing and data anonymization approaches: A review. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 139–142, 2017.

[GMT08] Aristides Gionis, Arnon Mazza, and Tamir Tassa. k-Anonymization revisited. In *IEEE 24th International Conference on Data Engineering*, pages 744–753, 2008.

[GMW$^+$12] Michaela Gotz, Ashwin Machanavajjhala, Guozhang Wang, Xiaokui Xiao, and Johannes Gehrke. Publishing search logs—a comparative study of privacy guarantees. *IEEE Transactions on Knowledge and Data Engineering*, 24(3):520–532, 2012.

[Gol91] David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys (CSUR)*, 23(1):5–48, 1991.

[GT09] Jacob Goldberger and Tamir Tassa. Efficient anonymizations with enhanced utility. In *IEEE International Conference on Data Mining Workshops*, pages 106–113, 2009.

[Hig02] Nicholas J Higham. *Accuracy and stability of numerical algorithms.* Society for Industrial and Applied Mathematics, 2002.

[IEE08] IEEE Standards Committee et al. 754-2008 IEEE standard for floating-point arithmetic. *IEEE Computer Society Standard*, 2008.

[Iye02] Vijay S Iyengar. Transforming data to satisfy privacy constraints. In *International Conference on Knowledge Discovery and Data Mining*, pages 279–288, 2002.

[JBD+21] Carolin EM Jakob, Stefan Borgmann, Fazilet Duygu, Uta Behrends, Martin Hower, Uta Merle, Anette Friedrichs, et al. First results of the "Lean European Open Survey on SARS-CoV-2-Infected Patients (LEOSS)". *Infection*, 49(1):63–73, 2021.

[JLE14] Zhanglong Ji, Zachary Chase Lipton, and Charles Elkan. Differential privacy and machine learning: A survey and review. *CoRR*, abs/1412.7584, 2014.

[JYC15] Zach Jorgensen, Ting Yu, and Graham Cormode. Conservative or liberal? Personalized differential privacy. In *IEEE International Conference on Data Engineering*, pages 1023–1034, 2015.

[KHC+16] Jinkyu Kim, Heonseok Ha, Byung-Gon Chun, Sungroh Yoon, and Sang K Cha. Collaborative analytics for data silos. In *IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 743–754, 2016.

[KHZ17] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open university learning analytics dataset. *Scientific Data*, 4(1):1–8, 2017.

[KL20] Jonathan Katz and Yehuda Lindell. *Introduction to modern cryptography*. CRC Press, 2020.

[Knu98] Donald E Knuth. *The Art of Computer Programming: Sorting and Searching*. Addison Wesley, 1998.

[Koh16] Florian Kohlmayer. *Datenschutz und biomedizinische Forschung: Konzepte und Lösungen für Anonymität*. PhD thesis, Technische Universität München, 2016.

[KPE+12] Florian Kohlmayer, Fabian Prasser, Claudia Eckert, Alfons Kemper, and Klaus A Kuhn. Flash: Efficient, stable and optimal k-anonymity. In *International Conference on Privacy, Security, Risk and Trust (PASSAT) and International Conference on Social Computing (SocialCom)*, pages 708–717, 2012.

[KS08] Shiva Prasad Kasiviswanathan and Adam Smith. A note on differential privacy: Defining resistance to arbitrary side information. *CoRR*, abs/0803.3946, 2008.

[Lam93] Diane Lambert. Measures of disclosure risk and harm. *Journal of Official Statistics – Stockholm*, 9:313–313, 1993.

[LDR05] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *International Conference on Management of Data*, pages 49–60, 2005.

[LDR06] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *International Conference on Data Engineering*, pages 25–25, 2006.

[LDR08] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Workload-aware anonymization techniques for large-scale datasets. *ACM Transactions on Database Systems*, 33(3):1–47, 2008.

[LLV07] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-Closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd Interational Conference on Data Engineering*, 2007.

[LLV09] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. Closeness: A new privacy measure for data publishing. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):943–956, 2009.

[LQS11] Ninghui Li, Wahbeh H Qardaji, and Dong Su. Provably private data anonymization: Or, k-anonymity meets differential privacy. *CoRR*, abs/1101.2604, 2011.

[LQS12] Ninghui Li, Wahbeh Qardaji, and Dong Su. On sampling, anonymization, and differential privacy: Or, k-anonymization meets differential privacy. In *ACM Symposium on Information, Computer and Communications Security*, pages 32–33, 2012.

[MBDD$^+$09] Jean-Michel Muller, Nicolas Brisebarre, Florent De Dinechin, Claude-Pierre Jeannerod, Vincent Lefevre, Guillaume Melquiond, Nathalie Revol, Damien Stehlé, Serge Torres, et al. *Handbook of floating-point arithmetic*. Springer Science & Business Media, 2009.

[MBDP21] Thierry Meurers, Raffael Bild, Kieu-Mi Do, and Fabian Prasser. A scalable software solution for anonymizing high-dimensional biomedical data. *GigaScience*, 10(10):giab068, 2021.

[McS09] Frank D McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *International Conference on Management of Data*, pages 19–30, 2009.

[MEO16] Elaine Mackey, Mark Elliot, and Kieron O'Hara. *The anonymisation decision-making framework*. UKAN Publications, 2016.

[Mir12] Ilya Mironov. On significance of the least significant bits for differential privacy. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, pages 650–661, 2012.

[MKGV07] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-Diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[Mon08] David Monniaux. The pitfalls of verifying floating-point computations. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 30(3):1–41, 2008.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *IEEE Symposium on Foundations of Computer Science*, pages 94–103, 2007.

[NAC07] Mehmet Ercan Nergiz, Maurizio Atzori, and Chris Clifton. Hiding the presence of individuals from shared databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 665–676, 2007.

[NC07] M Ercan Nergiz and Chris Clifton. Thoughts on k-anonymization. *Data & Knowledge Engineering*, 63(3):622–645, 2007.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

[NT15] Guillermo Navarro-Arribas and Vicenç Torra. *Advanced Research in Data Privacy*. Springer, 2015.

[Off] Office for National Statistics. Statistical disclosure control for 2011 census. http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/statistical-methodology/statistical-disclosure-control-for-2011-census.pdf. [Online; accessed 24-Oct-2023].

[oHfCR02] US Department of Health and Human Services Office for Civil Rights. Standards for privacy of individually identifiable health information. final rule. *Federal Register*, 67(157):53181, 2002.

[Opf93] Gerhard Opfer. *Numerische Mathematik für Anfänger*. Springer, 1993.

[oS] Personal Data Protection Commission of Singapore. Guide to basic data anonymization techniques. https://iapp.org/resources/article/guide-to-basic-data-anonymization-techniques/. [Online; accessed 26-Oct-2023].

[PBE$^+$16] Fabian Prasser, Raffael Bild, Johanna Eicher, Helmut Spengler, Florian Kohlmayer, and Klaus A Kuhn. Lightning: Utility-driven anonymization of high-dimensional data. *Transactions on Data Privacy*, 9(2):161–185, 2016.

[PES$^+$20] Fabian Prasser, Johanna Eicher, Helmut Spengler, Raffael Bild, and Klaus A Kuhn. Flexible data anonymization using arx—current status and challenges ahead. *Software: Practice and Experience*, 50(7):1277–1304, 2020.

[PGDL$^+$14] Giorgos Poulis, Aris Gkoulalas-Divanis, Grigorios Loukides, Spiros Skiadopoulos, and Christos Tryfonopoulos. SECRETA: A system for evaluating and comparing relational and transaction anonymization algorithms. In *International Conference on Extending Database Technology*, pages 620–623, 2014.

[PGW$^+$17] Fabian Prasser, James Gaupp, Zhiyu Wan, Weiyi Xia, Yevgeniy Vorobeychik, Murat Kantarcioglu, Klaus A Kuhn, and Brad Malin. An open source tool for game theoretic health data de-identification. In *AMIA Annual Symposium Proceedings*, pages 1430–1439, 2017.

[PK15] Fabian Prasser and Florian Kohlmayer. Putting statistical disclosure control into practice: The ARX data anonymization tool. In *Medical Data Privacy Handbook*, pages 111–148. Springer, 2015.

[PKK16a] Fabian Prasser, Florian Kohlmayer, and Klaus A. Kuhn. Efficient and effective pruning strategies for health data de-identification. *BMC Medical Informatics and Decision Making*, 16(1):1–14, 2016.

[PKK16b] Fabian Prasser, Florian Kohlmayer, and Klaus A Kuhn. The importance of context: Risk-based de-identification of biomedical data. *Methods of Information in Medicine*, 55(4):347–355, 2016.

[PLGDS13] Giorgos Poulis, Grigorios Loukides, Aris Gkoulalas-Divanis, and Spiros Skiadopoulos. Anonymizing data with relational and transaction attributes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 353–369, 2013.

[RE15] Felix Ritchie and Mark Elliott. Principles- versus rules-based output statistical disclosure control in remote access environments. *IASSIST Quarterly*, 39(2):5–13, 2015.

[RMGM08] Aric Rindfleisch, Alan J Malter, Shankar Ganesan, and Christine Moorman. Cross-sectional versus longitudinal survey research: Concepts, findings, and guidelines. *Journal of Marketing Research*, 45(3):261–279, 2008.

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[SCDF13] Jordi Soria-Comas and Josep Domingo-Ferrert. Differential privacy via t-closeness in data publishing. In *IEEE Conference on Privacy, Security and Trust*, pages 27–35, 2013.

[SCDFSM15] Jordi Soria-Comas, Josep Domingo-Ferrer, David Sanchez, and Sergio Martinez. t-Closeness through microaggregation: Strict privacy with enhanced utility preservation. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):3098–3110, 2015.

[Sha01] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[SKH16] Sebastian Stammler, Stefan Katzenbeisser, and Kay Hamacher. Correcting finite sampling issues in entropy l-diversity. In *International Conference on Privavy in Statistical Databases*, pages 135–146, 2016.

[Sta18] Standardization Administration of China. GB/T 35273-2017 Information Technology – Personal Information Security Specification, 2018.

[Swe97] Latanya Sweeney. Datafly: A system for providing anonymity in medical data. In *Proceedings of the 11th International Conference on Database Security*, 1997.

[Swe02a] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.

[Swe02b] Latanya Sweeney. k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[TMK08] Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. Privacy-preserving anonymization of set-valued data. *Proceedings of the VLDB Endowment*, 1(1):115–125, 2008.

[USMN17] Giske Ursin, Sagar Sen, Jean-Marie Mottu, and Mari Nygård. Protecting privacy in large datasets—first we assess the risk; then we fuzzy the data. *Cancer Epidemiology and Prevention Biomarkers*, 26(8):1219–1224, 2017.

[VSBH13] Jaideep Vaidya, Basit Shafiq, Anirban Basu, and Yuan Hong. Differentially private naive bayes classification. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, pages 571–576, 2013.

[WDW96] Leon Willenborg and Ton De Waal. *Statistical disclosure control in practice.* Springer Science & Business Media, 1996.

[WE05] Ian H Witten and Frank Eibe. *Data mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2005.

[WE18] Isabel Wagner and David Eckhoff. Technical privacy metrics: A systematic survey. *ACM Computing Surveys (CSUR)*, 51(3):1–38, 2018.

[Whi04] Stephen A White. Introduction to BPMN. *IBM Cooperation*, 2004.

[Wil94] James Hardy Wilkinson. *Rounding errors in algebraic processes.* Courier Corporation, 1994.

[WVX$^+$15] Zhiyu Wan, Yevgeniy Vorobeychik, Weiyi Xia, Ellen Wright Clayton, Murat Kantarcioglu, Ranjit Ganta, Raymond Heatherly, and Bradley A Malin. A game theoretic framework for analyzing re-identification risk. *PloS One*, 10(3):e0120592, 2015.

[WVX$^+$17] Zhiyu Wan, Yevgeniy Vorobeychik, Weiyi Xia, Ellen Wright Clayton, Murat Kantarcioglu, and Bradley Malin. Expanding access to large-scale genomic data while promoting privacy: A game theoretic approach. *The American Journal of Human Genetics*, 100(2):316–322, 2017.

[WYC04] Ke Wang, Philip S Yu, and Sourav Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *Fourth IEEE International Conference on Data Mining*, pages 249–256, 2004.

[XJC$^+$15] Lei Xu, Chunxiao Jiang, Yan Chen, Yong Ren, and KJ Ray Liu. Privacy or utility in data collection? A contract theoretic approach. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1256–1269, 2015.

[XT07] Xiaokui Xiao and Yufei Tao. m-Invariance: Towards privacy preserving re-publication of dynamic datasets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 689–700, 2007.

[YJ14] Fei Yu and Zhanglong Ji. Scalable privacy-preserving data sharing methodology for genome-wide association studies: An application to iDASH healthcare privacy protection challenge. *BMC Medical Informatics and Decision Making*, 14(1):1–8, 2014.

APPENDIX A

Contributed Original Works

Contents

A.1 Differentially Private Data Anonymization

Full Title

SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees

Authors

Raffael Bild, Klaus A. Kuhn and Fabian Prasser

Published In

Proceedings on Privacy Enhancing Technologies, 2018(1):67-87, 2018

Copyright

CC BY-NC-ND 3.0 (see Appendix C for legal code. Source: `https://creativecommons.org/licenses/by-nc-nd/3.0/` Accessed: 11.07.2024)

Raffael Bild*, Klaus A. Kuhn, and Fabian Prasser

SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees

Abstract: Methods for privacy-preserving data publishing and analysis trade off privacy risks for individuals against the quality of output data. In this article, we present a data publishing algorithm that satisfies the differential privacy model. The transformations performed are truthful, which means that the algorithm does not perturb input data or generate synthetic output data. Instead, records are randomly drawn from the input dataset and the uniqueness of their features is reduced. This also offers an intuitive notion of privacy protection. Moreover, the approach is generic, as it can be parameterized with different objective functions to optimize its output towards different applications. We show this by integrating six well-known data quality models. We present an extensive analytical and experimental evaluation and a comparison with prior work. The results show that our algorithm is the first practical implementation of the described approach and that it can be used with reasonable privacy parameters resulting in high degrees of protection. Moreover, when parameterizing the generic method with an objective function quantifying the suitability of data for building statistical classifiers, we measured prediction accuracies that compare very well with results obtained using state-of-the-art differentially private classification algorithms.

Keywords: Data privacy, differential privacy, anonymization, disclosure control, classification.

DOI 10.1515/popets-2018-0004
Received 2017-05-31; revised 2017-09-15; accepted 2017-09-16.

1 Introduction

There is a strong tension between opportunities to leverage ever-growing collections of sensitive personal data for business or research on one hand, and potential dangers to the privacy of individuals on the other. Methods for privacy-preserving data publishing and analysis aim to find a balance between these conflicting goals by trading privacy risks off against the quality of data [2].

Data published by statistical agencies usually describes a sample from a specific population. The sampling process performed during data acquisition, as well as additional random sampling sometimes performed before data is released, provides an intuitive but weak notion of privacy protection [53]. In addition, statistical data is typically sanitized using methods of *disclosure control* which includes modifying, summarizing, or perturbing (i.e. randomizing) the data. In this process, “principles-based” approaches defined by experts and rules of thumb are typically used [50].

An additional line of research, which we will call *data anonymization*, has formulated *syntactic requirements* for mitigating risks in the form of *privacy models.* The most well-known model is k*-anonymity*, which requires that each record in a dataset is indistinguishable from at least $k - 1$ other records regarding attributes which could be used for re-identification attacks [52].

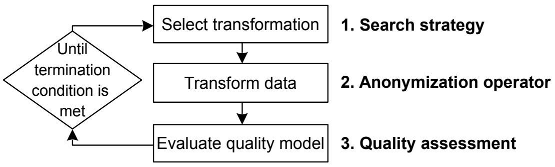


Fig. 1. Common components of data anonymization algorithms.

Based on such formal requirements, privacy protection can be implemented with *anonymization algorithms* which transform data to ensure that the requirements are met while reductions in data quality are quantified and minimized [2]. As is sketched in Figure 1, anonymization algorithms can be modelled as a process in which a set of available data transformations is being *searched*, while an anonymization operator is used to make sure that privacy requirements are *satisfied* and quality is *assessed* to guide the search process. We emphasize that this is a very high-level overview and that the design of concrete algorithms often depends on the privacy models, quality models, and, most importantly, the types of data transformation implemented.

Differential privacy [10] takes a different approach to privacy protection, as it is not a property of a dataset,

***Corresponding Author: Raffael Bild:** Technical University of Munich, Germany, E-mail: raffael.bild@tum.de
Klaus A. Kuhn: Technical University of Munich, Germany, E-mail: klaus.kuhn@tum.de
Fabian Prasser: Technical University of Munich, Germany, E-mail: fabian.prasser@tum.de

BY-NC-ND

but a property of a data processing method. Informally, it guarantees that the probability of any possible output of a probabilistic algorithm (called *mechanism*) does not change "by much" if data of an individual is added to or removed from input data. Implementing differential privacy does not require making strong assumptions about the background knowledge of attackers, e.g. about which attributes could be used for re-identification. Moreover, differential privacy provides strong protection, while syntactic models are much less reliable [9].

Differential privacy, however, has also been criticized for various reasons. First, implementations are often non-truthful, i.e. perturbative, as they rely on noise addition [5, 6]. Truthfulness can be a desirable property in many fields [3]. Examples include governmental or industrial applications [21] and the medical domain, in which implausible data created by perturbation (e.g. combinations or dosages of drugs which are harmful for a patient) have led to challenges for introducing noise-based mechanisms [6]. Second, the semantics of differential privacy are complex and it has been argued that the approach is much more difficult to explain to decision makers, e.g. to ethics committees and policy makers, than the idea of *hiding in a crowd* often implemented by syntactic models [6]. Finally, differentially private mechanisms are typically special-purpose algorithms developed for specific applications, see e.g. [17, 31, 32]. Many of them serve the *interactive* scenario, i.e. they provide perturbed answers to (limited sets of) queries. In contrast, microdata publishing methods aim to release a sanitized dataset that supports a variety of use cases. The development of such *non-interactive* methods which satisfy differential privacy while retaining sufficient data quality has remained challenging.

1.1 Contributions and Outline

Previous work has shown that algorithms which draw a random sample of data followed by k-anonymization can fulfill differential privacy [26, 39, 40]. These results are notable, as they combine statistical disclosure control, data anonymization and differential privacy.

In this article, we build upon this approach to implement a traditional data anonymization algorithm (see Figure 1) with differentially private components. The result is a practical method for non-interactive microdata publishing that fulfills differential privacy. The method is truthful, as randomization is implemented via sampling only and attribute values are transformed with truthful methods. Moreover, it is intuitive, as privacy is protected by sampling records and reducing the uniqueness of their features. Finally, the approach employs a flexible search strategy which can be parameterized with a wide variety of data quality models to optimize its output towards different applications. While developing the approach, we had to overcome multiple challenges.

On the theoretical level, we have completed and extended the proofs presented in [39] and [40] to develop a method for obtaining the exact privacy guarantees obtained by the approach instead of loose upper bounds. This enabled us to strengthen a theorem about the privacy guarantees provided, to study the relationships between random sampling, k-anonymization and differential privacy in more detail and to show that the approach can be used with reasonable parameterizations providing strong privacy protection. Moreover, we have transformed six common data quality models into a form suitable for integration into the approach.

On the practical level, we have performed an extensive experimental evaluation and a comparison with related work. We have evaluated general-purpose data quality and, as an application example, performed experiments with statistical classification. Our evaluation shows that the approach is practical in terms of runtime complexity and output quality. Moreover, when our generic method is parameterized with an according data quality model, it can be used to create classifiers which are en-par with, and sometimes significantly outperform, state-of-the-art approaches. This is notable, as these competitors are pertubative special-purpose implementations of the differential privacy model.

The remainder of this paper is structured as follows: We provide background information in Section 2. Then, we give a high-level overview of the method in Section 3. The anonymization operator is presented in Section 4. Section 5 describes the objective functions. In Section 6 we introduce the search strategy. Section 7 presents analytical evaluations of the method. In Section 8 we present results of experimental analyses, including comparisons with related approaches. Section 9 reviews related work. Section 10 concludes and summarizes this article and Section 11 discusses future work.

2 Background and Preliminaries

2.1 Dataset

For an arbitrary dataset D with m attributes we will denote the domains of attribute 1 to m by $\Omega_1, ..., \Omega_m$. Then, we can regard D to be a multiset $D \subseteq \Omega_1 \times ... \times \Omega_m$, and we will denote the universe of all datasets $D \subseteq \Omega_1 \times ... \times \Omega_m$ with $\mathcal{D}_m$. Analogously to other articles we will assume that each individual who contributed data to a dataset is represented by exactly one record $r = (r_1, ..., r_m) \in D$ and refer to such datasets as *microdata.*

2.2 Transformation Models

Data anonymization is typically performed by reducing the distinguishability of records. Common methods for doing so are clustering and aggregation of data items [28], the introduction of noise and the generalization and suppression of attribute values [2].

In this paper we focus on attribute generalization through user-specified hierarchies, which describe rules for replacing values with more general but semantically consistent values on increasing *levels* of generalization. Figure 2 shows two examples.

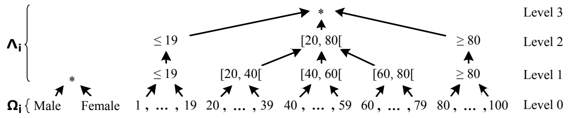


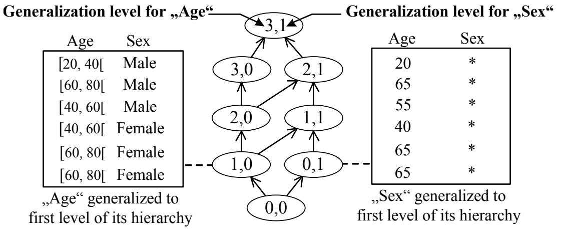
Fig. 2. Example generalization hierarchies.

Without loss of generality we will assume that a generalization hierarchy is provided for each attribute $i = 1, ..., m$ so that the values on level 0 form the domain Ω_i while we denote the set of values on levels greater than 0 by Λ_i. For a given value $r'_i \in \Omega_i \cup \Lambda_i$ we will call each value on level 0 which is an element of the subtree rooted at r'_i a *leaf node* of r'_i. For example, the leaf nodes of "$[20, 80[$" in Figure 2 are "20", ..., "79". We will indicate the removal of a record by replacing it with the placeholder $* = (*, ..., *)$. Since generalizing a value to the highest level effectively suppresses the value we will also denote the root values of all hierarchies with $*$.

2.3 Solution Spaces and Search Strategies

Most anonymization algorithms can be described as search algorithms through all possible outputs defined by the data transformation model. While they are obviously not always implemented this way (e.g. clustering algorithms typically use heuristics to guide the clustering process [28]) search algorithms are often implemented in combination with generalization hierarchies. The exact nature of the search space then depends on the generalization method.

For example, *full-domain* generalization generalizes all values of an attribute to the same level. With *subtree* generalization different values of an attribute can be generalized to different levels [2]. In this article we will focus on full-domain generalization, which results in search spaces that can be described with *generalization lattices.* An example is shown in Figure 3. An arrow denotes that a transformation is a direct *successor* of a more specialized transformation, i.e. it can be derived from its *predecessor* by incrementing the generalization level of exactly one attribute. The number of transformations in a generalization lattice grows exponentially with the number of attributes [15] and a wide variety of globally-optimal and heuristic search algorithms for generalization lattices have been proposed [15, 33–35].


Generalization level for „Age“
Generalization level for „Sex“
3,1
3,0
2,1
2,0
1,1
1,0
0,1
0,0
„Age“ generalized to first level of its hierarchy
„Sex“ generalized to first level of its hierarchy

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

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

Fig. 3. Example generalization lattice and output datasets.

In this article we will use the following notion. A *generalization scheme* is a function $g : \Omega_1 \times ... \times \Omega_m \to (\Omega_1 \cup \Lambda_1) \times ... \times (\Omega_m \cup \Lambda_m)$ mapping records to (possibly) generalized records. Obviously, every transformation which performs full-domain generalization can be formalized as a generalization scheme. Unless otherwise noted, we define the solution space $\mathcal{G}_m$ to be the set of all full-domain generalization schemes which is determined by the generalization hierarchies of all attributes of a given dataset.

2.4 Anonymization Operators

An anonymization operator implements a privacy model. It determines whether or not a record or dataset satisfies the privacy requirements and may also modify the data. For example, in clustering algorithms, the anonymization operator may merge the records within a cluster into an equivalence class that satisfies k-anonymity, which we define as follows:

Definition 1 (k-Anonymity [52])**.** For a given dataset $D \subseteq (\Omega_1 \cup \Lambda_1) \times ... \times (\Omega_m \cup \Lambda_m)$, we define an *equivalence class* $E \subseteq D$ to be the multiset of all records in D which share a given combination of attribute values. An equivalence class E satisfies k-anonymity if $|E| \geq k$ holds. D satisfies k-anonymity if each record $r \in D$ cannot be distinguished from at least $k - 1$ other records, i.e. if all equivalence classes $E \subseteq D$ are k-anonymous.

As a part of algorithms implementing full-domain generalization, the anonymization operator typically suppresses records which do not satisfy the privacy requirements [52]. This principle can not only be implemented for k-anonymity but also for further privacy models, including l-diversity [1], t-closeness [28] and δ-presence [47], which have been proposed for protecting data from threats that go beyond re-identification.

2.5 Quality Assessment

Measuring reductions in data quality due to anonymization is non-trivial as usefulness depends on the use case.

When it is unknown in advance how the data will be used, *general-purpose* quality models can be employed. They typically estimate data quality by quantifying the amount of information loss, e.g. by measuring similarities between the input and the output dataset [2]. Models can roughly be classified as measuring information loss on the attribute-level, cell-level or record-level. Typical examples for changes on these levels are differences in the distributions of attribute values (attribute-level), reductions in the granularity of data (cell-level) and differences in the sizes of equivalence classes (record-level).

Special-purpose (or *workload-aware*) quality models quantify data quality for a specific application scenario, e.g. statistical classification. Thereby the task is to predict the value of a predefined *class attribute* from a given set of values of *feature attributes*. This is implemented with *supervised learning* where a model is created from a *training set* [54]. Specific quality models have been developed for optimizing data for this purpose [25, 37].

2.6 Differential Privacy

Differential privacy requires that any output of a mechanism is almost as likely, independent of whether or not a record is present in the input dataset [10]. (ϵ, δ)-*Differential privacy* can be formally defined with respect to two datasets D_1 and D_2 satisfying $|D_1 \oplus D_2| = 1$, which means that D_2 can be obtained from D_1 by either adding or removing one record:

Definition 2 ((ϵ, δ)-differential privacy [6])**.** A randomized function $\mathcal{K}$ provides (ϵ, δ)-differential privacy if for all datasets $D_1, D_2 \in \mathcal{D}_m$ with $|D_1 \oplus D_2| = 1$, and all measurable $S \subseteq \mathit{Range}(\mathcal{K})$,

$$P[\mathcal{K}(D_1) \in S] \leq \exp(\epsilon) \cdot P[\mathcal{K}(D_2) \in S] \qquad (1)$$

holds with a probability of at least $1 - \delta$.

$(\epsilon, 0)$-Differential privacy is usually just called ϵ-differential privacy. For $\delta > 0$, (ϵ, δ)-differential privacy is then a relaxation of ϵ-differential privacy.

Sequences of differentially private computations are also differentially private:

Theorem 1. *For $i = 1, ..., n$, let the mechanism $\mathcal{M}_i$ provide ϵ_i-differential privacy. Then the sequence $\mathcal{M}_1^{r_1}(D), ..., \mathcal{M}_n^{r_n}(D)$, where $\mathcal{M}_i^{r_i}$ denotes mechnism $\mathcal{M}_i$ supplied with the outcomes of $\mathcal{M}_1, ..., \mathcal{M}_{i-1}$, satisfies $\left(\sum_{i=1}^{n} \epsilon_i\right)$-differential privacy [45].*

A common method to achieve differentially privacy is the *exponential mechanism* [44]. It ranks all potential outputs $r \in \mathcal{R}$ for a given input dataset D using a real-valued *score function* s. It then randomly chooses one according to a specific probability distribution which assigns higher probabilities to outputs with higher scores:

Definition 3 (Exponential mechanism [44])**.** For any function $s : (\mathcal{D}_m \times \mathcal{R}) \to \mathbb{R}$, the *exponential mechanism* $\mathcal{E}_s^{\epsilon}(D, \mathcal{R})$ chooses and outputs an element $r \in \mathcal{R}$ with probability proportional to $\exp\left(\frac{s(D,r)\epsilon}{2\Delta s}\right)$, where the *sensitivity* Δs of the function s is defined as

$$\Delta s := \max_{r \in \mathcal{R}} \max_{D_1, D_2 \in \mathcal{D}_m : |D_1 \oplus D_2| = 1} |s(D_1, r) - s(D_2, r)|.$$

It can be seen that it is important to use score functions which assign higher scores to outputs with higher quality while having a low sensitivity. The privacy guarantees provided are as follows:

Theorem 2. *For any function $s : (\mathcal{D}_m \times \mathcal{R}) \to \mathbb{R}$, $\mathcal{E}_s^{\epsilon}(D, \mathcal{R})$ satisfies ϵ-differential privacy [44].*

3 Overview of the Approach

Prior work has shown that randomization via sampling can be used to achieve (ϵ, δ)-differentially privacy [26, 39, 40]. We build upon and extend these results to implement the *SafePub* algorithm. It comprises a search strategy, an anonymization operator and various methods for quality assessment, similar to many anonymization algorithms. The overall privacy budget ϵ is split up into two parts ϵ_{anon}, which is used by the anonymization operator, and ϵ_{search}, which is used by the search strategy. SafePub satisfies $(\epsilon_{anon} + \epsilon_{search}, \delta)$-differential privacy, where δ and the number of iterations performed by the search strategy (*steps*) can also be specified.

Figure 4 shows the high-level design of the approach. It also indicates the parameters which are relevant for the individual steps. First, SafePub performs pre-processing by random sampling, selecting each

```
Input: Dataset $D$, Parameters $\epsilon_{anon}$, $\epsilon_{search}$, $\delta$, $steps$
Output: Dataset $S$
 1: Draw a random sample $D_s$ from $D$                        ▷ $(\epsilon_{anon})$
 2: Initialize set of transformations $G$
 3: for (Int $i \leftarrow 1, ..., steps$) do
 4:     Update $G$
 5:     for ($g \in G$) do
 6:         Anonymize $D_s$ using $g$                          ▷ $(\epsilon_{anon}, \delta)$
 7:         Assess quality of resulting data
 8:     end for
 9:     Probabilistically select solution $g \in G$            ▷ $(\epsilon_{search})$
10: end for
11: return Dataset $D_s$ anonymized using                      ▷ $(\epsilon_{anon}, \delta)$
    the best solution selected in Line 9
```

Fig. 4. High-level design of the SafePub mechanism. The search strategy is implemented by the loop in lines 3 to 10.

record independently with probability $\beta = 1 - e^{-\epsilon_{anon}}$. This leads to provable privacy guarantees as we will see in the next section. Then, a search through the space of all full-domain generalization schemes is performed. It comprises multiple iterations which are implemented by the for-loop in lines 3 to 10. In each iteration the sample is anonymized using every full-domain generalization scheme in the set G. The quality of the resulting data is assessed and a good solution is selected in a probabilistic manner. Finally, the mechanism returns the best transformation which has been selected during all iterations. In the following sections we will describe each component in greater detail.

4 Anonymization Operator

An overview of the anonymization operator is shown in Figure 5. It builds upon prior work by Li et al. [39] which we have extended with a parameter calculation so that the operator satisfies (ϵ, δ)-differential privacy for arbitrary user-specified parameters. The operator first generalizes the (sampled) input dataset using the generalization scheme g, and then suppresses every record which appears less than k times. Thereby the integer k is derived from the privacy parameters δ and ϵ_{anon}. We will simply denote ϵ_{anon} with ϵ in this section.

Every output of the operator obviously satisfies k-anonymity. Moreover, Li et al. have shown that:

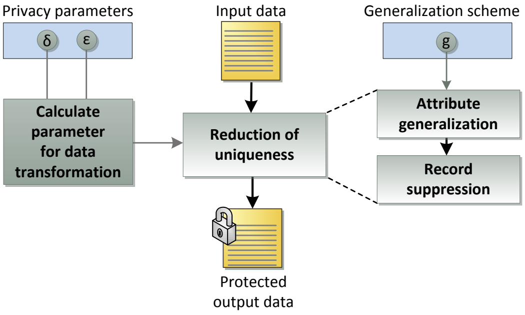


Fig. 5. Overview of the anonymization operator.

Theorem 3. *Random sampling with probability β followed by attribute generalization and the suppression of every record which appears less than k times satisfies (ϵ, δ)-differential privacy for every $\epsilon \geq -\ln(1-\beta)$ with*

$$\delta = d(k, \beta, \epsilon) := \max_{n:n \geq n_m} \sum_{j>\gamma n}^{n} f(j; n, \beta) \tag{2}$$

where $n_m := \left\lceil \frac{k}{\gamma} - 1 \right\rceil$, $\gamma := \frac{e^{\epsilon} - 1 + \beta}{e^{\epsilon}}$ and $f(j; n, \beta) := \binom{n}{j} \beta^j (1-\beta)^{n-j}$, which is the probability mass function of the binomial distribution [39].

It can be seen that the calculation of β described in Section 3 follows from Theorem 3:

$$\epsilon \geq -\ln(1-\beta) \Rightarrow \beta \leq 1 - e^{-\epsilon} := \beta_{max}$$

We will explain why we set $\beta = \beta_{max}$ in Section 8.2.

To derive a practical anonymization operator from Theorem 3, it is necessary to calculate a value for k from given values of ϵ, δ and β. This is not trivial since Equation (2) requires to find the maximum of an infinite non-monotonic sequence. In the following we will show how this is implemented in SafePub. To do so, we will first introduce some definitions for notational convenience and recapitulate some important prior results.

For ease of notation we define the sequence:

$$a_n := \sum_{j>\gamma n}^{n} f(j; n, \beta). \tag{3}$$

It follows that $d(k, \beta, \epsilon) = \max_{n:n \geq n_m} a_n$. Furthermore, we will use the following sequence:

$$c_n := e^{-n(\gamma \ln(\frac{\gamma}{\beta}) - (\gamma - \beta))}. \tag{4}$$

Li et al. have shown in [40] that c_n is strictly monotonically decreasing with $\lim_{n \to \infty} c_n = 0$ and that it is an upper bound for a_n, i.e. it satisfies:

$$\forall n \in \mathbb{N} : a_n \leq c_n. \tag{5}$$

From these results we can conclude:

$$\delta = d(k, \beta, \epsilon) = \max_{n:n \geq n_m} a_n \underset{(5)}{\leq} \max_{n:n \geq n_m} c_n \leq c_{n_m}. \tag{6}$$

The sequence a_n consists of sums which are, except for multiplicative factors, partial sums of a row in Pascal's triangle. For such sums no closed-form expressions are known [23]. However, we will show that the function d can still be evaluated by using the following simplified representation:

Theorem 4. *The function d has the representation*

$$d(k, \beta, \epsilon) = \max\{a_{n_m}, ..., a_{\tilde{n}}\}$$

where $\tilde{n} \coloneqq \min\{N \geq n_m : c_N \leq a_{n_m}\}$.

Proof. From $\lim_{n\to\infty} c_n = 0$ and $a_{n_m} > 0$ we can conclude:

$$\begin{aligned} &\forall \xi > 0\ \exists N \geq n_m\ \forall n \geq N:\ c_n \leq \xi \\ \Rightarrow\ &\exists N \geq n_m\ \forall n \geq N:\ c_n \leq a_{n_m} \\ \Rightarrow\ &\exists N \geq n_m:\ c_N \leq a_{n_m}. \end{aligned}$$

Hence $\tilde{n}$ exists. Since the sequence c_n is monotonically decreasing with increasing n it follows that:

$$\forall n > \tilde{n}:\ a_n \underset{(5)}{\leq} c_n \leq c_{\tilde{n}} \leq a_{n_m} \leq \max\{a_{n_m}, ..., a_{\tilde{n}}\}.$$

We can conclude:

$$\max\{a_{n_m}, ..., a_{\tilde{n}}\} = \max_{n: n \geq n_m} a_n = d(k, \beta, \epsilon). \qquad \square$$

Theorem 4 allows to derive δ from k, β and ϵ by calculating both a_n and c_n for increasing values of $n \geq n_m$ until an index $\tilde{n}$ satisfying $c_{\tilde{n}} \leq a_{n_m}$ is reached. δ is then the maximum of the finite sequence $a_{n_m}, ..., a_{\tilde{n}}$. This strategy is schematically illustrated in Figure 6.

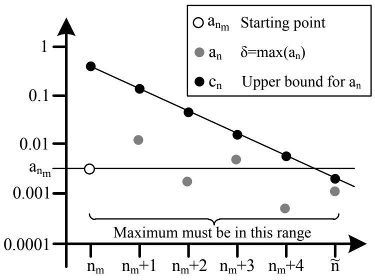


Fig. 6. Schematic plot of a_n and c_n in the range n_m to $\tilde{n}$.

For fixed values of β and ϵ we obtain the function $d(\cdot, \beta, \epsilon) : \mathbb{N} \to [0, 1]$. In order to use this function to compute a value of k so that (ϵ, δ)-differential privacy is provably satisfied, we will first prove that $d(\cdot, \beta, \epsilon)$ converges:

Theorem 5. *For arbitrary $\epsilon > 0$ and $0 < \beta < 1$, $\lim_{k\to\infty} d(k, \beta, \epsilon) = 0$ is satisfied.*

Proof. Note that n_m is a function of k which satisfies:

$$n_m = n_m(k) = \left\lceil \frac{k}{\gamma} - 1 \right\rceil \to \infty,\ k \to \infty.$$

Using the strict monotonicity of c_n we can conclude:

$$\begin{aligned} 0 &\leq d(k, \beta, \epsilon) = \max\left\{a_{n_m(k)}, ..., a_{\tilde{n}}\right\} \\ &\underset{(5)}{\leq} \max\left\{c_{n_m(k)}, ..., c_{\tilde{n}}\right\} = c_{n_m(k)} \to 0,\ k \to \infty. \end{aligned}$$

The claim follows according to the squeeze theorem. $\square$

From this result we can conclude:

$$\forall \delta > 0\ \exists k \in \mathbb{N} : d(k, \beta, \epsilon) \leq \delta.$$

In order to find the smallest such k for a given value of δ, we can evaluate $d(k, \beta, \epsilon)$ as described above for increasing values of $k \in \mathbb{N}$ until $d(k, \beta, \epsilon) \leq \delta$ is satisfied. More formally, k can be computed using the function:

$$d'(\delta, \beta, \epsilon) \coloneqq \min\{k \in \mathbb{N} : d(k, \beta, \epsilon) \leq \delta\}.$$

We denote the output of the operator with $S(D) \coloneqq$ *suppress*$(g(D), k)$, where $g(D) \coloneqq \bigcup_{r\in D}\{g(r)\}$ and *suppress* denotes a function that suppresses every record which appears less than k times.

5 Quality Assessment

The output of the anonymization operator must be assessed to determine a good solution. For this purpose the search strategy employs the exponential mechanism. In this section we will present implementations of common quality models as score functions and discuss their sensitivities. They comprise five general-purpose models which are frequently used in the literature [28, 56] and which have been recommended in current data de-identification guidelines [11] as well as a special-purpose model for building statistical classifiers. For proofs we refer to Appendix B.

5.1 Granularity and Intensity

Data Granularity is a cell-level, general-purpose model. It measures the extent to which the values in a dataset cover the domains of the respective attributes [25]. Since the model already has a low sensitivity, we can multiply its results with -1 to obtain a score function which measures data quality rather than information loss:

Definition 4. For $i = 1, ..., m$, let $leaves_i : \Omega_i \cup \Lambda_i \to \mathbb{N}$ denote the function which returns the number of leaf nodes for each value r'_i within the generalization hierarchy of the i-th attribute. For every $k \in \mathbb{N}$, we define the score function $gran_k : (\mathcal{D}_m \times \mathcal{G}_m) \to \mathbb{R}$ as follows:

$$gran_k(D, g) := - \sum_{(r'_1,\ldots,r'_m)\in S(D)} \sum_{i=1}^{m} \frac{leaves_i(r'_i)}{|\Omega_i|}.$$

The sensitivity of $gran_k$ is as follows (see Appendix B.1):

Theorem 6. *For every $k \in \mathbb{N}$, the following holds:*

$$\Delta gran_k \leq \begin{cases} (k-1)m, & \text{if } k > 1 \\ m, & \text{if } k = 1 \end{cases} .$$

Generalization Intensity is another cell-level, general-purpose model which sums up the relative generalization level of values in all cells [52]. A score function $intensity_k : (\mathcal{D}_m \times \mathcal{G}_m) \to \mathbb{R}$ which is tailored to this model, and which has the same sensitivity as $gran_k$, can be constructed analogously.

5.2 Discernibility

Discernibility is a record-level, general-purpose model which penalizes records depending on the size of the equivalence class they belong to [3]. Let $EQ(D)$ denote the set of all equivalence classes of D, except of $\{* \in D\}$, which contains the suppressed records in D. We first define the following normalized variant of the model:

$$\phi(D) := \left(\sum_{E\in EQ(D)} \frac{|E|^2}{|D|} \right) + |\{* \in D\}|. \qquad (7)$$

We note that suppressed records are considered separately from the other records in Equation (7) as this improves the sensitivity of the function. The score function $disc_k : (\mathcal{D}_m \times \mathcal{G}_m) \to \mathbb{R}$ is defined as follows:

Definition 5. $disc_k(D, g) := -\phi(S(D))$.

The sensitivity of $disc_k$ is as follows (see Appendix B.2):

Theorem 7. *For every $k \in \mathbb{N}$, the following holds:*

$$\forall k \in \mathbb{N} : \Delta disc_k \leq \begin{cases} 5, & \text{if } k = 1 \\ \frac{k^2}{k-1} + 1, & \text{if } k > 1 \end{cases} .$$

5.3 Non-Uniform Entropy

Non-Uniform Entropy is an attribute-level, general-purpose model which quantifies the amount of information that can be obtained about the input dataset by observing the output dataset [7]. According to this model information loss increases with increasing homogeneity of attribute values in the output dataset. Hence we will base the score function on a measure of homogeneity.

Let $p_i(D)$ denote the projection of D to its i-th attribute. We can then measure the homogeneity of attribute values in D using the function ϕ (see Equation (7)) by calculating $\sum_{i=1}^{m} \phi(p_i(D))$ and thus define:

Definition 6. For every $k \in \mathbb{N}$, the score function $ent_k : (\mathcal{D}_m \times \mathcal{G}_m) \to \mathbb{R}$ is defined as:

$$ent_k(D, g) := -\sum_{i=1}^{m} \phi(p_i(S(D))).$$

The sensitivity of ent_k is as follows (see Appendix B.3):

Theorem 8. *For every $k \in \mathbb{N}$, we have*

$$\Delta ent_k \leq \begin{cases} 5m, & \text{if } k = 1 \\ (\frac{k^2}{k-1} + 1)m, & \text{if } k > 1 \end{cases} .$$

5.4 Group Size

Group Size is a record-level, general-purpose model which measures the average size of equivalence classes [36]. We derive a score function which is inversely correlated to this model as follows:

Definition 7. For every $k \in \mathbb{N}$, the score function $groups_k : (\mathcal{D}_m \times \mathcal{G}_m) \to \mathbb{R}$ is defined as:

$$groups_k(D, g) := |EQS(D)|.$$

Since the addition of a single record can lead to at most one additional equivalence class, it is easy to see that $\forall k \in \mathbb{N} : \Delta groups_k \leq 1$ holds.

5.5 Statistical Classification

Iyengar has proposed a special-purpose model which measures the suitability of data as a training set for statistical classifiers [25]. It penalizes records which do not contain the most frequent combination of feature and class attribute values. Since the model already has a low sensitivity, we can derive a practical score function by giving weights to records which are not penalized:

Definition 8. For every $k \in \mathbb{N}$, the score function $class_k : (\mathcal{D}_m \times \mathcal{G}_m) \to \mathbb{R}$ is defined as follows:

$$class_k(D, g) := \sum_{r'\in S(D)} w(S(D), r') .$$

Let $fv(r')$ denote the the sub-vector of a record r' which consists of the feature attribute values in r'. The record

r' is given a weight by the function w if $fv(r')$ is not suppressed and if the class attribute value $cv(r')$ of r' is equal to the most frequent class value $cv_{maj}(S(D), r')$ among all records in $S(D)$ which share the same combination of feature values. More precisely, we define:

$$w(S(D), r') := \begin{cases} 1, & \text{if } fv(r') \text{ is not suppressed and} \\ & cv(r') = cv_{maj}(S(D), r') \text{ holds} \\ 0, & \text{otherwise} \end{cases} .$$

The sensitivity of $class_k$ is as follows (see Appendix B.4):

Theorem 9. *For every $k \in \mathbb{N}$, $\Delta class_k \leq k$ holds.*

6 Search Strategy

The search strategy implements a (randomized) top-down search through the generalization lattice using the scores which are calculated according to the given quality model. Traversal is implemented by iterative applications of the exponential mechanism which exponentially favors transformations with high scores, and thus likely returns transformations resulting in good output data quality (see Section 2.6). For ease of notation we will denote ϵ_{search} with ϵ in this section.

```
Input: Dataset $D \in \mathcal{D}_m$, Real $\epsilon$, Integer $steps$,
 1: ScoreFunction $s : (\mathcal{D}_m \times \mathcal{G}_m) \rightarrow \mathbb{R}$
Output: Scheme $g \in \mathcal{G}_m$
 2: Real $\tilde{\epsilon} \leftarrow \epsilon / steps$
 3: Scheme $pivot \leftarrow \top$
 4: Scheme $optimum \leftarrow \top$
 5: SchemeSet $candidates \leftarrow \{\top\}$
 6: for (Int $i \leftarrow 1, ..., steps$) do
 7:     $candidates \leftarrow candidates \cup predecessors(pivot)$
 8:     $candidates \leftarrow candidates \setminus \{pivot\}$
 9:     $pivot \leftarrow \mathcal{E}_s^{\tilde{\epsilon}}(D, candidates)$
10:     if ($s(D, pivot) > s(D, optimum)$) then
11:         $optimum \leftarrow pivot$
12:     end if
13: end for
14: return $optimum$
```

Fig. 7. Detailed presentation of the search strategy.

Figure 7 shows a more detailed presentation of the search strategy which is also outlined in the high-level overview in Figure 4 (the loop in lines 6 to 13 of Figure 7 corresponds to the loop in lines 3 to 10 of Figure 4). The function *predecessors* maps a transformation to the set of its direct predecessors. The search starts with the transformation $\top \in \mathcal{G}_m$ which generalizes every attribute to the highest level available. The scores of all direct predecessors of $\top$ are calculated and the transformations are put into the set *candidates*. In each iteration a pivot element is selected from the set using the exponential mechanism with a privacy budget of $\tilde{\epsilon} = \epsilon / steps$, the scores of all its direct predecessors are calculated, and the predecessors are being added to *candidates*. The pivot element is then removed from the set. After a predefined number of steps the method returns the pivot element with the best score.

We note that using $steps = 0$ is possible but impractical, as this results in the deterministic selection of the transformation $\top$ that suppresses all data.

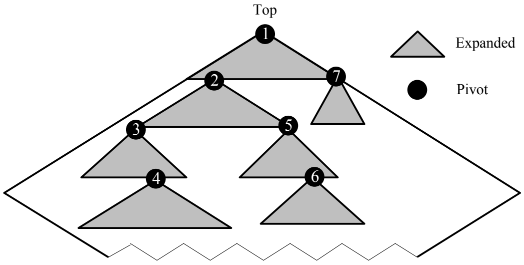


Fig. 8. Schematic illustration of the search strategy.

Figure 8 schematically illustrates the method. A black circle represents a pivot element and the gray triangle below it represents its direct predecessors. The method is likely to perform a best-first search, following a path of transformations with increasing score values (e.g. the path from transformation no. 1 to no. 4). We note that it is not likely that the algorithm will be trapped in a local minimum, i.e. that it continues following a path of elements with non-optimal score values. The reason is that the predecessors of all previously selected pivot elements are left in the set *candidates*. For example, if all predecessors of pivot element no. 4 have a lower score than transformation no. 5, then transformation no. 5 will likely be selected as the next pivot element. Moreover, following a non-optimal path is unlikely to negatively affect the quality of the overall output, as the final solution is selected deterministically. The privacy guarantees provided are as follows:

Theorem 10. *For every parameter $steps \in \mathbb{N}_0$ and $\epsilon > 0$, the search strategy satisfies ϵ-differential privacy.*

Proof. If $steps = 0$ holds the search strategy returns $\top$ in a deterministic manner and hence trivially satisfies ϵ-differential privacy. In the following we will assume that $steps > 0$ holds. We note that the only instructions which modify the content of the variables *pivot* and *candidates* are located in lines seven to nine.

For every iteration $i \in \{1, ..., steps\}$ of the enclosing loop, let $\mathcal{M}_i^{r_i}(D)$ denote the sequence of operations performed by these three lines during the i-th iteration. Let $r_i = (pivot_i, candidates_i)$ denote the content of the variables *pivot* and *candidates* before the i-th iteration of the loop. Then each r_i is determined by $\mathcal{M}_1^{r_1}(D), ..., \mathcal{M}_{i-1}^{r_{i-1}}(D)$ and supplied to $\mathcal{M}_i^{r_i}(D)$ which outputs r_{i+1} in a manner that satisfies $\tilde{\epsilon}$-differential privacy according to Theorem 2. We can conclude from Theorem 1 that the sequence $\mathcal{M}_1^{r_1}(D)$, ..., $\mathcal{M}_{steps}^{r_{steps}}(D)$ satisfies ϵ-differential privacy since $\sum_{i=1}^{steps} \tilde{\epsilon} = \epsilon$ holds. Finally, the algorithm returns the generalization scheme with the highest score value amongst all pivot elements selected by the differentially private operations $\mathcal{M}_1^{r_1}(D)$, ..., $\mathcal{M}_{steps}^{r_{steps}}(D)$ in a deterministic manner. Hence the algorithm satisfies ϵ-differential privacy. □

7 Analytical Evaluation

7.1 Complexity Analysis

Let $n = |D|$ denote the number of records, each consisting of m attributes. Each basic operation, i.e. drawing a random sample, executing the anonymization operator and evaluating a score function, has a runtime complexity of $O(n \cdot m)$. In each step of the search process, the anonymization operator and the score function are being evaluated once for at most m predecessors of the current pivot element. Hence each step has a time complexity of $O(n \cdot m^2)$. The number of steps performed is a user-defined parameter and we will derive recommendations experimentally in Section 8.

We note that the method for calculating the parameters of the algorithm described in Section 4 is of non-trivial runtime complexity. Unfortunately a detailed analysis is complex and out of the scope of this work. We have, however, performed experimental evaluations using a wide variety of common parameterizations which showed that the approach is practical. We will present the results in the next section.

7.2 Parameter Analysis

In this section we analyze dependencies between parameters of SafePub. We will focus on ϵ_{anon} and δ since they determine k and β in a non-trivial manner. For ease of of notation, we will denote ϵ_{anon} with ϵ.

Figure 9 shows the values of β and k obtained for various values of ϵ and δ as described in Section 4. We focus on common values of ϵ [6]. Later we will set δ to 10^{-m} with $m \in \mathbb{N}$ such that $\delta < 1/n$, where n is the size of the dataset, and at least $\delta \leq 10^{-4}$ holds. This is a recommended parameterization [38, 40]. We focus on ranges of δ relevant to our evaluation datasets (see Section 8.1).

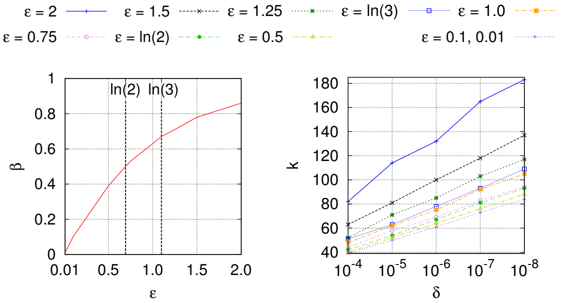


Fig. 9. Overview of values for k and β derived from ϵ and δ.

As can be seen, for fixed values of ϵ, decreasing δ increases k and thus potentially reduces data quality. Decreasing ϵ, however, has two consequences with possibly opposing impacts: On one hand, β decreases, but on the other hand, for fixed values of δ, k decreases as well. The value of β decreases rapidly for smaller values of ϵ which indicates that our approach is not practical with such parameterizations. When ϵ increases, the increase of β flattens, while k increases further.

We also measured the time required to calculate β_{max} and k for every ϵ discussed here and $10^{-4} \leq \delta \leq 10^{-20}$ on a desktop PC with a quad-core 3.1 GHz Intel Core i5 CPU. We measured between 0.1s and 37s with an average of 4.5s. This shows that the method presented in Section 4 terminates quickly for realistic privacy parameters.

7.3 Smooth Privacy

While the (ϵ, δ)-differential privacy model guarantees that the bound $\exp(\epsilon)$ in Inequation (1) may be exceeded with a probability of at most δ, it does not restrict the permitted degree of exceedance.

Li et al. have suggested that the mechanism studied here has the property that the higher such an exceedance is, the more unlikely it is to occur [40]. However, their results just provide upper bounds for these probabilities based on Inequation (6) which are very conservative: For example, for the values $\epsilon = 1$, $\beta = 0.632$ and $k = 75$, they overestimate δ by more than four orders of magnitude ($3.7 \cdot 10^{-2}$ vs. 10^{-6}). Based on our results, we can calculate the exact probabilities:

Theorem 11 (Smoothness property)**.** *For arbitrary parameters $\epsilon = \epsilon_{anon} > 0$ and $\delta > 0$, let β and k be the parameters derived as described in Section 4. Then the* *combination of random sampling with probability β and the anonymization operator satisfies $(\epsilon', d(k, \beta, \epsilon'))$-differential privacy simultaneously for every $\epsilon' \geq \epsilon$ while $d(k, \beta, \epsilon')$ is monotonically decreasing when ϵ' increases.*

The proof can be found in Appendix A.

Figure 10 illustrates the smoothness property for $\epsilon = 1$ and various values of δ. As can be seen, the probability of exceeding ϵ decreases exponentially for increasing degrees of exceedance. The smaller δ, the steeper are the curves, which means that the smoothness effect is stronger. Hence, when δ is set based on the size of the dataset as described in Section 7.2, the degree of protection increases with increasing size of the dataset.

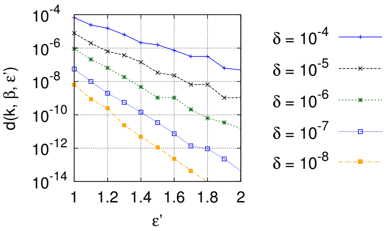


Fig. 10. Semi-log plot showing the smoothness property for $\epsilon = 1$ and various values of δ.

8 Experimental Evaluation

We have implemented our method using the open source ARX Data Anonymization Tool[1]. In this section we present experimental analyses of each individual component of SafePub and develop recommendations for parameterizations. Furthermore, we present results of comparisons with related methods.

8.1 Datasets and Setup

We used four different datasets (see [48]) in our experiments: *1) US Census (USC)*, an excerpt of records from the 1994 U.S. Census database which is often used for evaluating anonymization algorithms, *2) Crash statistics (CS)*, a database about fatal traffic accidents, *3) Time use survey (TUS)*, a dataset consisting of responses to a survey on individual time use in the U.S. and *4) Health interviews (HI)*, a database of records from a survey on the health of the U.S. population.

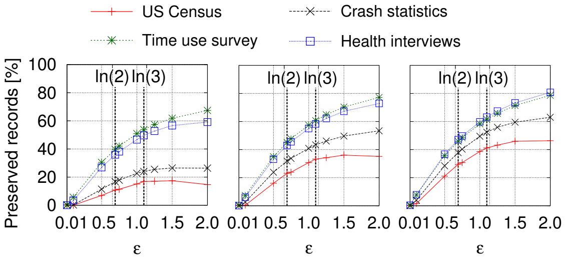
Label	No. of Attributes	No. of Records	Size of Lattice	$\epsilon = 1$ δ	$\epsilon' = 2$ δ'
USC	9	$30,162$	$19,440$	10^{-5}	1×10^{-9}
CS	8	$100,937$	$15,552$	10^{-6}	2×10^{-11}
TUS	9	$539,253$	$34,992$	10^{-6}	2×10^{-11}
HI	8	$1,193,504$	$14,580$	10^{-7}	4×10^{-14}

Table 1. Overview of the evaluation datasets.

The datasets have increasing volumes, ranging from about $30,000$ to more than a million records. All include sensitive data such as demographics (e.g. sex, age) or health-related data. Table 1 provides an overview of the datasets and parameterizations we used in our experiments. It also shows results of the smoothness property, i.e. the probability δ' of violating 2-differential privacy.

8.2 Analysis of the Anonymization Operator

First we examine the amount of records preserved (i.e. not removed by random sampling or record suppression) by the anonymization operator, which is a generic utility estimate. We set $\epsilon_{search} = 0$ and used three full-domain generalization schemes defining *low*, *medium* or *high* relative generalization levels for the attributes in the datasets. We note that the parameter ϵ determines the degree of privacy provided together with δ while the relative generalization level balances the loss of information resulting from generalization against the loss of information resulting from record suppression – the higher the degree of generalization is, the more records are likely to become indistinguishable, and hence the fewer records have to be removed for violating k-anonymity. We focus on the parameters also investigated in Section 7.2. Figure 11 shows averages of 10 executions. All standard deviations were less than 1%.

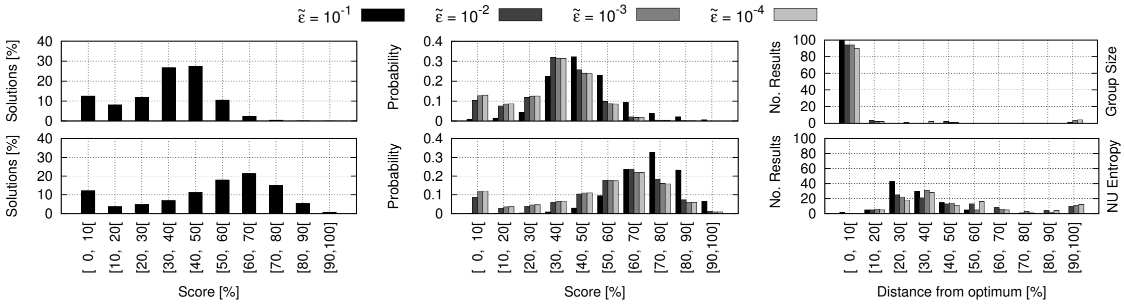


(a) Low **(b)** Medium **(c)** High

Fig. 11. Average number of records preserved by SafePub for each generalization scheme using various values of ϵ.

As can be seen, lower values of ϵ, and the resulting reduction of β and k (see Section 7.2), tendentially led to fewer records being preserved. Only for small datasets, low degrees of generalization and high values of ϵ the decrease of k did outweigh the decrease of β so that more records were preserved (see the "US Census" and "Crash statistics" datasets for $\epsilon = 2$ and $\epsilon = 1.5$). In all other cases the lower sampling probability dominated, especially for realistic values of $\epsilon \leq 1.5$.

[1] http://arx.deidentifier.org/



(a) Distribution of quality **(b)** Probability mass function **(c)** Distribution of output quality

Fig. 12. (a) Distribution of data quality for different score functions. (b) Probability mass functions obtained with varying values of $\tilde{\epsilon}$. (c) Distributions of the quality of results of 100 executions of the exponential mechanism using various values of $\tilde{\epsilon}$.

We note that SafePub uses the highest possible value of β (see Section 3) for a given privacy budget (see Theorem 3). The results presented here justify this choice. They also indicate that values of ϵ_{anon} in the order of one are a good choice. Unless noted otherwise, we will use an overall budget of $\epsilon = 1$ in the following sections, which is a common setup [10, 46] and, as we will show, a good parameterization for our method as well.

8.3 Analysis of the Optimization Functions

We now investigate the effectiveness of the score functions and the quality of transformations selected by the exponential mechanism. We focus on Non-Uniform Entropy and Group Size, because the results obtained for the other score functions lied in between. We further focus on "US Census" and point out differences obtained using the larger datasets where applicable.

Figure 12a shows the distribution of (normalized) scores within the solution space. We note that the y-axis represents the probability of selecting a transformation with a score value in a given range when drawing from the uniform distribution. For the other datasets, the fraction of transformations with higher scores increased with growing volume, because the more records are contained, the less records are likely to be suppressed because they appear less than k times.

Figure 12b shows the probability mass functions used by the exponential mechanism when drawing a solution from the whole solution space using $\epsilon_{anon} = 1$ and various values of $\tilde{\epsilon}$ between 10^{-1} and 10^{-4}. We focus on relatively small values since the search strategy executes the exponential mechanism several times so that higher budgets for each execution would add up to an unusably high overall budget. For $\tilde{\epsilon} = 0.1$ the resulting probability distributions were significantly better than the distributions obtained when drawing from the uniform distribution (see Figure 12a). The improvements decreased with decreasing $\tilde{\epsilon}$ and increased significantly with increasing data volumes. The main reason is that larger datasets often lead to broader ranges of score values in the solution space so that the application of the exponential function according to Definition 3 yields higher differences between probabilities for good and bad solutions.

Figure 12c shows the results of 100 executions of the exponential mechanism. For each transformation selected, we calculated the difference to the optimal solution in terms of data quality using the model for which the score function has been designed. On average, we measured very good results of less than 4% for the Group Size model, even though solutions with a score in the range $[30\%, 50\%[$ were selected with the highest probability. This is because the according score function is not directly proportional to the quality model, but rather inversely proportional. Hence data quality increases significantly with increasing scores. The results for Non-Uniform Entropy were not as good with averages ranging from 31% ($\tilde{\epsilon} = 10^{-1}$) to 49% ($\tilde{\epsilon} = 10^{-4}$). The reason is that the according score function does not resemble the corresponding quality model as closely as the other score functions do. The results imply that a budget which is very small compared to the one required by the anonymization operator can suffice to achieve good results using the exponential mechanism.

8.4 Analysis of the Search Strategy

Next we analyze the influence of the number of steps performed by the search strategy on the quality of output data. We executed SafePub 10 times for each dataset and score function using varying numbers of steps. Since the previous results imply that a budget in the order of one is a good choice for the anonymization operator while a significantly smaller budget is sufficient for the exponential mechanism, we used an overall budget of $\epsilon = 1$ which we have split into various combinations of ϵ_{search} and ϵ_{anon}. Figure 14 shows the results

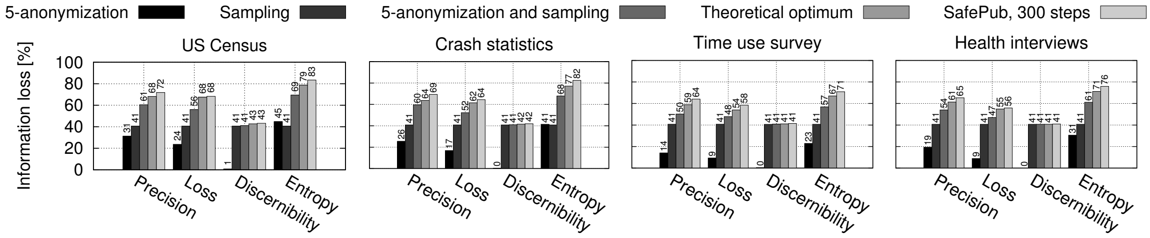


Fig. 13. Average information loss induced by SafePub for $\epsilon = 1$ compared with the (average) results of various baseline methods.

obtained for the "Health interviews" dataset, which are representative for the other datasets. The results for the Discernibility model were comparable to the results for the Granularity model. We normalized all values so that 0% corresponds to the input dataset and 100% to a dataset from which all information has been removed. All standard deviations were less than 12%.

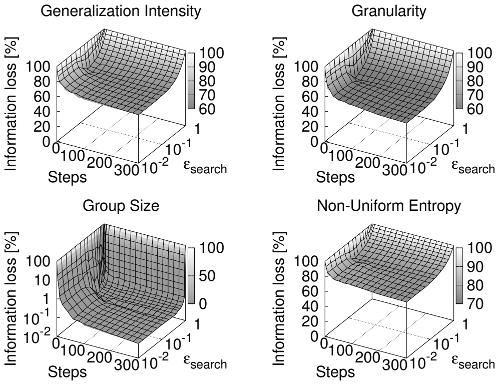


Fig. 14. Average information loss induced by SafePub for various step values and various values of ϵ_{search} with $\epsilon_{anon} = 1 - \epsilon_{search}$. The ϵ_{search} axis and the information loss values for the Group Size model are scaled logarithmically.

We note that increasing the number of steps performed by SafePub has two consequences: The number of executions of the exponential mechanism increases, while the budget $\tilde{\epsilon}$, which is used for each execution, decreases. It can be observed that the former effect tends to outweight the latter so that increasing the number of steps improves data quality. In all experiments the effect flattened at around 300 steps. Decreasing ϵ_{search} from 1 to 10^{-1} generally improved the results. Further reductions decreased data quality in some experiments and had no significant effects in the others. Hence, in the following sections, we will use a default parameterization of 300 steps, $\epsilon_{search} = 0.1$ and $\epsilon_{search} = 0.9$ unless noted otherwise. These values result in a budget of $\tilde{\epsilon} \approx 10^{-4}$ which did not perform as well as higher values when drawing from the whole solution space $\mathcal{G}_m$ (see Section 8.3). However, since the search strategy draws repeatedly out of subsets of $\mathcal{G}_m$, it can still select very good solutions as we will see in the next section.

8.5 Analysis of the Quality of Output

Here we analyze output data quality for the default parameterization and compare it with the quality obtained using various baseline methods: The optimal quality obtained with k-anonymization, by only using random sampling and by random sampling combined with k-anonymization. We also measured the quality of the theoretical optimum which can be obtained with SafePub by deterministically selecting the optimal generalization scheme rather than using the search strategy. Each of these methods constitutes a baseline in terms of output quality for (combinations of) transformations performed by SafePub, and hence illustrates their impact on output data quality. We note that none of them satisfies differential privacy but that all approaches have been implemented such that the optimal transformation according to a given quality model is selected. To establish a strict baseline we set $k = 5$, which is common in the literature [12, 13] but less conservative than other values, e.g. $k = 11$ which has been recommended by the European Medicines Agency (EMA) [14].

The results are shown in Figure 13. Numbers for the Group Size model are not included as we measured values of less than 2% for all approaches. It can be seen that SafePub removed a significant amount of information from the datasets, i.e. between 83% and 71% according to the Non-Uniform Entropy model and between 41% and 43% according to the Discernibility model. It can further be observed that random sampling contributed the most to these reductions (41%). The average difference between results of SafePub and the theoretical optimum was very small (less than 3%). We note that, even though SafePub produced near-optimal results, the fraction of the solution space which has been traversed by the search strategy was relatively small. When using the "Crash statistics" dataset, this fraction was about 5%. In the other cases, it was about 10%. This confirms that the search strategy performs very well using the default parameters. In particular, it also achieves very good results for the Non-Uniform Entropy model, for which the exponential mechanism alone did not perform as well as for the other models (see Section 8.3).

8.6 Analysis of the Utility of Output

As there is not necessarily a strong correlation between loss of information and the actual usefulness of data, we now evaluate the performance of statistical classifiers built with the output of SafePub. This is the most common benchmarking workload for methods of privacy-preserving data publishing. We have used the class attributes listed in Table 2, which resulted in both binomial and multinomial classification problems.

Dataset	Class attributes	Number of instances
USC	(1) Marital status	8
	(2) Salary class	2
CS	(1) Hispanic origin	10
	(2) Race	20
TUS	(1) Marital status	7
	(2) Sex	3
HI	(1) Marital status	10
	(2) Education	26

Table 2. Overview of the class attributes used in our evaluations.

For each dataset and class attribute, we executed 100 runs of SafePub with varying numbers of steps, varying values of ϵ_{anon} and $\epsilon_{search} = 0.1$. We focused on ϵ_{anon}, since the previous results showed that small values of ϵ_{search} are sufficient and that ϵ_{anon} thus primarily determines the overall trade-off between privacy and utility provided by SafePub. We configured SafePub to use the score function which optimizes output data for training statistical classifiers (see Section 5.5). All attributes besides the class attribute were used as features, and we used generalization schemes which do not generalize the class attribute.

As a classification method we used decision trees generated with the well-known C4.5 algorithm [49] because this is the most frequently used method in our context. We point out that it is obviously possible to use other classification methods with our approach and that we have obtained comparable results using logistic regression classifiers [54]. We created the classifiers from output data and evaluated their prediction accuracy with input data using the approach presented in [16] and 10-fold cross-validation. We report *relative prediction accuracies*, which means that all values have been normalized so that 0% represents the accuracy of the trivial ZeroR method, which always returns the most frequent value of the class attribute, while 100% corresponds to the accuracy of C4.5 decision trees trained on input data.

Figure 15 shows the results of varying ϵ_{anon} using 300 steps. As can be seen, the impact of ϵ_{anon} was relatively small considering the strong effect on the number of preserved records (see Section 8.2). As expected, small values of ϵ_{anon} often resulted in sub-optimal accuracies. Values of about $\epsilon_{anon} = 0.9$ generally resulted in good performance. Further increasing the parameter decreased the accuracies obtained. The reason is that, although increasing ϵ_{anon} increases the number of preserved records, k also increases, which eventually causes a high degree of generalization.

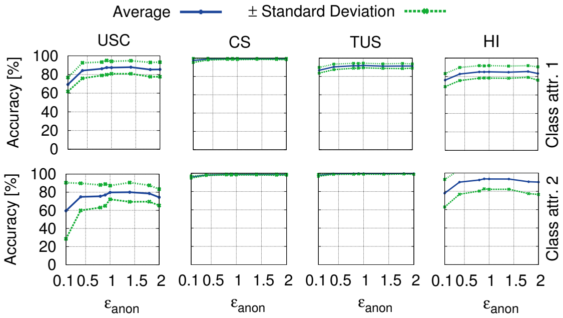


Fig. 15. Relative classification accuracies obtained using various values of ϵ_{anon}, $\epsilon_{search} = 0.1$ and 300 steps.

Figure 16 shows results obtained for varying numbers of steps and $\epsilon_{anon} = 0.9$. As can be seen, the performance of the classifiers improved with an increasing number of steps. The average accuracies obtained using 300 steps ranged from 82% when predicting the second class attribute of "US Census" to about 99% when predicting the class attributes of "Crash statistics". Results were rather stable (standard deviations of about 5%).

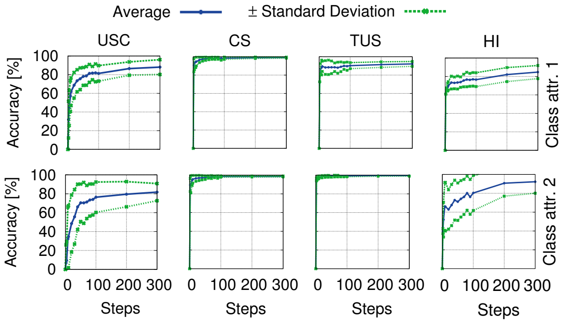


Fig. 16. Relative classification accuracies obtained using various numbers of steps, $\epsilon_{anon} = 0.9$ and $\epsilon_{search} = 0.1$.

Using the same setup, we also evaluated classification accuracies obtained using the output of the baseline methods discussed in Section 8.5 (sampling only, k-anonymization only), also optimized for building classifiers. All accuracies achieved were at least 97%.

In summary, these experiments justify our default parameterization, and we conclude that the differences between the performance of classifiers trained with unmodified input or the output of baseline methods and classifiers trained with the output of SafePub are small. This indicates that although SafePub removes a significant amount of information, it does so in a controlled manner which preserves frequent patterns hidden in the data.

8.7 Comparison With Prior Work

In this section we will put our method into perspective by experimentally comparing it to related approaches. We have performed all experiments using the default configuration (300 steps, $\epsilon = 1$) and we have calculated δ as described in Section 7. Where applicable, we used the same hierarchies as in the previous experiments.

8.7.1 Comparison With Other Approaches for Differentially Private Statistical Classification

We compared SafePub to the following state-of-the-art algorithms: DiffGen [46], DiffP-C4.5 [19], LDA [55], SDQ [57] and DPNB [29]. We have exactly replicated the setups reported in the respective publications and refer to them for exact specifications. All evaluations used (variants of) the "US Census" dataset (see Section 8.1) and the "Nursery" dataset [42]. We point out that the other methods implement ϵ-differential privacy while SafePub satisfies the slight relaxation (ϵ, δ)-differential privacy, which potentially allows for higher data quality. However, unlike the other methods which output classifiers or synthetic microdata, SafePub outputs truthful microdata using a less flexible but truthful transformation technique.

Algorithm	DiffP-C4.5	LDA	DPNB	DPNB	SDQ
Dataset	US Census			Nursery	
Competitor	82.1%	80.8%	82%	90%	79.9%
SafePub	80.9%	81.5%	81.2%	83.7%	83.8%

Table 3. Comparison of absolute prediction accuracies for $\epsilon = 1$.

The results for all mechanisms except DiffGen, which we will address below, are listed in Table 3. As can be seen, the accuracies obtained using C4.5 and SafePub were comparable to the results of DiffP-C4.5, LDA and DPNB for the "US Census" dataset. For the "Nursery" dataset, SafePub outperformed SDQ, while DPNB outperformed SafePub by 6.3%. In all experiments, we measured standard deviations of $< 2\%$.

DiffGen is particularly closely related to SafePub because it also produces microdata using concepts from data anonymization (i.e. attribute transformation based on generalization hierarchies). Hence we have performed a more detailed analytical and experimental comparison. DiffGen employs a more flexible transformation model, subtree generalization, where values of an attribute can be transformed to different generalization levels (see Section 2.3). Analogously to SafePub, it also selects a transformation based on a user-specified number of iterative applications of the exponential mechanism (steps). However, in contrast to our approach, it does not achieve differential privacy by random sampling and k-anonymization, but rather by probabilistically generating synthetic records.

Using the implementation provided by the authors and our evaluation datasets we compared SafePub and DiffGen using C4.5 decision trees which were evaluated using 2/3 of the records as training data and the remaining 1/3 as test data (as proposed by the authors of DiffGen [46]). We used a privacy budget of $\epsilon = 1$ for both methods and increasing numbers of steps. The number of steps DiffGen can perform has a limit which depends on the heights of the generalization hierarchies and which was around 20 in our setup. For SafePub, we used between 0 and 300 steps since higher values did not improve the quality of results (see Section 8.4). We performed every experiment 20 times. Table 4 lists average execution times and standard deviations for the maximal number of steps measured on the hardware described in Section 7.2. Moreover, we included the optimal accuracies obtained using any number of steps.

Label	Class Attribute	Execution times		Max. Accuracies	
		SafePub	DiffGen	SafePub	DiffGen
USC	1	**4.8 ± 1.0s**	**16.2 ± 0.7s**	**92.0%**	**85.0%**
	2	**5.1 ± 1.3s**	**21.9 ± 0.6s**	**87.3%**	**79.2%**
CS	1	**8.8 ± 0.7s**	**18.5 ± 1.6s**	**99.7%**	**97.9%**
	2	**8.9 ± 0.6s**	**6.5 ± 2.5s**	**99.9%**	**98.3%**
TUS	1	**54.2 ± 4.5s**	**28.7 ± 0.7s**	**93.6%**	**91.0%**
	2	**55.3 ± 2.0s**	**30.9 ± 0.6s**	**99.9%**	**99.7%**
HI	1	**98.0 ± 5.8s**	**61.1 ± 2.2s**	**87.7%**	**94%**
	2	**103.5 ± 9.2s**	**65.0 ± 2.1s**	**99.1%**	**64.0%**

Table 4. Comparison of absolute execution times and maximal relative accuracies achieved for $\epsilon = 1$.

SafePub outperformed DiffGen regarding maximal accuracies in seven out of eight experiments. The accuracies obtained by SafePub when predicting the second class attribute of "Health interviews" were 35% higher than the results obtained by DiffGen. The minimal and maximal execution times of SafePub varied from between 4s and 7s ("US Census") to between 90s and 128s ("Health interviews"). The corresponding times of DiffGen varied from between 15s and 18s to between 62s and 70s. In summary, SafePub was faster than DiffGen for smaller datasets while DiffGen was faster than SafePub for larger datasets.

A more detailed analysis is provided in Figure 17, which shows execution times and relative accuracies obtained using different numbers of steps.

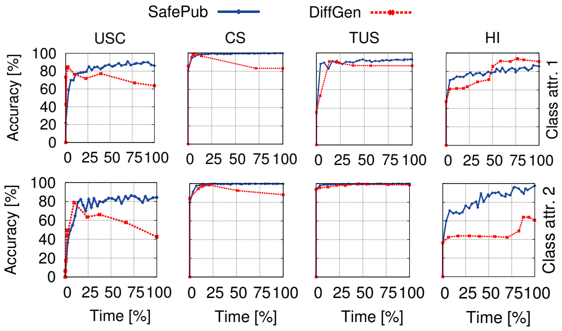


Fig. 17. Relative classification accuracies obtained for $\epsilon = 1$.

It can be seen that the accuracies achieved by SafePub improved monotonically over time (apart from minor fluctuations which are a result of randomization) while no such relationship can be observed for DiffGen. We explain this by the fact that SafePub is not likely to be trapped in a local minimum (see Section 6) while DiffGen can only keep on specializing a transformation once it has been selected. This implies that SafePub is easier to parameterize and enables trading execution times off against data quality.

8.7.2 Comparison With the Approach by Fouad et al.

We conclude our experimental evaluation by presenting a comparison with the approach which is most closely related to ours. Fouad et al. have also proposed a truthful (ϵ, δ)-differentially private microdata release mechanism using random sampling and generalization [18, 43].

Their algorithm replaces each record independently with a generalized record which is *t-frequent*, i.e. a generalization of at least t records from the input dataset. The authors show that the mechanism satisfies (ϵ, δ)-dif-ferential privacy, however, with unknown δ. They further show that an *upper bound* for δ can be calculated when t is chosen greater than a threshold $\lfloor T \rfloor$ [43, Theorem 4]. Knowing δ is, however, crucial for guaranteeing a known degree of privacy.

We analyzed $\lfloor T \rfloor$ and the resulting values of δ for various common input parameters. We emphasize that we chose all parameters in such a way that $\lfloor T \rfloor$ is as small as possible. Figure 18 shows the results for $\epsilon = 1$. As can be seen, δ decreases very quickly for an increasing number of attributes, while $\lfloor T \rfloor$ increases exponentially. For datasets with three attributes, $\lfloor T \rfloor$ equals 76, while for datasets with seven attributes, $\lfloor T \rfloor$ equals $1,217$ already. Hence, a very high degree of generalization is required to obtain known privacy guarantees.

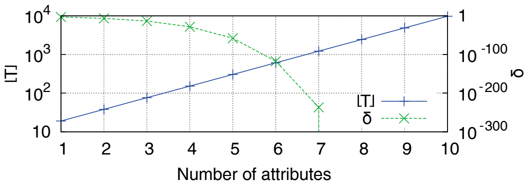


Fig. 18. Analysis of the approach by Fouad et al. The figure shows $\lfloor T \rfloor$ and corresponding values of δ for $\epsilon = 1$ and $h = 2$.

We experimentally evaluated the method choosing $\epsilon = 1$ and $t = \lfloor T \rfloor + 1$ so that the approach satisfies (ϵ, δ)-differential privacy. We performed the experiments ten times and report average results (standard deviations $< 1\%$). All information was removed from all datasets but "Health interviews" for which some information was preserved. However, 68% of records were removed and seven out of eight attributes were completely suppressed. With the models considered in this article, we measured reductions in data quality of between 97% and 99%, which renders the approach impractical.

9 Related Work

Other works have also investigated relationships between syntactic privacy models and differential privacy. Domingo-Ferrer and Soria-Comas have shown that there is a theoretical relationship between ϵ-differential privacy and a stochastic extension of t-closeness and that satisfying t-closeness can imply ϵ-differential privacy under certain assumptions [8]. Moreover, Soria-Comas et al. and Jafer et al. have also combined k-anonymity and differential privacy [27, 30]. While our approach uses k-anonymity in order to create a differentially private mechanism, these works employ k-anonymization to reduce the amount of noise that must be added.

Moreover, further differential privacy mechanisms have been proposed that use random sampling. Fan and Jin [17] as well as Jorgensen et al. [32] have used non-uniform random sampling to produce aggregate data. Hong et al. have used random sampling for protecting search logs [24]. These are all special-purpose mechanisms while SafePub is a generic microdata release algorithm.

For further differentially private microdata release methods see the surveys [6, 38]. Unlike SafePub, most of them are not truthful or use methods that are very different from those typically used in data anonymization. We have compared our approach to the notable exception, i.e. the approach by Fouad et al. [18, 43], in Section 8.7.2 and found that it is not practical.

Differentially private machine learning is also an ongoing field of research (see the surveys [31, 51]). We have compared our approach to five different state-of-the-art methods in Section 8.7.1. We have performed a detailed experimental comparison with DiffGen [46] because of its conceptual similarities to our approach. Our results showed that our method, which is the only generic and truthful approach in the field, achieves accuracies that compare well to those of special-purpose mechanisms.

Gehrke et al. have also studied the approach by Li et al. [26], albeit from a purely theoretical perspective. They showed that it satisfies a privacy model called crowd-blending privacy. Informally, this model guarantees that every record r from the input dataset either blends into a "crowd" of at least k records or that r is essentially being ignored by the mechanism. Their work also indicates that the mechanism satisfies a relaxation of another model called zero-knowledge privacy [22].

10 Summary and Discussion

In this paper we have presented a flexible differentially private data release mechanism that produces truthful output data, which is important in some data publishing scenarios [3] and domains such as medicine [6]. While it has been argued that differential privacy is difficult to explain to non-experts the approach offers an intuitive notion of privacy protection: with a probability determined by ϵ the data of an individual will not be included at all and even if it is included it will only be released in a generalized form such that it cannot be distinguished from the similarly generalized data of at least $k-1$ other individuals, where k is determined by ϵ and δ.

Our evaluation showed that the method is practical and that values in the order of $\epsilon = 1$ are a good parameterization. The current implementation uses full-domain generalization and the k-anonymity privacy model, methods which have frequently been criticized for being too inflexible and too strict to produce output data of high quality [2]. However, our experiments have shown that statistical classifiers trained with the output of the generic method parameterized with an appropriate objective function perform as well as non-truthful differential privacy mechanism designed specifically for this use case. The reason is that while the approach indeed removes a significant amount of information it does so in a controlled manner which extracts frequent patterns. Compared to prior work, however, our approach provides slightly lower privacy guarantees.

While developing the score functions introduced in Section 5, we learned that optimization functions which have the form of sums to which every record or cell contributes a non-negative summand tend to have a low sensitivity. According score functions can often be obtained easily (see score functions for Data Granularity, Intensity and Classification). If the sensitivity is high, it can be possible to reduce it by division through the size of the dataset or by forming reciprocals (see score functions for Discernibility and Group Size). If this is not the case, it can be worthwhile to try to find functions with lower sensitivities which have related properties (see score function for Non-Uniform Entropy).

11 Future Work

An interesting line of future research is to develop score functions tailored to further quality models which address learning tasks such as regression or time-to-event analysis [54]. Based on our experiences presented in the previous section we are confident that, for example, the workload-aware quality models presented by LeFevre et al. in [37] can be integrated into the method.

Another potential direction for further work is to try to consider the effects of random sampling which may have been performed during data acquisition to reduce the amount of explicit random sampling that needs to be used by the mechanism.

In its current form SafePub is suited for protecting dense data of low to medium dimensionality as high-dimensional data is often sparse and hence cannot be k-anonymized while retaining sufficient data quality. We plan to investigate methods for vertically partitioning high-dimensional data, such that disassociated subsets of correlated attributes can be processed independently. Moreover, future work could investigate the crowd-blending and the zero-knowledge privacy models which provide other means of formalizing the notion of "hiding in a group" than our implementation. We point out that these models can also make it possible to prefer certain records, e.g. for publishing control or test data using random sampling which is slightly biased [26].

Finally, a variety of unified frameworks have been proposed for comparing the trade-off between privacy and utility provided by algorithms which implement privacy models, including syntactic ones and ϵ-differential privacy [4, 20, 41]. As the mechanism presented here is the first practical implementation of differential privacy for the release of truthful microdata, it would be interesting to compare it to other methods using such frameworks.

References

[1] A. Machanavajjhala et al. l-diversity: Privacy beyond k-anonymity. *Transactions on Knowledge Discovery from Data*, 1(1):3, 2007.

[2] B. C. M. Fung et al. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. CRC Press, 2010.

[3] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *International Conference on Data Engineering*, pages 217–228, 2005.

[4] J. Brickell and V. Shmatikov. The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 70–78, 2008.

[5] C. Clifton and T. Tassa. On syntactic anonymity and differential privacy. In *International Conference on Data Engineering Workshops*, pages 88–93, 2013.

[6] F. K. Dankar and K. El Emam. Practicing differential privacy in health care: A review. *Transactions on Data Privacy*, 6(1):35–67, 2013.

[7] T. de Waal and L. Willenborg. Information loss through global recoding and local suppression. *Netherlands Official Statistics*, 14:17–20, 1999.

[8] J. Domingo-Ferrer and J. Soria-Comas. From t-closeness to differential privacy and vice versa in data anonymization. *Knowledge-Based Systems*, 74:151–158, 2015.

[9] C. Dwork. An ad omnia approach to defining and achieving private data analysis. In *International Conference on Privacy, Security, and Trust in KDD*, pages 1–13, 2008.

[10] C. Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19, 2008.

[11] K. El Emam and L. Arbuckle. *Anonymizing Health Data*. O'Reilly Media, 2013.

[12] K. El Emam and F. K. Dankar. Protecting privacy using k-anonymity. *Jama-J Am. Med. Assoc.*, 15(5):627–637, 2008.

[13] K. El Emam and B. Malin. Appendix b: Concepts and methods for de-identifying clinical trial data. In *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*, pages 1–290. National Academies Press (US), 2015.

[14] European Medicines Agency. External guidance on the implementation of the european medicines agency policy on the publication of clinical data for medicinal products for human use. EMA/90915/2016, 2016.

[15] F. Prasser et al. Lightning: Utility-driven anonymization of high-dimensional data. *Transactions on Data Privacy*, 9(2):161–185, 2016.

[16] F. Prasser et al. A tool for optimizing de-identified health data for use in statistical classification. In *IEEE International Symposium on Computer-Based Medical Systems*, 2017.

[17] L. Fan and H. Jin. A practical framework for privacy-preserving data analytics. In *International Conference on World Wide Web*, pages 311–321, 2015.

[18] M. R. Fouad, K. Elbassioni, and E. Bertino. A supermodularity-based differential privacy preserving algorithm for data anonymization. *IEEE Transactions on Knowledge and Data Engineering*, 26(7):1591–1601, 2014.

[19] A. Friedman and A. Schuster. Data mining with differential privacy. In *International Conference on Knowledge Discovery and Data Mining*, pages 493–502, 2010.

[20] G. Cormode et al. Empirical privacy and empirical utility of anonymized data. In *IEEE International Conference on Data Engineering Workshops*, pages 77–82, 2013.

[21] G. Poulis et al. Secreta: a system for evaluating and comparing relational and transaction anonymization algorithms. In *International Conference on Extending Database Technology*, pages 620–623, 2014.

[22] J. Gehrke, E. Lui, and R. Pass. Towards privacy for social networks: A zero-knowledge based definition of privacy. In *Theory of Cryptography Conference*, pages 432–449, 2011.

[23] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley publishing company, 2nd edition, 1994.

[24] Y. Hong, J. Vaidya, H. Lu, and M. Wu. Differentially private search log sanitization with optimal output utility. In *International Conference on Extending Database Technology*, pages 50–61, 2012.

[25] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *International Conference on Knowledge Discovery and Data Mining*, pages 279–288, 2002.

[26] J. Gehrke et al. Crowd-blending privacy. In *Advances in Cryptology*, pages 479–496. Springer, 2012.

[27] J. Soria-Comas et al. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *VLDB J.*, 23(5):771–794, 2014.

[28] J. Soria-Comas et al. t-closeness through microaggregation: Strict privacy with enhanced utility preservation. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):3098–3110, 2015.

[29] J. Vaidya et al. Differentially private naive bayes classification. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, pages 571–576, 2013.

[30] Y. Jafer, S. Matwin, and M. Sokolova. Using feature selection to improve the utility of differentially private data publishing. *Procedia Computer Science*, 37:511–516, 2014.

[31] Z. Ji, Z. C. Lipton, and C. Elkan. Differential privacy and machine learning: a survey and review. *CoRR*, abs/1412.7584, 2014.

[32] Z. Jorgensen, T. Yu, and G. Cormode. Conservative or liberal? personalized differential privacy. In *IEEE International Conference on Data Engineering*, pages 1023–1034, April 2015.

[33] K. El Emam et al. A globally optimal k-anonymity method for the de-identification of health data. *J. Am. Med. Inform. Assn.*, 16(5):670–682, 2009.

[34] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper, and K. A. Kuhn. Flash: efficient, stable and optimal k-anonymity. In *2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Conference on Social Computing (SocialCom)*, pages 708–717, 2012.

[35] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *International Conference on Management of Data*, pages 49–60, 2005.

[36] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *International Conference on Data Engineering*, pages 25–25, 2006.

[37] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Workload-aware anonymization techniques for large-scale datasets. *ACM Transactions on Database Systems*, 33(3):1–47, 2008.

[38] D. Leoni. Non-interactive differential privacy: A survey. In *International Workshop on Open Data*, pages 40–52, 2012.

[39] N. Li, W. Qardaji, and D. Su. On sampling, anonymization, and differential privacy: Or, k-anonymization meets differential privacy. In *ACM Symposium on Information, Computer and Communications Security*, pages 32–33, 2012.

[40] N. Li, W. H. Qardaji, and D. Su. Provably private data anonymization: Or, k-anonymity meets differential privacy. *CoRR*, abs/1101.2604, 2011.

[41] T. Li and N. Li. On the tradeoff between privacy and utility in data publishing. In *International Conference on Knowledge Discovery and Data Mining*, pages 517–526, 2009.

[42] M. Lichman. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2013.

[43] M. R. Fouad, K. Elbassioni, and E. Bertino. Towards a differentially private data anonymization. CERIAS Tech Report 2012-1, Purdue Univ., 2012.

[44] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *IEEE Symposium on Foundations of Computer Science*, pages 94–103, 2007.

[45] F. D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *International Conference on Management of Data*, pages 19–30, 2009.

[46] N. Mohammed et al. Differentially private data release for data mining. In *International Conference on Knowledge Discovery and Data Mining*, pages 493–501, 2011.

[47] M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In *International Conference on Management of Data*, pages 665–676, 2007.

[48] F. Prasser, F. Kohlmayer, and K. A. Kuhn. The importance of context: Risk-based de-identification of biomedical data. *Methods of information in medicine*, 55(4):347–355, 2016.

[49] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., 1993.

[50] F. Ritchie and M. Elliott. Principles- versus rules- based output statistical disclosure control in remote access environments. *IASSIST Quarterly*, 39(2):5–13, 2015.

[51] A. D. Sarwate and K. Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE Signal Processing Magazine*, 30(5):86–94, 2013.

[52] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588, Oct. 2002.

[53] L. Willenborg and T. De Waal. *Statistical disclosure control in practice*. Springer Science & Business Media, 1996.

[54] I. H. Witten and F. Eibe. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[55] X. Jiang et al. Differential-private data publishing through component analysis. *Transactions on Data Privacy*, 6(1):19–34, Apr. 2013.

[56] Z. Wan et al. A game theoretic framework for analyzing re-identification risk. *PloS one*, 10(3):e0120592, 2015.

[57] N. Zhang, M. Li, and W. Lou. Distributed data mining with differential privacy. In *IEEE International Conference on Communications*, pages 1–5, 2011.

A Proof of Theorem 11

Proof. For the purpose of this proof we will use the following representation of the function d which is obtained as an intermediate result in the proof of [40, Theorem 1]:

$$d(k, \beta, \epsilon') = \max_{n \in \mathbb{N}} \sum_{\{j \in \mathbb{N} \mid j \geq k \wedge j > \gamma n \wedge j \leq n\}} f(j; n, \beta).$$

Let us fix an arbitrary $\epsilon' \geq \epsilon$ and recall that $\gamma = \gamma(\epsilon')$ is actually a function of ϵ'. It is easy to see that $\epsilon' \geq \epsilon$ implies:

$$\gamma(\epsilon') = \frac{\mathrm{e}^{\epsilon'} - 1 + \beta}{\mathrm{e}^{\epsilon'}} \geq \frac{\mathrm{e}^{\epsilon} - 1 + \beta}{\mathrm{e}^{\epsilon}} = \gamma(\epsilon).$$

Hence we have:

$$\forall n \in \mathbb{N} : \{j \in \mathbb{N} \mid j \geq k \wedge j > \gamma(\epsilon') n \wedge j \leq n\} \subseteq \{j \in \mathbb{N} \mid j \geq k \wedge j > \gamma(\epsilon) n \wedge j \leq n\}.$$

This implies

$$\begin{aligned} d(k, \beta, \epsilon') &= \max_{n \in \mathbb{N}} \sum_{\{j \in \mathbb{N} \mid j \geq k \wedge j > \gamma(\epsilon') n \wedge j \leq n\}} f(j; n, \beta) \\ &\leq \max_{n \in \mathbb{N}} \sum_{\{j \in \mathbb{N} \mid j \geq k \wedge j > \gamma(\epsilon) n \wedge j \leq n\}} f(j; n, \beta) \\ &= d(k, \beta, \epsilon) \end{aligned}$$

which proofs the monotonicity. Furthermore, we have $\epsilon' \geq \epsilon = -\ln(1 - \beta)$ so that $(\epsilon', d(k, \beta, \epsilon'))$-differential privacy is indeed satisfied according to Theorem 3. □

B Proofs of Sensitivities

B.1 Granularity (Theorem 6)

Proof. Let $k \in \mathbb{N}$ be an arbitrary integer, let $g \in \mathcal{G}_m$ be an arbitrary generalization scheme and let $D_1, D_2 \in \mathcal{D}_m$ be arbitrary datasets satisfying $|D_1 \oplus D_2| = 1$. Without loss of generality we assume $D_1 = D_2 \cup \{r\}$. We will use the notation $g(r) = (\tilde{r}_1, ..., \tilde{r}_m)$ and point out that $\forall i = 1, ..., m : \ 0 \leq \frac{leaves_i(\tilde{r}_i)}{|\Omega_i|} \leq \frac{leaves_i(*)}{|\Omega_i|} = 1$ holds.

– If $g(r) = *$ holds or $g(r) \neq *$ appears less than k times in $g(D_1)$, then $g(r)$ is suppressed in both $S(D_1)$ and $S(D_2)$ with $S(D_1) = S(D_2) \cup \{*\}$. We can conclude:

$$\begin{aligned} &|gran_k(D_1, g) - gran_k(D_2, g)| = \\ &\left| \left(\sum_{(r'_1, ..., r'_m) \in S(D_2)} \sum_{i=1}^{m} \frac{leaves_i(r'_i)}{|\Omega_i|} \right) + \left(\underbrace{\sum_{i=1}^{m} \frac{leaves_i(*)}{|\Omega_i|}}_{=1} \right) \right. \\ &\left. - \left(\sum_{(r'_1, ..., r'_m) \in S(D_2)} \sum_{i=1}^{m} \frac{leaves_i(r'_i)}{|\Omega_i|} \right) \right| = m \,. \end{aligned}$$

– If $g(r) \neq *$ appears k times in $g(D_1)$, then it is not suppressed in $S(D_1)$ but in $S(D_2)$ with

$$S(D_1) = (S(D_2) \setminus \underbrace{\{*, ..., *\}}_{k-1-times}) \cup \underbrace{\{g(r), ..., g(r)\}}_{k-times}.$$

We can conclude:

$$\begin{aligned}
&|gran_k(D_1, g) - gran_k(D_2, g)| \\
&= \left| \left(\sum_{j=1}^{k} \sum_{i=1}^{m} \frac{leaves_i(\tilde{r}_i)}{|\Omega_i|} \right) - \left(\sum_{j=1}^{k-1} \sum_{i=1}^{m} \underbrace{\frac{leaves_i(*)}{|\Omega_i|}}_{=1} \right) \right| \\
&= \left| \underbrace{\left(\sum_{j=1}^{k} \sum_{i=1}^{m} \frac{leaves_i(\tilde{r}_i)}{|\Omega_i|} \right)}_{=:\sigma \in [0, km]} - (k-1)m \right| \\
&\leq \begin{cases} (k-1)m, & if \sigma \in [0, (k-1)m) \\ m, & if \sigma \in [(k-1)m, km] \end{cases}.
\end{aligned}$$

– If $g(r) \neq *$ appears more than k times in $g(D_1)$, then $g(r)$ is not suppressed in both $S(D_1)$ and $S(D_2)$ with $S(D_1) = S(D_2) \cup \{g(r)\}$. We can conclude:

$$|gran_k(D_1, g) - gran_k(D_2, g)| = \sum_{i=1}^{m} \underbrace{\frac{leaves_i(\tilde{r}_i)}{|\Omega_i|}}_{\leq 1} \leq m\ .$$

In summary we have:

$$|gran_k(D_1, g) - gran_k(D_2, g)| \leq \begin{cases} (k-1)m, & if k > 1 \\ m, & if k = 1 \end{cases}$$

□

B.2 Discernibility (Theorem 7)

In the following we will frequently employ the triangle inequality and indicate its application with (T). In order to prove the sensitivity of the Discernibility score function we will first propose two lemmas:

Lemma 12. *For all $D_1, D_2 \subseteq (\Omega_1 \cup \Lambda_1) \times ... \times (\Omega_m \cup \Lambda_m)$ with $D_1 = D_2 \cup \{r'\}$ the following holds: $|\phi(D_1) - \phi(D_2)| \leq 5$.*

Proof. If $D_2 = \emptyset$ holds we have $D_1 = \{r'\}$ and can conclude:

$$|\phi(D_1) - \phi(D_2)| = |1 - 0| = 1.$$

In the following we will assume $D_2 \neq \emptyset$ and define $c := |D_1|$, $n := |\{r' \in D_1\}|$, $y := |\{* \in D_1\}|$ and

$$x := \sum_{E \in EQ(D_1): r' \notin E} |E|^2 = \sum_{E \in EQ(D_2): r' \notin E} |E|^2.$$

– If $r' \neq *$ holds we have $|\{r' \in D_2\}| = n - 1$ and $|\{* \in D_2\}| = y$. Moreover, $x + n^2 = \sum_{E \in EQ(D_1)} |E|^2 = \sum_{r \in D_1 : r \neq *} |\{r \in D_1\}| \leq \sum_{r \in D_1} |D_1| = c^2$ holds. We can conclude:

$$\begin{aligned}
&|\phi(D_1) - \phi(D_2)| \\
&= \left| \frac{x + n^2 + yc}{c} - \frac{x + (n-1)^2 + y(c-1)}{c-1} \right| \\
&= \left| \frac{-x - n^2 + 2nc - c}{c(c-1)} \right| \underset{(T)}{\leq} \frac{x + n^2}{c(c-1)} + \frac{2n-1}{c-1} \\
&\leq \frac{c^2}{c(c-1)} + \frac{2c-1}{c-1} = \underbrace{\frac{3c-1}{c-1}}_{\searrow, c \nearrow\ \wedge\ c \geq 2} \leq 5.
\end{aligned}$$

– If $r' = *$ holds we have $|\{* \in D_2\}| = y - 1$ and we can conclude using $x = \sum_{E \in EQ(D_1)} |E|^2 = \sum_{r \in D_1 : r \neq *} |\{r \in D_1\}| \leq \sum_{r \in D_1 \setminus \{*\}} |D_1| \leq c(c-1)$:

$$\begin{aligned}
|\phi(D_1) - \phi(D_2)| &= \left| \frac{x + yc}{c} - \frac{x + (y-1)(c-1)}{c-1} \right| \\
&= \left| \frac{-x + c^2 - c}{c(c-1)} \right| \underset{(T)}{\leq} \frac{x}{c(c-1)} + 1 \leq 2.
\end{aligned}$$

In summary we have $|\phi(D_1) - \phi(D_2)| \leq 5$. □

Lemma 13. *For every integer $k \geq 2$ and all $D_1, D_2 \subseteq (\Omega_1 \cup \Lambda_1) \times ... \times (\Omega_m \cup \Lambda_m)$ satisfying*

$$D_1 = (D_2 \setminus \underbrace{\{*, ..., *\}}_{k-1-times}) \cup \underbrace{\{r', ..., r'\}}_{k-times}$$

with $r' \neq *$ the following holds:

$$|\phi(D_1) - \phi(D_2)| \leq \frac{k^2}{k-1} + 1.$$

Proof. With the definitions

$$\begin{aligned}
c &:= |D_1|, \\
n &:= |\{r' \in D_1\}| = |\{r' \in D_2\}| + k, \\
y &:= |\{* \in D_1\}| = |\{* \in D_2\}| - k + 1, \\
x &:= \sum_{E \in EQ(D_1): r' \notin E} |E|^2 = \sum_{E \in EQ(D_2): r' \notin E} |E|^2
\end{aligned}$$

and using $x + n^2 \leq \sum_{r \in D_1} |D_1| = c^2$ we have:

$$\begin{aligned}
&|\phi(D_1) - \phi(D_2)| \\
&= \left| \frac{x + n^2 + yc}{c} - \frac{x + (n-k)^2 + (y+k-1)(c-1)}{c-1} \right| \\
&= \left| \frac{-x - n^2 + 2knc - k^2c - kc^2 + kc + c^2 - c}{c(c-1)} \right| \\
&\underset{(T)}{\leq} \left| \frac{-x - n^2}{c(c-1)} \right| + k \left| \frac{2n - k - c + 1}{c-1} \right| + 1 \\
&\leq \frac{c}{c-1} + k \left| \frac{2n + 1 - (k + c)}{c-1} \right| + 1. \qquad (8)
\end{aligned}$$

- If $2n+1 \geq k+c$ holds we can conclude:

$$\left|\frac{2n+1-(k+c)}{c-1}\right| = \frac{2n+1-(k+c)}{c-1}$$
$$\underset{n\leq c}{\leq} \frac{2c+1-(k+c)}{c-1} = \frac{c+1-k}{c-1} \underset{k\geq 2}{\leq} 1.$$

- Otherwise we have:

$$\left|\frac{2n+1-(k+c)}{c-1}\right| = \frac{k+c-(2n+1)}{c-1}$$
$$\underset{n\geq k}{\leq} \frac{k+c-(2k+1)}{c-1} = \frac{c-1-k}{c-1} \leq 1.$$

We can conclude from Inequation (8):

$$|\phi(D_1)-\phi(D_2)| \leq \underbrace{\frac{c}{c-1}}_{\searrow, c\nearrow \,\wedge\, c\geq k} +k+1 = \frac{k^2}{k-1}+1. \square$$

We can now prove Theorem 7 as follows:

Proof. Let $k \in \mathbb{N}$ be an arbitrary integer, let $g \in \mathcal{G}_m$ be an arbitrary generalization scheme and let $D_1, D_2 \in \mathcal{D}_m$ be arbitrary datasets satisfying $|D_1 \oplus D_2| = 1$. Without loss of generality we assume $D_1 = D_2 \cup \{r\}$.

- If $S(D_1) = S(D_2) \cup \{*\}$ or $S(D_1) = S(D_2) \cup \{g(r)\}$ holds (which is always satisfied in the case of $g(r) = *$ or $k = 1$) we can conclude using Lemma 12:

$$|disc_k(D_1,g) - disc_k(D_2,g)| \leq 5.$$

- If $k \geq 2$ and $g(r) \neq *$ hold and $g(r)$ is suppressed in $S(D_2)$ but not in $S(D_1)$ we have

$$S(D_1) = (S(D_2) \setminus \underbrace{\{*,...,*\}}_{k-1-times}) \cup \underbrace{\{g(r),...,g(r)\}}_{k-times}$$

and can conclude using Lemma 13:

$$|disc_k(D_1,g) - disc_k(D_2,g)| \leq \frac{k^2}{k-1}+1.$$

In summary we can conclude:

$$|disc_k(D_1,g) - disc_k(D_2,g)| \leq \begin{cases} 5, & ifk = 1 \\ \frac{k^2}{k-1}+1, & ifk > 1 \end{cases} \quad \square$$

B.3 Non-Uniform Entropy (Theorem 8)

We can prove Theorem 8 using the two lemmas proposed in Appendix B.2 as follows:

Proof. Let $k \in \mathbb{N}$ be an arbitrary integer, let $g \in \mathcal{G}_m$ be an arbitrary generalization scheme and let $D_1, D_2 \in \mathcal{D}_m$ be arbitrary datasets satisfying $|D_1 \oplus D_2| = 1$. Without loss of generality we assume $D_1 = D_2 \cup \{r\}$. Then we have:

$$|ent_k(D_1,g) - ent_k(D_2,g)|$$
$$= \left|\sum_{i=1}^{m} \phi(p_i(S(D_1))) - \phi(p_i(S(D_2)))\right|$$
$$\underset{(T)}{\leq} \sum_{i=1}^{m} |\phi(p_i(S(D_1))) - \phi(p_i(S(D_2)))|. \qquad (9)$$

Let us fix an arbitrary $i = 1,...m$, define $g(r) =: (r'_1,...,r'_m)$ and regard $p_i(S(D_1))$ and $p_i(S(D_2))$ as datasets with one attribute.

- If $p_i(S(D_1)) = p_i(S(D_2)) \cup \{*\}$ or $p_i(S(D_1)) = p_i(S(D_2)) \cup \{r'_i\}$ holds (which is always satisfied in the case of $r'_i = *$ or $k = 1$) we can conclude using Lemma 12:

$$|\phi(p_i(S(D_1))) - \phi(p_i(S(D_2)))| \leq 5.$$

- If $k \geq 2$ and $r'_i \neq *$ hold and $g(r)$ is suppressed in $S(D_2)$ but not in $S(D_1)$ we have

$$p_i(S(D_1)) = (p_i(S(D_2)) \setminus \underbrace{\{*,...,*\}}_{k-1-times}) \cup \underbrace{\{r'_i,...,r'_i\}}_{k-times}$$

and can conclude using Lemma 13:

$$|\phi(p_i(S(D_1))) - \phi(p_i(S(D_2)))| \leq \frac{k^2}{k-1}+1.$$

In summary we have

$$|\phi(p_i(S(D_1))) - \phi(p_i(S(D_2)))| \leq \begin{cases} 5, & ifk = 1 \\ \frac{k^2}{k-1}+1, & ifk > 1 \end{cases}$$

and can conclude from Inequation (9):

$$|ent_k(D_1,g) - ent_k(D_2,g)| \leq \begin{cases} 5m, & ifk = 1 \\ (\frac{k^2}{k-1}+1)m, & ifk > 1 \end{cases} \quad \square$$

B.4 Statistical Classification (Theorem 9)

Proof. Let $k \in \mathbb{N}$ be an arbitrary integer, let $g \in \mathcal{G}_m$ be an arbitrary generalization scheme and let $D_1, D_2 \in \mathcal{D}_m$ be arbitrary datasets satisfying $|D_1 \oplus D_2| = 1$. Without loss of generality we assume $D_1 = D_2 \cup \{r\}$. For ease of notation we define $w_1(\cdot) := w(S(D_1), \cdot)$ and $w_2(\cdot) := w(S(D_2), \cdot)$. Moreover, we define FV to be the subset of all records in $S(D_2)$ which have the same combination of feature attribute values as $g(r)$, i.e. $FV := \{r' \in S(D_2) \mid fv(r') = fv(g(r))\}$.

If $fv(g(r))$ is suppressed as a consequence of generalization then $S(D_1)$ and $S(D_2)$ differ only in records with a weight of zero in either set, i.e. we have $C \coloneqq \{r' \in S(D_1) : w_1(r') = 1\} = \{r' \in S(D_2) : w_2(r') = 1\}$ which implies:

$$\begin{aligned} class_k(D_1,g) &= \sum_{r'\in S(D_1)} w_1(r') = \sum_{r'\in C} w_1(r') \\ &= \sum_{r'\in C} w_2(r') = \sum_{r'\in S(D_2)} w_2(r') = class_k(D_2,g). \end{aligned}$$

In the following we will regard the case that $fv(g(r))$ is not suppressed, which implies $g(r) \neq *$.

- If $g(r)$ appears less than k times in $g(D_1)$ then it is suppressed in $S(D_1)$ with $S(D_1) = S(D_2) \cup \{*\}$. We can argue as above: $class_k(D_1,g) = class_k(D_2,g)$.
- If $g(r)$ appears k times in $g(D_1)$ then it is not suppressed in $S(D_1)$ while we have $g(r) \notin S(D_2)$, in particular $g(r) \notin FV$, and

  $$S(D_1) = (S(D_2) \setminus \underbrace{\{*,...,*\}}_{k-1-times}) \,\dot\cup\, \underbrace{\{g(r),...,g(r)\}}_{k-times}.$$

  Moreover, all records in $S(D_2)$ which have the same feature values as $g(r)$ are also contained in $S(D_1)$, i.e. $FV \subseteq S(D_1) \cap S(D_2)$ holds, and these are the only records contained in both $S(D_1)$ and $S(D_2)$ which may have different weights in these sets, i.e. $\forall r' \in (S(D_1) \cap S(D_2)) \setminus FV : w_1(r') = w_2(r')$ holds. We can conclude:

  $$\begin{aligned} &|class_k(D_1,g) - class_k(D_2,g)| = \\ &\left|\left(k \cdot w_1(g(r)) + \sum_{r'\in S(D_1)\cap S(D_2)} w_1(r')\right) - \right. \\ &\left.\left((k-1)\cdot \underbrace{w_2(*)}_{=0} + \sum_{r'\in S(D_1)\cap S(D_2)} w_2(r')\right)\right| = \\ &\left|k \cdot w_1(g(r)) + \sum_{r'\in FV} \big(w_1(r') - w_2(r')\big)\right|. \end{aligned} \tag{10}$$

  Let r'_{maj} denote the record with the most frequent class attribute value among all records in $S(D_1)$ which have the same feature values as $g(r)$.
  If $r'_{maj} \neq g(r)$ holds we have $w_1(g(r)) = 0$ and r'_{maj} is also the record with the most frequent class attribute value in FV with

  $$\forall r' \in FV : w_1(r') = w_2(r') = \begin{cases} 1, & \text{if } r' = r'_{maj} \\ 0, & \text{otherwise} \end{cases}.$$

  Using Equation (10) we can conclude:

  $$|class_k(D_1,g) - class_k(D_2,g)| = 0.$$

  If $r'_{maj} = g(r)$ holds we have $w_1(g(r)) = 1$ and $\forall r' \in FV : w_1(r') = 0$. Moreover, the record $\tilde{r}_{maj}$ with the most frequent class value in FV can appear at most k times in FV (because otherwise, $\tilde{r}_{maj} \in FV \subseteq S(D_1)$ would have a class value more frequent than the one of $g(r)$ in $S(D_1)$, which contradicts $r'_{maj} = g(r)$). Hence, we have:

  $$0 \leq \sum_{r'\in FV} \underbrace{w_2(r')}_{=1 \text{ } \mathit{iff} \text{ } r'=\tilde{r}_{maj}} \leq k.$$

  We can conclude using Equation (10):

  $$|class_k(D_1,g) - class_k(D_2,g)| = k - \sum_{r'\in FV} w_2(r') \leq k.$$

- If $g(r)$ appears $l > k$ times in $g(D_1)$, then it appears $l-1 \geq k$ times in $g(D_2)$. It follows that $g(r)$ is not suppressed in both $S(D_1)$ and $S(D_2)$ with $S(D_1) = S(D_2) \cup \{g(r)\}$. Moreover, $g(r) \in FV \subseteq S(D_2) \subseteq S(D_1)$ holds, and the records in FV are the only ones which may have a different weight in $S(D_1)$ and $S(D_2)$, i.e. $\forall r' \in S(D_2) \setminus FV : w_1(r') = w_2(r')$ holds. We can conclude:

  $$\begin{aligned} &|class_k(D_1,g) - class_k(D_2,g)| \;= \\ &\left|w_1(g(r)) + \sum_{r'\in S(D_2)} w_1(r') - \sum_{r'\in S(D_2)} w_2(r')\right| = \\ &\left|w_1(g(r)) + \sum_{r'\in FV} \big(w_1(r') - w_2(r')\big)\right|. \end{aligned} \tag{11}$$

  If $r'_{maj} \neq g(r)$ holds we can argue similar as above:

  $$|class_k(D_1,g) - class_k(D_2,g)| = 0.$$

  If $r'_{maj} = g(r)$ holds we have

  $$\forall r' \in FV : w_1(r') = \begin{cases} 1, & \text{if } r' = g(r) \\ 0, & \text{otherwise} \end{cases},$$

  $|\{g(r) \in FV\}| = l - 1$ (so that the record with the most frequent class value appears at least $l-1$ times in FV) and $\forall r' \in FV, r' \neq g(r) : |\{r' \in S(D_2)\}| \leq l$ (because otherwise, there would exist a record $\tilde{r}_{maj} \in FV \subseteq S(D_1), \tilde{r}_{maj} \neq g(r)$ with a class value which is more frequent than the one of $g(r)$ in $S(D_1)$, which contradicts $r'_{maj} = g(r)$). Hence we have:

  $$l - 1 \leq \sum_{r'\in FV} w_2(r') \leq l.$$

  We can conclude using Equation (11):

  $$|class_k(D_1,g) - class_k(D_2,g)| = l - \sum_{r'\in FV} w_2(r') \leq 1.$$

In summary we can conclude:

$$|class_k(D_1,g) - class_k(D_2,g)| \leq k \qquad \square$$

A.2 Efficient Protection of Numeric Attributes

Full Title

Efficient Protection of Health Data from Sensitive Attribute Disclosure

Authors

Raffael Bild, Johanna Eicher and Fabian Prasser

Published In

Digital Personalized Health and Medicine: Proceedings of Medical Informatics Europe, 270:193-197, 2020

Copyright

CC BY-NC 4.0 (see Appendix D for legal code. Source: `https://creativecommons.org/licenses/by-nc/4.0/` Accessed: 11.07.2024)

Efficient Protection of Health Data from Sensitive Attribute Disclosure

Raffael BILD [a,1], Johanna EICHER [a] and Fabian PRASSER [b,c]
[a] *University hospital rechts der Isar, Technical University of Munich, Germany*
[b] *Charité - Universitätsmedizin Berlin, Berlin, Germany*
[c] *Berlin Institute of Health (BIH), Berlin, Germany*

Abstract. Biomedical research has become data-driven. To create the required big datasets, health data needs to be shared or reused out of the context of its initial purpose. This leads to significant privacy challenges. Data anonymization is an important protection method where data is transformed such that privacy guarantees can be provided according to formal models. For applications in practice, anonymization methods need to be integrated into scalable and robust tools. In this work, we focus on the problem of scalability.

Protecting biomedical data from inference attacks is challenging, in particular for numeric data. An important privacy model in this context is t-closeness, which has also been defined for attribute values which are totally ordered. However, directly implementing a scalable algorithmic representation of the mathematical definition of the model proves difficult. In this paper we therefore present a series of optimizations that can be used to achieve efficiency in production use. An experimental evaluation shows that our approach reduces execution times of anonymization processes involving t-closeness by up to a factor of two.

Keywords. data protection, anonymization, inference attacks, scalability

1. Introduction

Biomedical research, e.g. in the field of precision medicine which tailors healthcare to characteristics of individuals, is increasingly data-driven and leveraging methods from data science such as machine learning [1]. However, when creating the required big datasets, stringent privacy protection is mandated by laws and regulations. Hence, a wide range of safeguards has to be applied, including organizational and technical measures.

Data anonymization is an important building block for implementing technical privacy protection. The basic idea is to transform data in such a manner that formal guarantees, e.g. regarding the risk of singling out, linkage or inference, can be provided [2]. These formal guarantees are captured by so called *privacy models*. t-Closeness is a state-of-the-art model for protecting data from inference attacks. The model requires that the distribution of sensitive attribute values in a set of indistinguishable data records is not too different from the distribution of sensitive information in the overall dataset [3].

[1] Corresponding Author: Raffael Bild, Institute of Medical Informatics, Statistics and Epidemiology, University Hospital rechts der Isar, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany; E-mail: raffael.bild@tum.de.

2. Objective

The ARX Data Anonymization Tool is among the few software solutions for quantitative data anonymization, that have found wide-spread adoption. ARX focuses on data transformation methods which have been specifically recommended for applications to health data and it implements models for protecting data from singling out, linkage and inference [4]. t-Closeness is amongst the models supported.

t-Closeness has been specified in different variants that apply to variables with different scales of measure. One of these variants focuses on variables which are totally ordered. This model is particularly relevant in practice, as it is one of the few privacy models which have been proposed for protecting sensitive numeric variables.

When using ARX to protect complex datasets using the t-closeness model, however, we realized that the initial implementation is not scalable. Upon further inspection, we realized that directly implementing a scalable algorithmic representation of the mathematical definition of the model t-closeness proves difficult in general. In this paper we therefore present a series of optimizations that we have developed to achieve efficiency in productive use. All of them have been integrated into ARX.

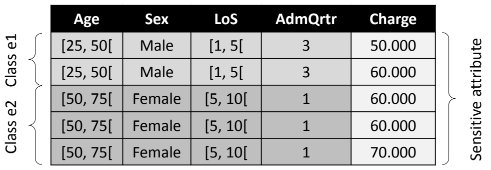
3. Methods

3.1. Problem Definition

t-Closeness is a condition that applies to equivalence classes, i.e. groups of records which are indistinguishable regarding attributes that could be used for linking records. Let $P(e) = (p_1, p_2, ..., p_m)$ be the relative frequency distribution of sensitive values in a given equivalence class e and let $Q = (q_1, q_2, ..., q_m)$ be the relative frequency distribution of sensitive values in the whole dataset. $D[P(e), Q]$ is the distance between the distributions $P(e)$ and Q [3]. It is defined as follows [3]:

$$D[P(e),Q] = \frac{1}{m-1}\sum_{i=1}^{m}\left|\sum_{j=1}^{i}(p_j - q_j)\right|.$$

A dataset fulfills t-closeness with numerical ground distance if for all equivalence classes e, $D[P(e),Q] \leq t$ holds.



Class e1
Class e2
Sensitive attribute

Age	Sex	LoS	AdmQrtr	Charge
[25, 50[	Male	[1, 5[	3	50.000
[25, 50[	Male	[1, 5[	3	60.000
[50, 75[	Female	[5, 10[	1	60.000
[50, 75[	Female	[5, 10[	1	60.000
[50, 75[	Female	[5, 10[	1	70.000

Figure 1. Example discharge dataset. ”LoS” = Length of stay, ”AdmQrtr” = Admission quarter.

Figure 1 shows an example dataset with a sensitive attribute ”Charge” and two equivalence classes $e1$ and $e2$ defined by the values of the other attributes. The distribution of sensitive values is $Q = (\frac{1}{5}, \frac{3}{5}, \frac{1}{5})$ since the values 50.000, 60.000 and 70.000 appear 1, 3 and 1 times in the whole dataset, respectively. For the equivalence classes, we get $P(e1) = (\frac{1}{2}, \frac{1}{2}, 0)$ and $P(e2) = (0, \frac{2}{3}, \frac{1}{3})$. Consequently, we have

$$D[P(e1),Q] = \frac{1}{2}(|\frac{1}{2}-\frac{1}{5}|+|\frac{1}{2}-\frac{1}{5}+\frac{1}{2}-\frac{3}{5}|+|\frac{1}{2}-\frac{1}{5}+\frac{1}{2}-\frac{3}{5}-\frac{1}{5}|) = \frac{1}{4},$$

$$D[P(e2),Q] = \frac{1}{2}(|-\frac{1}{5}|+|-\frac{1}{5}+\frac{2}{3}-\frac{3}{5}|+|-\frac{1}{5}+\frac{2}{3}-\frac{3}{5}+\frac{1}{3}-\frac{1}{5}|) = \frac{1}{6}.$$

Hence, we can conclude that the dataset satisfies t-closeness with $t = max\{\frac{1}{4}, \frac{1}{6}\} = 0.25$.

A straight-forward implementation of t-closeness for fully ordered attributes would implement this by checking if the following inequality holds:

$$|r_1| + |r_1 + r_2| + |r_1 + r_2 + r_3| + ...|r_1 + ... + r_{m-1}| \leq t(m-1)$$

Each r_i has the form $r_i = p_i - q_i$, where p_i is the frequency of the attribute value number i in the currently considered equivalence class e of the transformed data set, and q_i the frequency of the attribute value number i in the entire input dataset.

As we will show in Section 4 this process is highly inefficient. The main reason is that it needs to iterate over all sensitive attribute values contained in the overall dataset. Given that this process needs to be executed for all equivalence classes, the worst-case complexity is $O(n^2)$ where n is the number of records in the dataset.

3.2. Optimization Approaches

In this section, we present three optimizations that we used to improve our initial, straight-forward implementation of the model.

Optimization 1 – Fibonacci hashing: The first optimization adresses the implementation level. One of the most time-consuming aspects of implementing a check for t-closeness is to dynamically group the sensitive attribute values in each class to determine their frequency. The standard data structure used for this purpose are hash tables. ARX already used an efficient implementation provided by the High Performance Primitive Collections for Java library [5]. However, these collections are still much more complex than required, as they for example support updating the data stored in a map. We therefore implemented a simplified hash table using Fibonacci hashing based on the golden ratio to reduce the number of CPU cycles required for adding and querying elements.

Optimization 2 – Check pruning: The second optimization addresses the mathematical definition of the model. Let us assume that the attribute values number $1...k$ in the equivalence class currently under consideration do not occur at all, then $p_1 = p_2 = ... = p_k = 0$ holds and the first k summands in the above condition have the form:

$$|-q_1| + |-q_1 - q_2| + ... + |-q_1 - ... - q_k|.$$

This partial sum depends only on the input dataset and can therefore be precalculated for every possible value of k before the individual equivalence classes of the transformed dataset are checked.

These precalculations can be performed in an initialization step, for ascending values of k until the corresponding subtotal is greater than the threshold, i.e. the following holds:

$$|-q_1| + |-q_1 - q_2| + ... + |-q_1 - ... - q_k| > t(m-1).$$

Let us denote the smallest value of k for which this inequality is fulfilled with x.

When checking whether t-closeness holds for a specific equivalence class, as a first step, we calculate the smallest index of any attribute value occurring in the class. We call this index y. When $y > x$ it can be inferred that the following summands are included in the relevant sum:

$$|-q_1| + |-q_1 - q_2| + ... + |-q_1 - ... - q_x|.$$

It follows that the threshold $t(m-1)$ will definitely be exceeded and computations can already be stopped at this point (concluding that privacy guarantees are not fulfilled).

Optimization 3 – Summand pruning: The third optimization adds an additional pruning mechanism using pre-computations.

It can be used in cases where the pruning strategy described above is not applicable, i.e. if $y \leq x$ holds. It works by starting the summation at position y, using an appropriate sum which has been pre-calcuated ahead of time for all $k \leq x$ as a starting point:

$$|-q_1| + |-q_1 - q_2| + ... + |-q_1 - ... - q_k|.$$

3.3. Experimental Design

To evaluate our approach, we used the following data from registries and health surveys from the U.S.: 100,937 records about traffic accidents from the NHTSA Fatality Analysis Reporting System (FARS), 539,253 records from the American Time Use Survey (ATUS) and 1,193,504 records from the Integrated Health Interview Series (IHIS). Moreover, we analyzed a subset of a synthetic discharge dataset which is particularly hard to protect from sensitive attribute disclosure (SPD) [6]. We also included two de-facto standard datasets for benchmarking anonymization methods: 30,162 records from the 1994 U.S. Census (ADULT) and 63,441 records from the 1998 KDD competition (CUP). For a detailed specification of the datasets we refer to [7].

We anonymized the datasets with attribute generalization and record suppression to produce output datasets which fulfill *t-closeness* for fully ordered attributes. We varied the risk threshold t $(0.5 \geq t \geq 0.1)$ to study the effect of our optimizations on different parameterizations. As a baseline, we used our original, unoptimized implementation. In the software, both pruning strategies are combined into a common implementation and we therefore present one measurement capturing both of them.

4. Results

Figure 2 shows the results of our experiments. As can be seen, our optimizations improved execution times by up two a factor of more than two. Each optimization had a positive effect in all setups while the degree of effectiveness of each optimization varied between setups. Pruning was possible in between 69% and 99% of the checks performed, but the effect on execution times varied.

The differences in the impact of optimizations on execution times can be explained by considering the distribution of sensitive attributes values in the different datasets. The impact was lower for datasets with a small number of distinct values, i.e. ADULT (7), ATUS (7) and IHIS (10), higher for datasets with more distinct sensitive values, i.e. CUP (81) and FARS (20). We measured the strongest effect for the SPD dataset, which has 101 different sensitive attribute values. When the number of sensitive attribute values is high, execution times are also higher, implying the our optimizations are more effective exactly in the cases where they are needed the most.

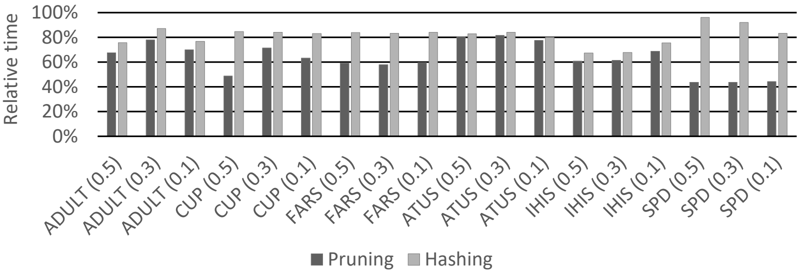


Figure 2. Execution times relative to the original implementation for various datasets and risk thresholds.

5. Conclusion

Due to its usability, flexibility and scalability, ARX is actively used in many areas, including commercial big data analytics platforms, medical research projects, clinical trial data sharing and for training purposes. An important reason for ARX's scalability are the many optimizations that have been integrated into the software. In prior work we have, for example, presented methods to improve the scalability of optimization algorithms for trading off risks vs. utility [8] and a versatile optimized runtime environment for anonymization algorithms [9].

In this paper, we have presented an optimization affecting a specific and important privacy model only. Our approach addresses the implementation level as well as the mathematical definition of the model implemented. Our solution utilizes high-performance data structures as well as pre-computation techniques. The model is particularly relevant in practice, as it is one of the few approaches which can be used to protect sensitive numeric data.

References

[1] V. Gligorijević et al., Integrative methods for analyzing big data in precision medicine, *Proteomics* **16** (2016), 741–758.

[2] B. C. M. Fung, K. Wang, A. W.-C. Fu and P. S. Yu, *Introduction to privacy-preserving data publishing: Concepts and techniques*, CRC Press, 1st ed., 2010, ISBN 9781420091489.

[3] N. Li, T. Li and S. Venkatasubramanian, t-closeness: Privacy beyond k-anonymity and l-diversity, *2007 IEEE 23rd International Conference on Data Engineering*, IEEE, 2007, 106–115.

[4] F. Prasser and F. Kohlmayer, Putting statistical disclosure control into practice: The ARX data anonymization tool, *Medical data privacy handbook*, Springer, 2015, 111–148.

[5] S. Osinski and D. Weiss, HPPC: High Performance Primitive Collections for Java, `https://labs.carrotsearch.com/hppc.html`, accessed: October 10, 2019.

[6] D. Sánchez, S. Martínez and J. Domingo-Ferrer, Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata", *Science* **351** (2016), 1274–1274.

[7] F. Prasser, F. Kohlmayer and K. A. Kuhn, A benchmark of globally-optimal anonymization methods for biomedical data, *2014 IEEE 27th International Symposium on Computer-Based Medical Systems*, IEEE, 2014, 66–71.

[8] F. Prasser, F. Kohlmayer and K. A. Kuhn, Efficient and effective pruning strategies for health data de-identification, *BMC medical informatics and decision making* **16** (2016), 49.

[9] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper and K. A. Kuhn, Highly efficient optimal k-anonymity for biomedical datasets, *2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, 2012, 1–6.

A.3 Reliable Data Anonymization

Full Title

Better Safe Than Sorry – Implementing Reliable Health Data Anonymization

Authors

Raffael Bild, Klaus A. Kuhn and Fabian Prasser

Published In

Digital Personalized Health and Medicine: Proceedings of Medical Informatics Europe, 270:68-72, 2020

Copyright

CC BY-NC 4.0 (see Appendix D for legal code. Source: `https://creativecommons.org/licenses/by-nc/4.0/` Accessed: 11.07.2024)

September 2019

Better Safe Than Sorry - Implementing Reliable Health Data Anonymization

Raffael BILD [a,1], Klaus A. KUHN [a] and Fabian PRASSER [b,c]

[a] *University hospital rechts der Isar, Technical University of Munich, Germany*
[b] *Charité – Universitätsmedizin Berlin, Berlin, Germany*
[c] *Berlin Institute of Health (BIH), Berlin, Germany*

Abstract. Modern biomedical research is increasingly data-driven. To create the required big datasets, health data needs to be shared or reused, which often leads to privacy challenges. Data anonymization is an important protection method where data is transformed such that privacy guarantees can be provided according to formal models. For applications in practice, anonymization methods need to be integrated into scalable and reliable tools. In this work, we tackle the problem of achieving reliability. Privacy models often involve mathematical definitions using real numbers which are typically approximated using floating-point numbers when implemented as software. We study the effect on the privacy guarantees provided and present a reliable computing framework based on fractional and interval arithmetic for improving the reliability of implementations. Extensive evaluations demonstrate that reliable data anonymization is practical and that it can be achieved with minor impacts on executions times and data utility.

Keywords. data protection, anonymization, reliable computing

1. Introduction

Modern data-driven biomedical research, e.g. in the field of precision medicine which tailors healthcare to the characteristics of individuals, increasingly leverages data science methods such as machine learning [1]. However, when creating the required big datasets, laws and regulations mandate stringent privacy protection. Hence, a wide range of safeguards has to be applied, ranging from organizational to technical measures.

Data anonymization is an important technical building block for implementing privacy protection. In this process, data is transformed in such a manner that formal guarantees, e.g. regarding the risk of singling out, linkage or inference, can be provided. Traditional models of privacy protection such as k-anonymity, ℓ-diversity and t-closeness specify syntactic constraints on output datasets, while more recent models like differential privacy formulate requirements for data processing methods [2].

2. Objective

All methods for implementing privacy models require performing changes to data which inevitably leads to a decrease of its utility. To balance a decrease in privacy risks with a

[1] Corresponding Author: Raffael Bild, Institute of Medical Informatics, Statistics and Epidemiology, School of Medicine, Technical University of Munich, Ismaninger Straße 22, 81675 Munich, Germany; E-mail: raffael.bild@tum.de.

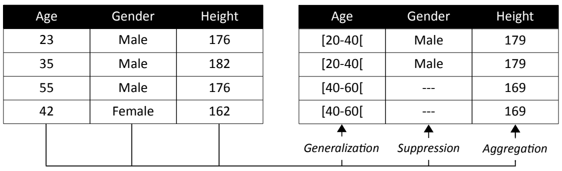
decrease of utility, models for quantifying both aspects have been developed. When implementing privacy models in practice, an important challenge lies in the need to reflect their mathematical definitions in software. Privacy models are often formulated over real numbers, which in software are approximated by floating-point numbers with limited precision (typically 64 bits). Computations can therefore yield results that differ significantly from the exact mathematical results [3]. This can make output data of anonymization tools vulnerable to privacy breaches. For example, it has been shown that straightfoward implementations of a common method for achieving differential privacy can be exploited to extract the entire content of a (presumably protected) dataset [4]. However, studies of the effects of floating-point errors on the privacy guarantees provided by other methods for data anonymization are still lacking in the literature.

In this article, we aim to fill this gap, with a focus on investigating further methods which are truthful (i.e. non-perturbative) and hence particularly well-suited for the biomedical domain [5]. For this, we discuss the numerical stability of implementations of various privacy models, including k-anonymity, ℓ-diversity, t-closeness and further methods for achieving differential privacy [2]. Moreover, we present a reliable computing framework, which we have integrated into the open source data anonymization tool ARX [6] to mitigate vulnerabilities resulting from the use of floating-point operations.

3. Methods

3.1. Data Anonymization and Floating-Point Arithmetic

Figure 1 shows an example transformation of an input dataset using a combination of generalization (i.e. the replacement of values with more general, but semantically consistent values), suppression (i.e. the removal of values) and aggregation (i.e. the replacement of values with an aggregate, such as their mean). The example output dataset satisfies 2-anonymity, which means that each combination of attribute values appears at least twice (see [2] for further details). Whether or not k-anonymity is satisfied is easy to determine by simply counting the size of groups of indistinguishable records. Implementing other privacy models, such as ℓ-diversity, t-closeness or differential privacy, requires evaluating mathematical expressions over real numbers, though (cf. Section 3.2).



Age	Gender	Height
23	Male	176
35	Male	182
55	Male	176
42	Female	162

Age	Gender	Height
[20-40[	Male	179
[20-40[	Male	179
[40-60[	---	169
[40-60[	---	169

Figure 1. Example of input data and transformed output data.

In computers, real numbers are typically approximated using floating-point numbers. The number of floating-point values which can be represented with a fixed number of bits (typically 64) is finite. Hence, there exists an infinite number of unrepresentable real numbers. Most implementations of floating-point arithmetic adopt the IEEE standard 754 [7]. It specifies that all floating-point operations have to be performed as if it was possible to perform the corresponding operation with infinite precision, and then to round the result to a representable number. This inevitably introduces rounding errors which add up during sequences of calculations [3].

3.2. Numerical Stability of Common Privacy Models

Implementing some privacy models supported by ARX, e.g. k-anonymity [2], doesn't require decimal numbers at all. Implementing others requires significant amounts of decimal arithmetic, though. Examples are (1) t-closeness which basically requires that the distribution of sensitive attribute values in a set of indistinguishable data records is not too different from the corresponding distribution in the overall dataset or (2) (entropy) ℓ-diversity which requires the distribution $(p_1,...,p_m)$ of sensitive attribute values in each group of indistinguishable records to satisfy $-\sum_{i=1}^{m} p_i \ln(p_i) \geq \ln(\ell)$ [2]. However, by studying possible effects of floating-point error propagation using forward analyses (see e.g. [3] for details), we were able to derive upper bounds for the resulting additive exceedances of the privacy parameters of these models. While a detailed presentation of these analyses is beyond the scope of this article, they showed that the resulting privacy violations are negligible in practice for all syntactic privacy models supported by ARX.

Differential privacy is not a property of a dataset, but a property of a data processing method. It basically guarantees that the probability of any possible output of a probabilistic algorithm does not change "by much" if data of one individual is added to or removed from the input dataset. The Laplacian mechanism and the exponential mechanism are important building blocks for implementating differentially private algorithms [8]. In [9], we have presented a process for implementing k-anonymity while fulfilling (ε,δ)-differential privacy. This approach uses random sampling to introduce non-determinism and the exponential mechanism to optimize the utility of output data. Consequently, unlike the majority of differentially private algorithms, it is truthful and therefore well-suited for processing health data [5].

We were able to calculate an upper bound on the rounding error induced by straight-forward floating-point implementations of the exponential mechanism. For this, we applied conservative methodologies described e.g. in [3] followed by an extension of the original proof of the privacy guarantees provided [8] which takes rounding errors into account. While a detailed presentation of this analysis is, again, out of the scope of this article, it showed that the additive exceedance of the expected privacy loss ε is negligible, with values of about 10^{-10} or less in practical settings.

However, the implementation of differential privacy in ARX requires complex calculations to determine the sampling fraction β and the parameter k for k-anonymity to guarantee the requested degree of privacy protection. Investigating our floating-point implementation, we found that the deviations of ε were in the order of 10^{-16} using common values of ε (e.g. 0.01, 0.1, 0.5, ln(2), 0.75, 1, ln(3), 1.25, 1.5 and 2). The actual values calculated for the parameter δ, however, deviated drastically, as is shown in Table 1.

Table 1. Relative error of δ for $\varepsilon = \ln(2)$ using a floating-point implementation.

Requested value of δ	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-8}	10^{-9}
Error of δ [%]	13.3	11.1	17.2	28.4	10.0	27.9

3.3. Design of ARX's Reliable Computing Framework

To solve this problem, we implemented a framework comprising different computing technologies that are reliable, i.e. offer strict guarantees for the accuracy of the calculated results: *(1)* Arithmetic using exact arbitrary-precision floating-point numbers. This can be used for calculations involving numbers with a finite amount of digits only. *(2)* Using representations as fractions with arbitrarily long integer enumerators and denominators. This approach can be used to perform exact calculations over rational numbers but it can become very slow. *(3)* Interval arithmetic [3], which dynamically computes strict bounds on the errors of mathematical operations. The basic idea is not to operate on (approximations of) real numbers, but rather on intervals which enclose the respective exact real numbers. Functions operating on such intervals yield intervals which are guaranteed to include the exact result. For example, addition can be performed by calculating

$$[x_1,x_2]+[y_1,y_2]=[x_1+y_1,x_2+y_2].$$

ARX is implemented in Java and arbitrary prevision arithmetic and fraction arithmetic is well supported by common programming libraries. The number of Java libraries for performing interval arithmetic is, however, known to be limited [10]. Hence, we implemented a basic interval arithmetic framework from scratch while focusing on a minimal amount of easily understandable and verifiable code. We implemented various operators, including the basic arithmetic operators. For more complex functions such as *exp*, *pow*, *log* and *sqrt* we invoke the respective implementations for floating-point values provided by the Java standard library which have clearly defined accuracies. We also implemented various comparison operators such as $\leq$ which allow for reliable comparisons by returning the result of comparing the upper and lower endpoint of their operands. These operators are guarded by checks which raise an error if the relation of their operands is not decidable, i.e. if the intervals are overlapping.

We used the methods and operators provided by this framework to implement the parameter calculation process for differential privacy reliably so that the actual degree of privacy protection provided can be at most more conservative than specified by the user.

4. Results

To evaluate the impact of the reliable parameter calculation on the strictness of the derived parameters we have compared it with a straight-forward floating-point implementation using common values of ε ranging from 0.01 to 2 and $\delta = 10^{-i}$ for $i = 1,...,9$.

The differences between the values of β obtained using both implementations were very small with values of about 10^{-16} in all cases. All values obtained for k were identical except for ten configurations using irrational parameters. This is because these numbers have more significant figures than the other values considered, which resulted in higher rounding errors and hence larger intervals during calculations. In these cases, the values of k obtained reliably were (slightly) higher. Using $\varepsilon = \ln(3)$, the values of k computed differed for $\delta = 10^{-5}$ and $\delta = 10^{-6}$ ($k = 63$ vs. $k = 66$ and $k = 78$ vs. $k = 82$, respectively). The results obtained when using $\varepsilon = \ln(2)$ are listed in Table 2. As can be seen, the absolute differences were at most two. Consequently, for decreasing values of δ, which correspond to increasing degrees of privacy protection, the relative differences between the values of k obtained by both implementations tended to become smaller.

Table 2. Values of k derived from various values of δ and $\varepsilon = \ln(2)$.

δ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}
Floating-Point	7	18	30	42	54	67	81	93	105
Reliable	8	20	32	44	56	68	81	95	107

In contrast to results obtained using the floating-point implementation (cf. Table 1), the actual values of ε and δ resulting from reliably calculated parameters k and β were at most more conservative than the privacy parameters specified by the user. In particular, increasing k was necessary to mitigate the violations of δ reported in Table 1. At the same time, the impacts on the intensity of data transformations applied and hence the potential reductions of data utility are negligible when using recommended parameterizations [9].

We also evaluated the execution times of both implementations on a PC with a quad-core 3.1 GHz CPU, Ubuntu Linux and an Oracle JVM. The results are shown in Figure 2. When decreasing both ε and δ (which corresponds to stronger degrees of protection), the relative execution times tended to increase. Using typical values of $\varepsilon \approx 1$ and $\delta \approx 10^{-6}$, the execution time of the reliable implementation was about four times higher than the time used by the floating-point implementation. In all experiments with $\varepsilon \geq 0.1$, the calculation of parameters terminated in less than one second using both implementations. This contains the range of parameters which is practical for the approach (cf. [9]).

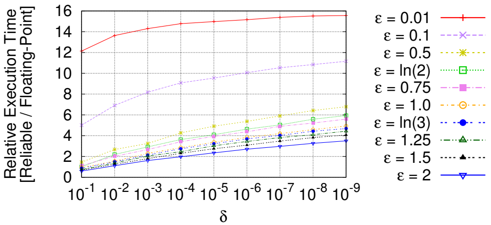


Figure 2. Execution times for deriving β and k from various values of δ and ε reliably relative to the floating-point implementation.

5. Conclusion

In this article, we studied how privacy violations resulting from floating-point implementations of anonymization algorithms can be mitigated. We discussed reliability issues resulting from arithmetic operations for a variety of privacy models including k-anonymity, ℓ-diversity and t-closeness as well as an implementation of differential privacy specifically suited for applications to health data [9]. Moreover, we presented a framework comprising reliable computing techniques, including interval and fractional arithmetic. All results have been integrated into the open source tool ARX. Finally, we examined the impacts of the reliable implementation on output data utility as well as execution times and found both to be negligible in practice when realistic parameters are being used.

References

[1] V. Gligorijević et al., Integrative methods for analyzing big data in precision medicine, *Proteomics* **16** (2016), 741–758.

[2] B. C. Fung et al., *Introduction to privacy-preserving data publishing: Concepts and techniques*, CRC Press, 2010.

[3] N. J. Higham, *Accuracy and stability of numerical algorithms*, vol. 80, Siam, 2002.

[4] I. Mironov, On significance of the least significant bits for differential privacy, *Proceedings of the 2012 ACM conference on computer and communications security*, ACM, 2012, 650–661.

[5] F. K. Dankar and K. El Emam, Practicing differential privacy in health care: A review., *Transactions on data privacy* **6** (2013), 35–67.

[6] F. Prasser and F. Kohlmayer, Putting statistical disclosure control into practice: The ARX data anonymization tool, *Medical data privacy handbook*, Springer, 2015, 111–148.

[7] IEEE Standards Committee et al., 754-2008 IEEE standard for floating-point arithmetic, *IEEE computer society std* **2008**.

[8] C. Dwork, A. Roth et al., The algorithmic foundations of differential privacy, *Foundations and trends® in theoretical computer science* **9** (2014), 211–407.

[9] R. Bild, K. A. Kuhn and F. Prasser, Safepub: A truthful data anonymization algorithm with strong privacy guarantees, *Proceedings on privacy enhancing technologies* **2018** (2018), 67–87.

[10] J. W. von Gudenberg, OOP and interval arithmetic–Language support and libraries, *Numerical software with result verification*, Springer, 2004, 1–14.

APPENDIX B

Further Publications

1. Thierry Meurers, **Raffael Bild**, Kieu-Mi Do, and Fabian Prasser. A Scalable Software Solution for Anonymizing High-Dimensional Biomedical Data. *GigaScience*, 10(10):giab068, 2021.

2. **Raffael Bild**, Martin Bialke, Karoline Buckow, Thomas Ganslandt, Kristina Ihrig, Roland Jahns, Angela Merzweiler, Sybille Roschka, Björn Schreiweis, Sebastian Stäubert, Sven Zenker, and Fabian Prasser. Towards a Comprehensive and Interoperable Representation of Consent-Based Data Usage Permissions in the German Medical Informatics Initiative. *BMC Medical Informatics and Decision Making*, 20(1):1-9, 2020.

3. Johanna Eicher, **Raffael Bild**, Helmut Spengler, Klaus A Kuhn, and Fabian Prasser. A Comprehensive Tool for Creating and Evaluating Privacy-Preserving Biomedical Prediction Models. *BMC Medical Informatics and Decision Making*, 20(1):1-14, 2020.

4. Fabian Prasser, Johanna Eicher, Helmut Spengler, **Raffael Bild**, and Klaus A Kuhn. Flexible Data Anonymization Using ARX — Current Status and Challenges Ahead. *Software: Practice and Experience*, 50(7):1277-1304, 2020.

5. Fabian Prasser, Helmut Spengler, **Raffael Bild**, Johanna Eicher, and Klaus A Kuhn. Privacy-Enhancing ETL-Processes for Biomedical Data. In *International Journal of Medical Informatics*, 126:72-81, 2019.

6. Fabian Prasser, Johanna Eicher, **Raffael Bild**, Helmut Spengler, and Klaus A Kuhn. A Tool for Optimizing Anonymized Health Data for Use in Statistical Classification. In *Proceedings of the IEEE International Symposium on Computer Based Medical Systems*, pages 169-174, 2017.

7. Marco Brandizi, Olga Melnichuk, **Raffael Bild**, Florian Kohlmayer, Benedicto Rodriguez-Castro, Helmut Spengler, Klaus A Kuhn, Wolfgang Kuchinke, Christian Ohmann, Timo Mustonen, Mikael Linden, Tommi Nyrönen, Ilkka Lappalainen, Alvis Brazma, and Ugis Sarkans. Orchestrating Differential Data Access For Translational Research: A Pilot Implementation. *BMC Medical Informatics and Decision Making*, 17(1):30, 2017.

8. Fabian Prasser, **Raffael Bild**, and Klaus A Kuhn. A Generic Method for Assessing the Quality of De-Identified Health Data. In *Medical Informatics Europe*, pages 312-316, 2016.

9. Fabian Prasser, **Raffael Bild**, Johanna Eicher, Helmut Spengler, Florian Kohlmayer, and Klaus A Kuhn. Lightning: Utility-Driven Anonymization of High-Dimensional Data. *Transactions on Data Privacy*, 9(2):161-185, 2016.

10. Roxana Merino-Martinez, Loreana Norlin, David van Enckevort, Gabriele Anton, Simone Schuffenhauer, Kaisa Silander, Linda Mook, Petr Holub, **Raffael Bild**, and Morris Swertz. Toward Global Biobank Integration by Implementation of the Minimum Information About Biobank Data Sharing (MIABIS 2.0 Core). *Biopreservation and Biobanking*, 14(4):298-306, 2016.

11. Klaus A Kuhn, **Raffael Bild**, Gabriele Anton, Simone Schuffenhauer, and H-Erich Wichmann. Connecting Biobanks of Large European Cohorts (EU Project BBMRI-LPC). *Bundesgesundheitsblatt*, 59(3):385–389, 2016.

APPENDIX C

Legal Code CC BY-NC-ND 3.0

Legal Code - Attribution-NonCommercial-NoDerivs 3.0 Unported - Creative Commons

Creative Commons public licenses provide a standard set of terms and conditions that creators and other rights holders may use to share original works of authorship and other material subject to copyright and certain other rights specified in the public license below. The following considerations are for informational purposes only, are not exhaustive, and do not form part of our licenses.

Our public licenses are intended for use by those authorized to give the public permission to use material in ways otherwise restricted by copyright and certain other rights. Our licenses are irrevocable. Licensors should read and understand the terms and conditions of the license they choose before applying it. Licensors should also secure all rights necessary before applying our licenses so that the public can reuse the material as expected. Licensors should clearly mark any material not subject to the license. This includes other CC-licensed material, or material used under an exception or limitation to copyright. More considerations for licensors.

By using one of our public licenses, a licensor grants the public permission to use the licensed material under specified terms and conditions. If the licensor's permission is not necessary for any reason–for example, because of any applicable exception or limitation to copyright–then that use is not regulated by the license. Our licenses grant only permissions under copyright and certain other rights that a licensor has authority to grant. Use of the licensed material may still be restricted for other reasons, including because others have copyright or other rights in the material. A licensor may make special requests, such as asking that all changes be marked or described. Although not required by our licenses, you are encouraged to respect those requests where reasonable. More considerations for the public.

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN "AS-IS" BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE.

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

a. **Adaptation** means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered an Adaptation for the purpose of this License.

b. **Collection** means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(f) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined above) for the purposes of this License.

c. **Distribute** means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.

d. **Licensor** means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.

e. **Original Author** means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

f. **Work** means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.

g. **You** means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

h. **Publicly Perform** means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.

i. **Reproduce** means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.

2. Fair Dealing Rights.

Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant.

Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

a. to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections;

b. to Distribute and Publicly Perform the Work including as incorporated in Collections

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(e) , all rights not expressly granted by Licensor are hereby reserved, including but not limited to the rights set forth in Section 4(d) .

4. Restrictions.

The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a. You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c) , as requested.

b. You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file-sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in con-nection with the exchange of copyrighted works.

c. If You Distribute, or Publicly Perform the Work or Collections, You must, unless a request has been made pursuant to Section 4(a) , keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work. The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

d. For the avoidance of doubt:

   i. **Non-waivable Compulsory License Schemes** . In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

   ii. **Waivable Compulsory License Schemes** . In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License if Your exercise of such rights is for a purpose or use which is otherwise than noncommercial as permitted under Section 4(c) , and otherwise waives the right to collect royalties through any statutory or compulsory licensing scheme; and,

   iii. **Voluntary License Schemes** . The Licensor reserves the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License that is for a purpose or use which is otherwise than noncommercial as permitted under Section 4(c) , .

e. Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING AND TO THE FULLEST EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO THIS EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability.

EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

a. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

b. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

a. Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

b. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

c. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

d. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

e. The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty

provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

Creative Commons is the nonprofit behind the open licenses and other legal tools that allow creators to share their work. Our legal tools are free to use.

APPENDIX D

Legal Code CC BY-NC 4.0

Legal Code - Attribution-NonCommercial 4.0 International - Creative Commons

Version 4.0 • See the errata page for any corrections and the date of change

About the license and Creative Commons

Creative Commons Corporation ("Creative Commons") is not a law firm and does not provide legal services or legal advice. Distribution of Creative Commons public licenses does not create a lawyer-client or other relationship. Creative Commons makes its licenses and related information available on an "as-is" basis. Creative Commons gives no warranties regarding its licenses, any material licensed under their terms and conditions, or any related information. Creative Commons disclaims all liability for damages resulting from their use to the fullest extent possible.

Creative Commons public licenses provide a standard set of terms and conditions that creators and other rights holders may use to share original works of authorship and other material subject to copyright and certain other rights specified in the public license below. The following considerations are for informational purposes only, are not exhaustive, and do not form part of our licenses.

Our public licenses are intended for use by those authorized to give the public permission to use material in ways otherwise restricted by copyright and certain other rights. Our licenses are irrevocable. Licensors should read and understand the terms and conditions of the license they choose before applying it. Licensors should also secure all rights necessary before applying our licenses so that the public can reuse the material as expected. Licensors should clearly mark any material not subject to the license. This includes other CC-licensed material, or material used under an exception or limitation to copyright. More considerations for licensors.

By using one of our public licenses, a licensor grants the public permission to use the licensed material under specified terms and conditions. If the licensor’s permission is not necessary for any reason–for example, because of any applicable exception or limitation to copyright–then that use is not regulated by the license. Our licenses grant only permissions under copyright and certain other rights that a licensor has authority to grant. Use of the licensed material may still be restricted for other reasons, including because others have copyright or other rights in the material. A licensor may make special requests, such as asking that all changes be marked or described. Although not required by our licenses, you are encouraged to respect those requests where reasonable. More considerations for the public.

Attribution-NonCommercial 4.0 International

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution-NonCommercial 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Section 1 – Definitions.

a. Adapted Material means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed relation with a moving image.
b. Adapter's License means the license You apply to Your Copyright and Similar Rights in Your contributions to Adapted Material in accordance with the terms and conditions of this Public License.
c. Copyright and Similar Rights means copyright and/or similar rights closely related to copyright including, without limitation, performance, broadcast, sound recording, and Sui Generis Database Rights, without regard to how the rights are labeled or categorized. For purposes of this Public License, the rights specified in Section 2(b)(1)-(2) are not Copyright and Similar Rights.
d. Effective Technological Measures means those measures that, in the absence of proper authority, may not be circumvented under laws fulfilling obligations under Article 11 of the WIPO Copyright Treaty adopted on December 20, 1996, and/or similar international agreements.
e. Exceptions and Limitations means fair use, fair dealing, and/or any other exception or limitation to Copyright and Similar Rights that applies to Your use of the Licensed Material.
f. Licensed Material means the artistic or literary work, database, or other material to which the Licensor applied this Public License.
g. Licensed Rights means the rights granted to You subject to the terms and conditions of this Public License, which are limited to all Copyright and Similar Rights that apply to Your use of the Licensed Material and that the Licensor has authority to license.
h. Licensor means the individual(s) or entity(ies) granting rights under this Public License.
i. NonCommercial means not primarily intended for or directed towards commercial advantage or monetary compensation. For purposes of this Public License, the exchange of the Licensed Material for other material subject to Copyright and Similar Rights by digital file-sharing or similar means is NonCommercial provided there is no payment of monetary compensation in connection with the exchange.
j. Sui Generis Database Rights means rights other than copyright resulting from Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, as amended and/or succeeded, as well as other essentially equivalent rights anywhere in the world.
k. You means the individual or entity exercising the Licensed Rights under this Public License. **Your** has a corresponding meaning.

Section 2 – Scope.

a. **License grant** .
   1. Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to:
      A. reproduce and Share the Licensed Material, in whole or in part, for NonCommercial purposes only; and
      B. produce, reproduce, and Share Adapted Material for NonCommercial purposes only.
   2. **Exceptions and Limitations** . For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions.
   3. **Term** . The term of this Public License is specified in Section 6(a) .
   4. **Media and formats; technical modifications allowed** . The Licensor authorizes You to exercise the Licensed Rights in all media and formats whether now

known or hereafter created, and to make technical modifications necessary to do so. The Licensor waives and/or agrees not to assert any right or authority to forbid You from making technical modifications necessary to exercise the Licensed Rights, including technical modifications necessary to circumvent Effective Technological Measures. For purposes of this Public License, simply making modifications authorized by this Section 2(a)(4) never produces Adapted Material.

5. Downstream recipients .
   - A. Offer from the Licensor – Licensed Material . Every recipient of the Licensed Material automatically receives an offer from the Licensor to exercise the Licensed Rights under the terms and conditions of this Public License.
   - B. No downstream restrictions . You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, the Licensed Material if doing so restricts exercise of the Licensed Rights by any recipient of the Licensed Material.
6. No endorsement . Nothing in this Public License constitutes or may be construed as permission to assert or imply that You are, or that Your use of the Licensed Material is, connected with, or sponsored, endorsed, or granted official status by, the Licensor or others designated to receive attribution as provided in Section 3(a)(1)(A)(i) .

b. **Other rights** .

1. Moral rights, such as the right of integrity, are not licensed under this Public License, nor are publicity, privacy, and/or other similar personality rights; however, to the extent possible, the Licensor waives and/or agrees not to assert any such rights held by the Licensor to the limited extent necessary to allow You to exercise the Licensed Rights, but not otherwise.
2. Patent and trademark rights are not licensed under this Public License.
3. To the extent possible, the Licensor waives any right to collect royalties from You for the exercise of the Licensed Rights, whether directly or through a collecting society under any voluntary or waivable statutory or compulsory licensing scheme. In all other cases the Licensor expressly reserves any right to collect such royalties, including when the Licensed Material is used other than for NonCommercial purposes.

Section 3 – License Conditions.

Your exercise of the Licensed Rights is expressly made subject to the following conditions.

a. **Attribution** .

1. If You Share the Licensed Material (including in modified form), You must:

   - A. retain the following if it is supplied by the Licensor with the Licensed Material:
     - i. identification of the creator(s) of the Licensed Material and any others designated to receive attribution, in any reasonable manner requested by the Licensor (including by pseudonym if designated);
     - ii. a copyright notice;
     - iii. a notice that refers to this Public License;
     - iv. a notice that refers to the disclaimer of warranties;
     - v. a URI or hyperlink to the Licensed Material to the extent reasonably practicable;
   - B. indicate if You modified the Licensed Material and retain an indication of any previous modifications; and
   - C. indicate the Licensed Material is licensed under this Public License, and include the text of, or the URI or hyperlink to, this Public License.
2. You may satisfy the conditions in Section 3(a)(1) in any reasonable manner based on the medium, means, and context in which You Share the Licensed Material. For example, it may be reasonable to satisfy the conditions by providing a URI or hyperlink to a resource that includes the required information.

3. If requested by the Licensor, You must remove any of the information required by Section 3(a)(1)(A) to the extent reasonably practicable.
4. If You Share Adapted Material You produce, the Adapter's License You apply must not prevent recipients of the Adapted Material from complying with this Public License.

Section 4 – Sui Generis Database Rights.

Where the Licensed Rights include Sui Generis Database Rights that apply to Your use of the Licensed Material:

a. for the avoidance of doubt, Section 2(a)(1) grants You the right to extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database for NonCommercial purposes only;
b. if You include all or a substantial portion of the database contents in a database in which You have Sui Generis Database Rights, then the database in which You have Sui Generis Database Rights (but not its individual contents) is Adapted Material; and
c. You must comply with the conditions in Section 3(a) if You Share all or a substantial portion of the contents of the database.

For the avoidance of doubt, this Section 4 supplements and does not replace Your obligations under this Public License where the Licensed Rights include other Copyright and Similar Rights.

Section 5 – Disclaimer of Warranties and Limitation of Liability.

a. **Unless otherwise separately undertaken by the Licensor, to the extent possible, the Licensor offers the Licensed Material as-is and as-available, and makes no representations or warranties of any kind concerning the Licensed Material, whether express, implied, statutory, or other. This includes, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of latent or other defects, accuracy, or the presence or absence of errors, whether or not known or discoverable. Where disclaimers of warranties are not allowed in full or in part, this disclaimer may not apply to You.**
b. **To the extent possible, in no event will the Licensor be liable to You on any legal theory (including, without limitation, negligence) or otherwise for any direct, special, indirect, incidental, consequential, punitive, exemplary, or other losses, costs, expenses, or damages arising out of this Public License or use of the Licensed Material, even if the Licensor has been advised of the possibility of such losses, costs, expenses, or damages. Where a limitation of liability is not allowed in full or in part, this limitation may not apply to You.**
c. The disclaimer of warranties and limitation of liability provided above shall be interpreted in a manner that, to the extent possible, most closely approximates an absolute disclaimer and waiver of all liability.

Section 6 – Term and Termination.

a. This Public License applies for the term of the Copyright and Similar Rights licensed here. However, if You fail to comply with this Public License, then Your rights under this Public License terminate automatically.

b. Where Your right to use the Licensed Material has terminated under Section 6(a), it reinstates:

   1. automatically as of the date the violation is cured, provided it is cured within 30 days of Your discovery of the violation; or

2. upon express reinstatement by the Licensor.

For the avoidance of doubt, this Section 6(b) does not affect any right the Licensor may have to seek remedies for Your violations of this Public License.

c. For the avoidance of doubt, the Licensor may also offer the Licensed Material under separate terms or conditions or stop distributing the Licensed Material at any time; however, doing so will not terminate this Public License.
d. Sections 1 , 5 , 6 , 7 , and 8 survive termination of this Public License.

Section 7 – Other Terms and Conditions.

a. The Licensor shall not be bound by any additional or different terms or conditions communicated by You unless expressly agreed.
b. Any arrangements, understandings, or agreements regarding the Licensed Material not stated herein are separate from and independent of the terms and conditions of this Public License.

Section 8 – Interpretation.

a. For the avoidance of doubt, this Public License does not, and shall not be interpreted to, reduce, limit, restrict, or impose conditions on any use of the Licensed Material that could lawfully be made without permission under this Public License.
b. To the extent possible, if any provision of this Public License is deemed unenforceable, it shall be automatically reformed to the minimum extent necessary to make it enforceable. If the provision cannot be reformed, it shall be severed from this Public License without affecting the enforceability of the remaining terms and conditions.
c. No term or condition of this Public License will be waived and no failure to comply consented to unless expressly agreed to by the Licensor.
d. Nothing in this Public License constitutes or may be interpreted as a limitation upon, or waiver of, any privileges and immunities that apply to the Licensor or You, including from the legal processes of any jurisdiction or authority.

About Creative Commons

Creative Commons is not a party to its public licenses. Notwithstanding, Creative Commons may elect to apply one of its public licenses to material it publishes and in those instances will be considered the "Licensor." The text of the Creative Commons public licenses is dedicated to the public domain under the CC0 Public Domain Dedication . Except for the limited purpose of indicating that material is shared under a Creative Commons public license or as otherwise permitted by the Creative Commons policies published at creativecommons.org/policies , Creative Commons does not authorize the use of the trademark "Creative Commons" or any other trademark or logo of Creative Commons without its prior written consent including, without limitation, in connection with any unauthorized modifications to any of its public licenses or any other arrangements, understandings, or agreements concerning use of licensed material. For the avoidance of doubt, this paragraph does not form part of the public licenses.

Creative Commons may be contacted at creativecommons.org .

Creative Commons is the nonprofit behind the open licenses and other legal tools that allow creators to share their work. Our legal tools are free to use.

- Learn more about our work
- **Learn more about CC Licensing**
- Support our work
- Use the license for your own material.

- Licenses List
- Public Domain List