Institute of Parallel and Distributed Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Masterarbeit

Enhancing Privacy in Car Data: Anonymization Techniques and Metrics Evaluation

Patrick Singer

Course of Study:	Informatik
Examiner:	Prof. Dr.-Ing Bernhard Mitschang
Supervisor:	Andrea Fieschi, M.Sc., Dr. rer. nat. Pascal Hirmer
Commenced:	May 15, 2023
Completed:	November 15, 2023

Abstract

With the advancement of connected vehicles in the automotive domain, data collection and sharing between vehicles or between vehicles and manufacturers reaches considerable proportions. The excellent possibilities of utilizing this data for product improvement are restrained by the fact that vehicle data contains personal information about drivers. Privacy-preserving technologies have been subject to research in various fields, but not as much in the automotive domain. Consequently, metrics evaluating the privacy and data quality provided by these technologies also remain scarce in this field. In this work, we apply different anonymization approaches to real-world vehicle data. We assess the performance of the approaches using a selection of metrics from the literature. Additionally, two domain-specific demonstrators are designed and implemented to analyze the privacy and data utility the approaches provide. The results show that privacy protection for vehicle data poses new challenges. We motivate the introduction of domain-specific metrics to evaluate the privacy and data quality of anonymization approaches in a useful way.

Contents

List of Figures

List of Tables

Acronyms

AOI Area of Interest. 51

DP Differential Privacy. 9, 16, 21

EC Equivalence Class. 18

EMD Earth Mover's Distance. 21

FL Federated Learning. 22

GDPR General Data Protection Regulation. 17

i.i.d. independent and identically distributed. 36

KL-divergence Kullback-Leibler divergence. 86

KPI Key Performance Indicator. 7, 9

ML Machine Learning. 22

PD Privacy Demonstrator. 9, 14, 41, 42, 43

PPDA Privacy Preserving Data Analysis. 17

PPDDP Privacy Preserving Dynamic Data Publishing. 27

PPDP Privacy Preserving Data Publishing. 17

UD Utility Demonstrator. 9, 14, 41, 42, 43

VGH Value Generalization Hierarchy. 7, 19

VIN Vehicle Identification Number. 7, 40

1 Introduction

Data generation, collection, and usage are omnipresent in every aspect of technology in the modern world. Businesses, organizations, and governments rely on data to improve products, make informed business decisions, and advance technologies. The automotive industry is not exempt from this development. Modern vehicles are getting smarter and more connected with every new model [CM16]. Consequently, the amount of data collected and shared by a vehicle grows considerably. The vast amounts of data enable car manufacturers to improve their vehicles, diagnose anomalies and technical faults faster, provide up-to-date navigation services, train Machine Learning models for autonomous driving, and much more [FHS+23].

These outstanding prospects are dampened by the fact that data collected in vehicles contains sensitive information about the drivers, like their position or driving behavior. Protecting the privacy of drivers is integral for car manufacturers to foster customer trust and adhere to legal regulations. However, privacy protection measures could not keep up with the rapid growth of data collection in the automotive field. A recent article published by the Mozilla Foundation finds that none of the 25 car brands they have researched adequately protects user privacy [JMZ23]. Therefore, measures of proper privacy protection in the automotive field are urgently needed.

Privacy protection is not a new topic, and data anonymization techniques have been around for over 20 years [Swe02b]. It is common knowledge in the field that the mere removal of identifying values, like names, addresses, or social security numbers, from the data does not suffice to protect the disclosure of sensitive information from the data to unauthorized third parties. In fact, Sweeney [Swe00] reports that about half of the U.S. population will likely be uniquely identified by a combination of their 5-digit ZIP code, gender, and date of birth. Literature in the field of anonymization is vast and spans privacy protection methods for text data, graph data, images, audio, and many more. Numerous anonymization approaches and privacy models exist. These approaches transform or distort the raw data to protect the privacy of the individuals, losing data quality in the process. Dwork [Dwo06] shows that "any statistical database with any non-trivial utility compromises a natural definition of privacy" [Dwo08]. Therefore, anonymization techniques have to balance a trade-off between the privacy protection of individuals in the data and the utility of the anonymized data for further processing. To this end, many different metrics have been proposed throughout the years to assess the utility and privacy guaranteed by anonymized data. Often, new metrics are introduced in correspondence with a novel anonymization approach, making them tailored to that specific approach and not applicable universally. Also, research in the field of anonymization for vehicle data remains sparse. These data show a variety of peculiarities, resulting in emerging challenges for anonymization.

This work aims to contribute to privacy preservation for vehicle data. The research question we address is the following: "What metrics exist that allow a comparison between different anonymization approaches on vehicle data?" To this end, we implement different anonymization approaches on real-world vehicle data collected by a large German car manufacturer. We identify

issues and challenges that arise when known approaches are applied to this data. We employ a selection of metrics from the literature to evaluate the utility and privacy the approaches provide. Additionally, a Utility Demonstrator and Privacy Demonstrator simulate a domain-specific data use case and adversarial attack on the anonymized data. We conduct experiments that evaluate the anonymization approaches using the metrics and demonstrators. The results of these experiments are discussed, and key takeaways are presented that suggest the usage of domain-specific measures to assess the quality of the anonymization of vehicle data.

This thesis is structured as follows. Chapter 2 outlines related work for this thesis. Chapter 3 introduces the basics and fundamental concepts of data anonymization. Chapter 4 presents the contribution of this work. Section 4.1 outlines peculiarities and challenges for data anonymization in the automotive domain. The selection of metrics for the experiments is presented in Section 4.2. The experiment design is illustrated in Section 4.3 and the implementation in Section 4.4. The results of the experiments are discussed in Section 4.5. Chapter 5 concludes this thesis and outlines future work. Additionally, Appendix A lists metrics encountered during the literature research but not selected for the experiments and an overview of existing software and tools for anonymization (Appendices A.1 and A.2, respectively).

2 Related Work

In this work, we evaluate anonymization techniques and metrics for privacy preservation in the automotive field. Related work is, therefore, comprised of (1) works in the field of privacy preservation for vehicle data, (2) works that list and evaluate metrics in the field of data anonymization, and (3) works that compare different data anonymization approaches. In the following, we present literature in the field for each of the abovementioned directions and motivate the scope of our work.

There is not much literature concerning privacy preservation in the automotive field. Coppola and Morisio [CM16] list upcoming issues with the emergence of connected vehicles. They stress the importance of privacy preservation but give no technical details or advice. Ghane et al. [GJK+21] address the topic of communication between multiple vehicles and potentially untrusted parties, where the communicated data contains sensitive information, like the velocity of the car or location data. The authors propose a framework that adds noise to the data on the vehicle level before communication. The work of Li et al. [LHSM23] addresses privacy protection for connected vehicles with a novel framework that allows drivers to set individual and situation-aware privacy demands. Drivers may wish to set different privacy regulations depending on the app they are using or their situation. Lastly, Fieschi et al. [FHS+23] explain different data use cases for product improvement in the automotive field. The authors characterize the different use cases according to their implications for privacy preservation. However, to the best of our knowledge, no work has been done on anonymizing real-world vehicle data containing information from multiple vehicles and utilizing metrics to evaluate the effectiveness of the anonymization approach.

Literature in the field of data anonymization presents a significant number of metrics that evaluate the utility and privacy of the anonymized data. Often, works present a novel approach to anonymization and define new metrics used to assess the effectiveness of the proposed approach. Studies that collect metrics and evaluate their expressiveness and applicability to different anonymization approaches are much scarcer. A quick literature research using the scientific platform Scopus [BV] yielded a mere 15 results when querying "anonymization AND metric AND survey" (effective 11 of November 2023). Of these 15, only three are relevant as related work. Kelly et al. [KRG+08] survey state-of-the-art metrics for network data. The authors compare ten different metrics regarding their applicability, complexity, and generality. However, the metrics are not tested in a practical experiment. Silva et al. [SMS21] provide an extensive survey of privacy concepts, privacy-enhancing algorithms and models, and privacy metrics in a cloud environment. Again, the metric properties are analyzed only theoretically. The same holds for the work of Majeed and Lee [ML21], who present anonymization approaches and metrics for tabular and graph data. An evaluation of these metrics concerning their expressiveness and performance on anonymized data remains an open topic.

Lastly, we turn our attention to works that evaluate different anonymization approaches. We again have the works of Silva et al. [SMS21] and Majeed and Lee [ML21] discussing different approaches only in theory. Ayala-Rivera et al. [AMCM14] present three anonymization approaches for k-anonymity and measure their efficiency concerning resource usage and effectiveness with

regard to data utility. The utility of the anonymized data is examined using three metrics. No metrics are considered that evaluate the privacy that the anonymized data guarantees. They conclude that multiple factors influence the best-performing approach, so there is no best-choice approach per se. Clifton and Tassa [CT13] discusses properties, advantages, and disadvantages of Differential Privacy and syntactic privacy theoretically. They conclude that both types of models have their place and that the two paradigms are not necessarily mutually exclusive. A paper by Soria-Comas et al. [SDSM14] shows that a synergy between k-anonymity and Differential Privacy can improve the data utility of the results of differentially private queries. This is because the noise that has to be applied to the result of a query in order to satisfy Differential Privacy can be significantly reduced if the query is executed on a k-anonymous dataset.

Some works in the literature aim to relate different privacy models directly through mathematical analysis. Ekenstedt et al. [EOL+22] show how Differential Privacy relates to t-closeness and vice-versa. They introduce an extension of the t-closeness model and a generalization of Differential Privacy to connect the two. Domingo-Ferrer and Soria-Comas [DS15] also relate t-closeness and Differential Privacy. They show that t-closeness implies Differential Privacy and that Differential Privacy implies a stochastic version of t-closeness. Other works that aim to compare different privacy models perform empirical evaluations of the anonymized datasets generated by the approaches. Sadhya and Chakraborty [SC22] perform experiments on real-world personal data using k-anonymity, l-diversity, and t-closeness as privacy models. The authors introduce a metric that consolidates utility and privacy into a normalized value to evaluate the results of the approaches. The experiments show that the t-closeness model performs well regarding the metric score. Cormode et al. [CPE+13] address the risk-utility trade-off between different privacy models with different parameters. The authors measure privacy empirically based on the posterior beliefs of an adversary. They evaluate data utility based on a workload of `COUNT` queries run on the anonymized data. This allowed the comparison of k-anonymity, l-diversity, t-closeness, and Differential Privacy with regard to the privacy/utility trade-off on the same scale. They conclude that the difference between Differential Privacy and syntactic models like k-anonymity is less dramatic than initially thought, although they find clear domination relations between them. The work by Almasi et al. [ASMH16] also aims to compare the privacy/utility trade-off of k-anonymity, l-diversity, t-closeness, and Differential Privacy on the same scale. They use multiple metrics to evaluate the trade-off, while [CPE+13] use only one evaluation metric for privacy and data utility, respectively. They observe that both t-closeness and Differential Privacy provide high utility while maintaining high privacy. In contrast to [CPE+13], they conclude that there are differences regarding the privacy guarantees the models make.

We conclude that a literature gap exists regarding the anonymization of datasets containing sensitive data in the automotive field. Furthermore, works that evaluate the applicability and expressiveness of metrics lack a practical realization of their analyses. Lastly, we observe that efforts regarding the comparison of different anonymization approaches lack a representative quantity of metrics for evaluation and are not present for vehicle data.

3 Fundamental Concepts

This chapter introduces fundamental concepts required for a discussion in the field of anonymization. Section 3.1 introduces important terminology required to discuss the existing techniques for privacy preservation presented in Section 3.2.

3.1 Terminology

Various terms that describe the privacy protection of individuals in data can be found in the literature. One of the most common ones throughout the years is "Statistical Disclosure Control". The term refers to the general goal of preventing third parties working with personal data from recognizing individuals in the data and learning private information [BL20; HDF+12; Tem17; WD01]. One aspect of Statistical Disclosure Control is exposing database access to third parties for data analysis, termed Privacy Preserving Data Analysis (PPDA) [DR13; Dwo06]. Another aspect is the process of publishing data while protecting the privacy of the respondents in the data, which is termed Privacy Preserving Data Publishing (PPDP) in the literature [CKLM09; ML21]. This is usually where the notion of "anonymization" is encountered.

The terms "anonymization" and "anonymized" appear in various fields of daily life. Whether it is software asking for permission to gather anonymized usage data, a survey collecting the results of participants anonymously, or anonymously browsing the internet using VPNs and onion routing. However, a technical definition of what it means to be anonymous is difficult to come by. One reason is that the definition of anonymity changes depending on the situation. Being anonymous in a network communication context means something different than being anonymous in a dataset [KRG+08]. This work addresses the anonymization of data. A common consensus is that data anonymization is irreversible, distinguishing it from approaches such as pseudonymization and encryption, where the raw data can be unveiled using external information (the mapping table of the pseudonyms and the encryption key, respectively). Pfitzmann and Köhntopp [PK01] define anonymity to be "the state of being not identifiable within a set of subjects, the anonymity set". A handout for the European General Data Protection Regulation (GDPR) defines anonymous data as "information which does not relate to an identified or identifiable natural person" [AEP23]. The concept of identifying an individual in a dataset is integral to many definitions of anonymization. These definitions extend to preserving privacy for companies, organizations, etc. For simplicity, we use the term "individual" in the remainder of this work for any entity that contributed to a dataset.

Information *identifies* an individual when it points to this and only this unique individual. Consider the raw medical data in Figure 3.1a. The name "Bob" uniquely identifies his entry in the table and allows anyone to conclude that Bob suffers from Apnea. Uniquely identifying an individual in a dataset is termed "identity disclosure" in the literature [LLV07]. Attributes in a dataset that clearly identify individuals are labeled *explicit identifiers*. Examples include social security numbers,

names, or addresses. *Quasi-identifying* attributes (or quasi-identifiers) can identify an individual when used in conjunction. These attributes appear benign in isolation but can single out individuals because their combination creates unique specifications. Considering Figure 3.1a once again, the attributes "Age" and "ZIP Code" (and their conjunction) can identify an individual in the data. Therefore, they are the quasi-identifiers of this dataset. An *Equivalence Class (EC)* in an anonymized table is a set of records with the same values for all quasi-identifiers, meaning they can not be distinguished using quasi-identifying values. Lastly, *sensitive attributes* contain non-public, private values. In the example table, "Condition" constitutes a sensitive attribute. The goal of anonymization for tabular data is to prevent the disclosure of the individuals' sensitive values.

3.2 Privacy-preserving Techniques

This section introduces prominent techniques for the privacy protection of tabular data. Some of the explained concepts are not direct approaches per se but privacy models that can be achieved using different methods.

3.2.1 k-anonymity

k-anonymity is a privacy model introduced by Sweeney [Swe02b]. The motivation behind k-anonymity is that even after removing explicit identifiers from a dataset, quasi-identifier values can still be combined with external data to identify an individual uniquely. To prevent this identity disclosure, k-anonymity demands that each record in a dataset is indistinguishable from at least $k-1$ other records with respect to the quasi-identifier values. In other words, each Equivalence Class in the dataset needs to be of size $\geq k$. Figure 3.1b shows a table satisfying k-anonymity for $k = 2$. The records in the ECs can not be distinguished using quasi-identifier values.

Although k-anonymity protects against identity disclosure, *attribute disclosure* is possible due to *homogeneity attacks* on the anonymized data. A homogeneity attack exploits the fact that all sensitive values in an EC are identical, which is the case for the second EC in the example table. These groups leak information because of the lack of diversity in the sensitive attribute. An attacker who knows David's age and ZIP code can identify the EC that must contain him. Although there is no way of telling exactly which entry is David's, the attacker can conclude that he was diagnosed with insomnia.

Name	Age	ZIP Code	Condition
Alice	25	10203	Gastritis
Bob	27	10204	Apnea
Charlotte	34	10701	Insomnia
David	45	10711	Insomnia

(a) The raw medical data.

Age	ZIP Code	Condition
[20, 29]	1020*	Gastritis
[20, 29]	1020*	Apnea
[30, 49]	107**	Insomnia
[30, 49]	107**	Insomnia

(b) The 2-anonymous release.

Figure 3.1: Two tables containing personalized medical information. The left table shows the raw data. The right table shows a 2-anonymous transformation of the raw data using generalization of attribute values.

Achieving k-anonymity

The main idea behind methods that achieve k-anonymity is transforming records such that groups of similar data points are created. Consequently, individual data points lose their uniqueness and identifiability in the data set, increasing the anonymity of respondents who contributed data. Another way of looking at these methods is as techniques that aim to bring together data points scattered in the quasi-identifier space [BF19]. However, this increase in privacy comes at the cost of a loss of precision, as data points have to be mutated to make them similar to each other.

Generalization One prominent way of achieving k-anonymity is through the generalization of quasi-identifier values. Generalization substitutes the precise attribute values with more general ones. For example, to generalize the dates of birth in a dataset, one could omit the day and month of the birthday and keep only the year, thus making the information less specific. Another example is a numerical value, such as a person's age, which can be generalized by bucketing it into an interval of values. An age of 22 could be generalized to the interval $[20, 25]$ or $[0, 50]$, for example. This process increases the ambiguity between records in the data set at the cost of data precision. One important property of generalization is that it is faithful, meaning any statement about the transformed data is also true about the raw data. The generalization of attribute values is done according to a predefined Value Generalization Hierarchy (VGH) [Sam01]. A VGH defines a generalization strategy for all values of an attribute. One example of such a hierarchy is shown in Figure 3.2 for the attribute "Marital Status". The topmost element in a VGH is the generalization of every element in the domain of the attribute and often denoted with $*$. There are two approaches to the generalization (or recoding) of attribute values [AMCM14]. *Global recoding* defines one generalized value for each original value in the attribute domain and applies this recoding to every record in the dataset. *Local recoding* generalizes the attribute values on a tuple by tuple basis. This means that the same original value can be recoded differently for two records.

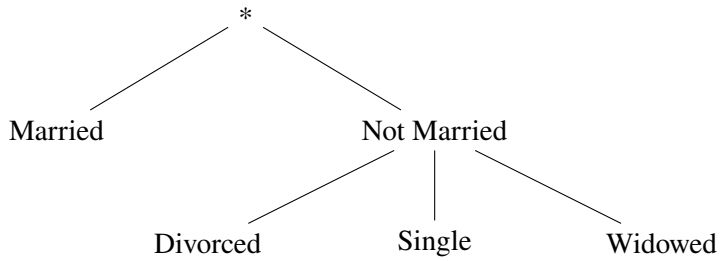


Figure 3.2: An example Value Generalization Hierarchy of the attribute "Marital Status".

Suppression k-anonymity can be achieved using solely generalization, but there exist cases where a few outlier tuples may require a high degree of generalization to achieve k-anonymity. In these cases, Samarati and Sweeney [SS98] suggest removing (or suppressing) these tuples to achieve k-anonymity with fewer sacrifices to data precision at the loss of completeness. Combining both provides a fitting means for using the minimal alteration of the data to achieve k-anonymity.

Clustering Clustering is introduced as a method to find suitable ECs for k-anonymity [BKBL07; LWFP06]. A clustering algorithm can group similar data points into clusters such that each cluster contains at least k elements. When the data points have been organized into clusters, their quasi-identifier values will be modified such that k-anonymity is achieved. To minimize information loss, the data points in one cluster should be as similar as possible.

Microaggregation The basic principle of microaggregation is the substitution of individual entries with entries computed on small aggregates (microaggregates) [DM02]. The approach is comprised of two steps [VFLS23]. First, records in the original data are partitioned into groups of at least k using a clustering approach. Then, an aggregation operator is applied to all records in one group. The operator summarizes the tuples in one cluster into an aggregated output, which then replaces the tuples in the respective cluster. The choice of the aggregation function depends on the data type of the values that are combined. Numerical values can be combined using the average or mean. For categorical values, one option is the group's most frequent value (mode). Another option is choosing the median of the values, which requires that an ordering is defined on the attribute. When microaggregation is performed on the quasi-identifiers, the resulting dataset is k-anonymous [SDSM14].

3.2.2 l-diversity

Machanavajjhala et al. [MGKV06] introduce l-diversity as an extension of k-anonymity that addresses the shortcomings of the privacy model regarding the possibility of attribute disclosure. l-diversity demands that the distribution of a sensitive attribute in each EC has at least l "well-represented" values. The paper gives different definitions that adhere to the demand "well-represented". A simple one is that at least l different values of the sensitive attribute must be present in each EC. The first EC in Figure 3.1b satisfies 2-diversity, but the overall table satisfies only 1-diversity because of the second EC.

l-diversity prevents an attacker from learning the precise sensitive attribute value of a specific record. However, there is still the possibility that an attacker can gain significant knowledge regarding the sensitive value for a tuple [BS08]. Suppose a table contains a tiny percentage of records that have the sensitive value S. Let an EC contain l different values of the sensitive attribute, where a significantly higher percentage has the sensitive value S. Any individual in this Equivalence Class will have a much higher probability of having value A than any other individual in the dataset, which is an information gain for an attacker.

3.2.3 t-closeness

Li et al. [LLV07] propose the t-closeness requirement as another improvement in data anonymization by grouping. This requirement aims to prevent the shortcoming of l-diversity that allows attackers to make probabilistic inferences on the sensitive attributes of tuples. The authors assume that there is always some base knowledge that an adversary has regarding the distribution of sensitive values in the table. As most generalization approaches do not alter the sensitive values in the table, the adversary will always know the probability distribution of sensitive values across the entire table, regardless of how strongly the quasi-identifier values have been changed in the anonymization

process. The rationale behind t-closeness is that the distribution of sensitive values in an EC should be "close" to this base knowledge of the adversary. If this is not the case, the adversary gains probabilistic information about the sensitive values of all individuals in this EC. Formally, an EC provides t-closeness if the difference between the distribution of sensitive values in this EC and the distribution across the entire table is at most t. A table provides t-closeness if every EC provides t-closeness. Li et al. [LLV07] use the Earth Mover's Distance (EMD) to calculate the difference between probability distributions. This measurement will be discussed in detail in Section 4.2.9.

3.2.4 Differential Privacy

Differential Privacy (DP) is a privacy model introduced by Dwork [Dwo06] that applies to algorithms on datasets rather than the datasets themselves. Roughly, an algorithm satisfies Differential Privacy if an observer seeing its output cannot tell whether a particular individual's information was used in the computation or not. The implication is that any harm done to an individual by an adversary observing the output of a function on the dataset is essentially the same whether the individual joins or refrains from joining the dataset. The formal definition of Differential Privacy operates on a stochastic function $\mathcal{K}$ on datasets, an $S \subseteq Range(\mathcal{K})$, and two neighboring datasets D_1 and D_2. Neighboring datasets are defined as two datasets that differ in the data of only one individual.

$$Pr[M(D_1) \in S] \leq e^{\epsilon} Pr[M(D_2) \in S]$$

The probability in the definition is taken over the randomness of the algorithm $\mathcal{K}$. The privacy parameter ϵ is also referred to as the *privacy budget* in the literature [ZLZY17]. The lower the budget, the stronger is the guarantee on privacy. One key feature of Differential Privacy (DP) is that it does not restrict the knowledge of a possible adversary. The mathematical property holds regardless of how strong the attacker is.

Achieving Differential Privacy

Dwork and Roth [DR13] give various approaches achieving DP on query functions. Randomized response is a mechanism that protects privacy by providing deniability of the result. For a given query that asks a yes or no question, flip a coin. If it lands on tails, respond truthfully. If it lands on heads, flip a second coin and respond with "Yes" if heads and "No" if tails. Wang et al. [WWH16] use the randomized response mechanism for differentially private data collection. Another established approach is the Laplace mechanism. This mechanism adds noise sampled from a Laplace distribution to the true value of the query result. The magnitude of the noise applied is controlled by the privacy parameter ϵ. As a final example, the exponential mechanism guarantees DP for queries where the output is associated with a "utility". It is used in settings where adding noise directly to the result would destroy its value, like auction prices. Therefore, the approach outputs query results with respect to a utility function that assigns each database/output pair a score [DR13].

Later works attempt to develop techniques that satisfy DP for Privacy Preserving Data Publishing [ZLZY17]. Some techniques perform multidimensional partitioning of the data to provide histograms describing the sensitive values [CPS+12; XXY10]. Bild et al. [BKP18] present a data

publishing algorithm that randomly draws samples from the raw data, which are then generalized to achieve k-anonymity. Bowen and Liu [BL20] present a study of data synthesis approaches that satisfy Differential Privacy.

3.2.5 Value Perturbation

Adding noise to perturb true results is applied at different stages of the data pipeline in the literature [CKLM09]. Evfimievski et al. [EGS03] apply perturbation for privacy-preserving mining of association rules. Blum et al. [BDMN05] propose a system that introduces noise to query responses in order to maintain privacy for the individuals in the dataset. The direct perturbation of attribute values is another approach to privacy protection for table-based data [HDF+12]. An anonymized attribute can be obtained by adding noise sampled from a normal distribution to the attribute values [DMB20].

3.2.6 Federated Learning

The idea of Federated Learning (FL) is that many devices contribute to the training of a central inference model. Each device performs local model training on its raw data and then transmits its trained model to a central server [BTM+20; KMA+21]. The server then aggregates the different models sent from the devices to obtain one model trained on all data held by the devices. This federated way of training a model improves privacy by default because the server never sees any raw data the devices hold. FL is different from traditional distributed learning, which is the parallelization of computing power to train a single model on multiple servers. In a Federated Learning environment, the training data on the devices may be imbalanced and of different sizes. In contrast, data are usually balanced and equally spread among compute nodes in a distributed learning scenario.

It should be noted, however, that FL on its own does not automatically guarantee that there is no information leakage. While it is true that the server never actually encounters any sensitive data, the trained model surprisingly can leak information. Shokri et al. [SSSS17] show that Machine Learning (ML) models are susceptible to membership attacks which, given a data record and access to a (black-box) ML model, were able to determine whether said data record was included in the model's training set. The authors furthermore note that Differential Privacy on the model training would, by construction, not be susceptible to the membership attack. This is because a specific output of any model training that uses a particular data point would be nearly equally probable as the same output without this data point in the training set. Because of this, recent work has integrated Differential Privacy in their FL approaches to provide an extra layer of privacy protection [BKMR21].

3.2.7 Data Synthesis

One approach to protecting the privacy of individuals in a dataset is to synthesize completely new data based on the original data [BL20; FWFY10]. For example, one could build a statistical model based on the original data and then sample synthetical data points from that model. The resulting synthetic dataset should show similar statistical properties to the original data. Similar to

the problem of ML models leaking information discussed in Section 3.2.6, it is essential to note that synthetic data is not automatically anonymous simply because it is not “real data”. Generative Adversarial Networks can be used for data generation where the generative network tries to fake data that shows the same characteristics as the original data [PMG+18].

4 Evaluation of Anonymization Techniques in the Automotive Domain

This chapter presents the main work of this thesis. We begin by discussing the emerging issues in data anonymization in the vehicle domain in Section 4.1. The selection of metrics used to evaluate the anonymization approaches is presented in Section 4.2. In Section 4.3, we introduce the data, the anonymization approaches, and the demonstrators for the experiments. Section 4.4 provides implementation details of the metrics and approaches. Finally, Section 4.5 presents the results of the experiments and a discussion of the findings.

4.1 Challenges and Peculiarities in the Automotive Field

The previous sections consider the problem of PPDP from a general viewpoint, agnostic of specific data domains. Since this thesis aims to contribute to the automotive field, we now focus on the characteristics and challenges of the automotive domain. In the following, we discuss why data is gathered in the automotive domain, what motivation there is to apply anonymization techniques to the collected data, what special characteristics data in the automotive field have, and finally, what challenges arise in the domain of anonymization for vehicle data.

Motivation for Data Collection Not only in the automotive domain but in general, data is mainly gathered to extract patterns and insights for product improvement. In the automotive domain, the collection and processing of vehicle information has been picking up speed with the emergence of connected vehicles. A connected car is capable of accessing the Internet, communicating with other smart devices, and leveraging vehicle-to-vehicle communication technologies [CM16]. Fieschi et al. [FHS+23] list some of the many data use cases done by a large German car manufacturer and classify them according to their requirements and challenges for anonymization. The use cases presented range from battery lifetime improvement over live traffic to Machine Learning for autonomous driving. These applications require vehicle data collected while the car is being driven. Because individual people drive cars, these data may contain sensitive information that the driver does not wish to be disclosed to unauthorized entities. While the data required for a battery lifetime improvement use case may not contain heavily personalized data, applications such as live traffic analysis already require accurate GPS locations of the drivers, a highly sensitive piece of information. Apart from information about the vehicle state, audiovisual data about the drivers themselves is also collected in the field of connected vehicles [LHSM23].

Motivation for Data Protection The motivation behind data protection in the automotive domain is twofold. Firstly, protecting the privacy of individuals in the data has marketing benefits. A company can use the confidence in its data protection capabilities to communicate this to potential customers openly. If the promise of excellent data protection stands the test of time, this company will profit from customer trust and a good image regarding the privacy protection of customer data. This goal is desirable for any company that collects and processes customer data. Secondly, the most forcing factor for data protection in any domain are regulations on the federal or supranational level. The GDPR is a data protection and privacy regulation enacted in the European Union. The regulation demands that "the principles of data protection should apply to any information concerning an identified or identifiable natural person" [EC16]. The regulation also notes that data that has undergone pseudonymization is not exempt from requiring data protection because it can be linked to a natural person with external information, i.e., the pseudonymization table. Anonymous information, on the other hand, is explicitly exempt from this regulation. The GDPR defines anonymous information as "information which does not relate to an identified or identifiable natural person" or as data where "the data subject is not or no longer identifiable". While this definition is intuitive, the weaknesses of some privacy models introduced in Section 3.2 show that seemingly anonymized data can still allow an adversary to identify an individual. The regulation lacks a technical definition of anonymity.

4.1.1 Data Characteristics and Emerging Challenges

Data in the automotive domain show several peculiarities compared to other fields. Firstly, there is a vast realm of data sources in the field. As mentioned, the data collected can take many forms, from GPS information over vehicle speed values and other sensory data to audiovisual data, including image, audio, and video material. The different data types require different approaches to anonymization to protect privacy. A consequence is that individuals are associated with many pieces of information in the raw data. This creates a detailed portrait of the individuals, which may contain considerable amounts of sensitive info. Different data sources can be joined to make statements like: Driver D was at position X at time T, going S kilometers per hour with a destination of Z while listening to artist A. The prevention of linking this information to an individual is ultimately the goal of anonymization of these data. Related to the issue of having different data types associated with an individual is the peculiarity of having multiple entries of *the same data type* associated with one individual. A vehicle picks up a variety of different sensor events throughout one trip, which are all associated with the same driver. As another example, an in-car passenger monitoring system for drowsiness detection collects multiple images or videos of the same driver. There exist works addressing the topic of privacy preservation for multiple records per person. Still, most of the literature assumes a one-to-one relationship between records and individuals [UKLU20].

Another issue in the field is that there are different stages in the data pipeline where data protection mechanisms may be employed. The vehicle data is generated in the car itself and may be transferred to the database of the car manufacturer, other IoT devices, or even surrounding cars and infrastructure. Data protection can be implemented at the vehicle level or when the data originating from multiple vehicles is consolidated in the database of the manufacturer. Federated Learning as introduced in section Section 3.2.6 is an interesting approach for data protection on the vehicle level because the notion of distributed clients contributing (sensitive) data to a central server is intrinsic to the modern automotive world. Each vehicle could train a statistical model on its data that is then sent to

and aggregated on a central server. The server will have a statistical model fitted on data that the server never had to see. Of course, this will only work for a use case that involves fitting a statistical model and is not a general approach to anonymization. On the other hand, perturbation-based anonymization approaches are a natural pick for privacy protection on the vehicle level. The car can apply noise to any data before sending it to a different entity. Of course, perturbation-based approaches (see Section 3.2.5) are also an option on the database level, along with grouping-based approaches. Data Synthesis (see Section 3.2.7) can also be performed on a server to generate anonymous vehicle data. However, the statistical model that is fitted and provides the samples to synthesize new data would have to allow the addition of new samples in the training process. A different point of application is the interaction of a data scientist with the database. The queries on the database available to the data scientist can be implemented in a privacy-preserving manner. This approach is referred to as Privacy-Preserving Data Analysis in the literature [ZLZY17].

Lastly, the dynamic nature of vehicle data is a significant factor in preserving privacy. Since vehicle data is collected continuously, the associated datasets will change continually. This means that new records will be added that require anonymization. The problem of sequentially publishing a dataset after it has been updated with insertions or deletions is named Privacy Preserving Dynamic Data Publishing (PPDDP) in the literature [ML21]. The difference between PPDDP and a one-time release of the dataset is that subsequent releases of data, although privacy-preserving *on their own*, may leak information when combined. As a straightforward example, imagine that a second data release contains only one more record than the first. An adversary that knows the first release can quickly determine which record has been added in the second release by comparing the two and can, therefore, learn sensitive information about the added individual. As a concrete example, consider a hospital that would like to release anonymous patient information every 6 months. The releases contain patient information spanning one year. Figure 4.1 shows two exemplary releases of data by the hospital. Alice and Bob form one EC in the first release, and Bob and Charlotte form one EC in the second release. Alice's stay at the hospital was already more than one year ago at the time of the second release, so she does not show up in the data. Imagine an attacker knows that Bob has to appear in both releases and knows Bob's Age and Zip Code. This attacker can easily identify Bob's EC in both tables and conclude from the first release that Bob has either Gastritis or Apnea and from the second release that he has either Apnea or Insomnia. Therefore, the attacker concludes that Bob suffers from Apnea, which would not have been possible by considering only one of the releases [XT07]. Literature in the field of PPDDP aims to provide models that guarantee privacy even for sequential data releases. Xiao and Tao [XT07] introduce the m-invariance model, which requires that the ECs containing a record in multiple releases must all contain the same sensitive values. The authors accomplish this by introducing counterfeit tuples into the data release.

Besides the discussed privacy issues of releasing continuous data, technical problems arise when handling sequential updates. Suppose a dataset of personalized data gathered from vehicles is transformed to satisfy k-anonymity. During the anonymization process, data quality is lost, and the raw data has been discarded due to data protection regulations. Integrating new records into the anonymized dataset poses some problems because it is not clear how they are incorporated with the others. One strategy is to add a new record to the EC that is "closest" to this entry. But after many iterations of extending the dataset with new data, we could have heavily skewed sizes of ECs. This problem would not emerge if the raw data could be considered as a whole again and anonymized to satisfy k-anonymity.

Name	Age	Zip	Condition
...	...		
Alice	[20, 30]	123*	Gastritis
Bob	[20, 30]	123*	Apnea
...	...		

(a) First 2-diverse, 2-anonymous release

Name	Age	Zip	Condition
...	...		
Bob	[20, 30]	123*	Apnea
Charlotte	[20, 30]	123*	Insomnia
...	...		

(b) Second 2-diverse, 2-anonymous release

Figure 4.1: Two releases of the same data that on their own satisfy 2-diversity and 2-anonymity but leak sensitive information when combined.

4.2 Selection of Metrics for Evaluation

This section presents the selection of metrics used in the experiments. Apart from choosing eleven metrics from existing literature, we define a handful of soft metrics introduced in Section 4.2.1. Table 4.1 gives an overview of the metrics presented in this section and their characteristics. The characteristics highlighted in the table are further explained in the following paragraphs.

The "Type" of a metric has one of four values. *Syntactic* metrics measure the structural properties of a dataset. These metrics do not consider the attribute values in the dataset at all, which is why they evaluate syntactic properties, not semantic ones. Metrics in this category can usually be calculated straightforwardly with low computational costs. *Distance-based* metrics measure a notion of distance between records or cell values. *Distribution-based* metrics are semantic measures. In contrast to syntactic metrics that focus on the structure of the anonymized table (for example, the sizes of ECs), these metrics consider the values and distributions of these values in the table. *Adversarial* metrics aim to model an adversarial attack by including hypothetical adversarial actions like the act of guessing the most probable value in an EC to be the sensitive value for an individual.

The property "Measure" indicates what aspect of the anonymized data is considered by the metric. Metrics that evaluate the performance of any given anonymization approach in the literature mainly examine data utility and privacy. Data utility relates to the ability to learn aggregate statistics about large groups of individuals. Privacy relates to the ability to extract information about specific individuals in the data [CPE+13; LL09].

The "Granularity" indicates how detailed the metric considers the individual entries in the table. Metrics that operate on a cell level consider the individual attribute values in the table, while record-based metrics operate on a tuple-by-tuple basis.

"Relative" indicates whether the metric outputs a score relative to a baseline dataset. Metrics that measure utility can incorporate the utility of the raw data in their calculation to capture information loss due to anonymization. Metris that measure privacy can incorporate the privacy of the trivially anonymized dataset, where all quasi-identifier values are removed [LL09]. Relative metrics can convey more information. For example, if the utility of the raw data is low, it will consequently also be low in the anonymized data. A relative metric, in contrast to an absolute one, can recognize that the utility of the anonymized data is not low because the anonymization approach destroyed it.

	Year	Type	Measure	Granularity	Relative	Normalized
Suppression Ratio	1998	Syntactic	P + U	Record	Y	Y
Minimum k	2002	Syntactic	P + U	Record	N	N
Minimum l	2006	Syntactic	P	Cell	N	N
Average Equivalence Class Size	2006	Syntactic	P + U	Record	N	N
Discernibility Penalty	2005	Syntactic	P + U	Record	N	N
In-Data Precision Loss	2002 (2006)	Distance	P + U	Cell	N	Y
Cross-Data Precision Loss	2015	Distance	P + U	Cell	Y	Y
Earth Mover's Distance	2007	Distribution	P	Cell	Y	Y
g-balance	2020	Distribution	P	Record	N	Y
h-affiliation	2020	Distribution	P	Record	N	Y
Adversarial Knowledge Gain	2008	Adversarial	P	Cell	Y	Y

Table 4.1: Overview of the selected metrics and their characteristics.

Lastly, "Normalized" indicates whether the metric values are adjusted to allow a comparison between different datasets. For example, removing five records from a dataset containing ten entries is a more significant loss of information than removing five records from a dataset containing 1000 entries. Normalization scales the raw value to allow a fair comparison.

Notation

This paragraph presents the mathematical notation used in the following chapters to ensure notational consistency in the remainder of this work. The notation is mostly consistent with the one that can be found in the literature, but some adjustments and harmonizations have been made.

#	Name	Age	Occupation
1	John Doe	26	Dentist
2	Alice Doe	26	Dentist
3	Eve Parker	31	Computer Scientist

Table 4.2: An example table with 3 attributes and 3 tuples

We define a table T with attribute set $A = \{A_1, \ldots, A_m\}$ to be a collection of records (or tuples) $t_1, \ldots, t_n$ and $|T| = n$ denoting the cardinality of T, meaning the number of tuples in T. The value of tuple $t \in T$ under attribute $A_i \in A$ is denoted by $t[A_i]$. Note that the term *value* in this context includes numerical values, categorical values, and even value intervals, which arise from the generalization of attributes. We use the term *interval* to describe a set of values such as $[2, 5]$ for example, and the term *range* to describe the difference between the largest and smallest value in the interval, which is 3 in this case. We define a quasi-identifier $QI \subseteq A$ to be a set of attributes $QI_1, \ldots, QI_k$ [Dal86; SS98]. The function $EC(T)$ returns all Equivalence Classes of table T as a set of sets of tuples. We also define $EC(t_i)$ to return the Equivalence Class that contains tuple t_i as a set of tuples.

Table 4.2 shows an example table $T = \{t_1, t_2, t_3\}$ with $|T| = 3$ numbered records and attribute set $A = \{Name, Age, Occupation\}$. Let $QI = \{Age, Occupation\}$ be the quasi-identifier. We have

$$t_1[Name] = John\ Doe$$
$$EC(T) = \{\{t_1, t_2\}, \{t_3\}\}$$
$$EC(t_1) = \{t_1, t_2\}$$

Throughout the explanation of the metrics, we set T to be the original table of microdata and T' to be the table that results from anonymizing T. We also refer to T' as the transformed table.

4.2.1 Soft Metrics

The first metrics we introduce are not mathematic measurements that output a specific number but high-level evaluations of the anonymization approaches. These soft metrics can be used as first considerations when deciding on a fitting anonymization approach.

Raw Data Permanence Some techniques for privacy preservation require permanent access to the raw, personalized data, which can pose problems regarding legal regulations. For example, techniques for Privacy Preserving Data Analysis tend to require permanent access to the raw data. As a reminder, these approaches often expose privacy-preserving interfaces to a database containing personal data. There exist approaches in the literature that allow a Machine Learning classifier training with DP [ACG+16]. If the data is intended for classification, a model can be trained on the raw data with guarantees of privacy preservation, and the raw sensitive data can be discarded afterward.

Required Amount of Data Preprocessing If an anonymization approach has strict requirements regarding the structure of data it expects, preprocessing the raw data may be a substantial task. The transformation of the raw data may introduce human error or result in information loss.

Data Use Case Flexibility Approaches for PPDP release an anonymized dataset to allow a wide range of data use cases. On the other hand, techniques for PPDA target very specific data use cases. This makes the usability of their individual outputs narrow but allows tailoring the mechanism exactly to a given use case. The selection of a fitting anonymization approach may be influenced by assessing whether the data will have broad or specific uses.

Intuitiveness of Privacy Effect Effective communication regarding the privacy measures taken to protect customer data greatly benefits companies. Customers who understand how their private data is managed build trust in the company. On the other hand, effective communication allows fast resolution of any inquiries coming from legal institutions. This metric captures how intuitive the privacy mechanism is to a non-specialist.

Required Amount of Domain Knowledge This soft metric assesses the difficulty level involved in implementing one of the anonymization approaches by someone unfamiliar with the data domain. A low requirement indicates that a data engineer can apply an anonymization approach to different types of data without the need to invest time to understand the semantics and structure of the data.

Suitability for on-line Data Collection Often, new data arrive continuously, and anonymization approaches should be able to handle the steady stream of new data. This metric assesses the ability of an anonymization approach to anonymize new data efficiently and effectively.

4.2.2 Suppression Ratio

One of the earliest metrics proposed is measuring the amount of suppressed tuples in the output of an anonymization algorithm [SS98]. This count can be normalized by dividing the result by the number of records in the input dataset.

$$M_{SR}(T, T') = \frac{|T| - |T'|}{|T|} = 1 - \frac{|T'|}{|T|}$$

Summary The Suppression Ratio is a syntactic, record-based measurement that considers both the original and the anonymized table to generate a normalized value in $[0, 1]$. Suppressing a record takes away utility from the data but also protects the privacy of the individual whose data is removed. Therefore, it is a metric that measures utility and privacy. Low values indicate that little utility is lost while little privacy is gained, and high values indicate high utility loss and high privacy gain. Although a basic metric, it can be found in more recent work [KYH+16].

4.2.3 Minimum k

The value k for which an anonymized table satisfies k-anonymity can be considered as a basic metric to evaluate the privacy as well as the utility of table T' [Swe02b].

$$M_k(T') = \min_{e \in EC(T')} |e|$$

Summary The minimum k in an anonymized table is a syntactic, record-based measurement that only operates on the anonymized table. The value lies in $[1, |T'|]$. Lower values indicate higher utility and lower privacy. Consider a dataset transformed to satisfy k-anonymity with a certain utility. This transformation automatically satisfies $(k-1)$-anonymity with the same utility. However, there could exist a different transformation of the original dataset satisfying $(k-1)$-anonymity with a higher utility. Thus, lower values of k are to be preferred regarding the utility of the dataset.

4.2.4 Minimum l

The l for which an anonymized table satisfies l-diversity can act as a measurement of privacy. Machanavajjhala et al. [MGKV06] present l-diversity as a privacy model that prevents homogeneity attacks. We use distinct l-diversity, which means that l for an EC is the number of distinct sensitive attributes in the group. Let $SV(e)$ return the set of unique sensitive values in EC e.

$$M_l(T') = \min_{e \in EC(T')} |SV(e)|$$

Summary The minimum l over all ECs in an anonymized table is a syntactic, cell-based privacy metric operating on the anonymized data. The value has a lower bound of 1 and an upper bound of the number of distinct sensitive values in the table.

4.2.5 Average Equivalence Class Size

Machanavajjhala et al. [MGKV06] propose the average EC size as a metric to get an intuition about how big the groups created by anonymization are.

$$M_{EC}(T') = \frac{|T'|}{|EC(T')|}$$

LeFevre et al. [LDR06] aim to add a normalizing factor to the metric calculation by dividing the value by k.

$$M_{NEC}(T') = \frac{M_{EC}(T')}{k} = \frac{|T'|}{|EC(T')| \cdot k}$$

This division does not normalize the values into the range $[0, 1]$ but signifies the factor by which the average EC is bigger than k.

Summary The Average Equivalence Classes Size is a syntactic, record-based metric only operating on the anonymized table. It outputs values in $[1, |T'|]$ for the originally proposed version and values between $[1, |T'|/k]$ for the version that incorporates the normalization factor. The authors introduce it as a cost metric on data utility, so smaller values signify higher utility. However, smaller values also indicate lower privacy.

4.2.6 Discernibility Penalty

Using the discernibility (distinctiveness) of tuples in the anonymized table as a metric of information loss is proposed by Bayardo and Agrawal [BA05]. The authors introduce the Discernibility Penalty, assigning a penalty to each tuple. This penalty is determined by the number of tuples in the transformed table that are identical to it. A suppressed tuple is assigned a penalty of $|T|$, the size of the input dataset. An unsuppressed tuple contributes a penalty that is as big as the size of its EC.

$$M_D(T, T') = \sum_{t \in T'} |EC(t)| + |T| \cdot (|T| - |T'|)$$

The first summand accounts for all unsuppressed tuples, and the second computes the penalty for all suppressed tuples. If the discernibility of tuples in a table is low, the Discernibility Penalty is high.

Summary The Discernibility Penalty is a syntactic, record-based metric that operates on the anonymized table. The lowest penalty possible for a table satisfying k-anonymity is k^2 times the number of ECs in the table. This is because each tuple contributes at least a penalty of k to the total value, and there need to be at least k tuples in any EC. The maximum value of the metric is $|T|^2$. A lower penalty means that fewer tuples are indistinguishable from each other, suggesting there is also less information loss and less privacy protection. The Discernibility Penalty is used in many works in the field [LDR06; NAC07; XWP+06].

4.2.7 In-Data Precision Loss

There are many uses of measuring in-data loss due to generalization in the literature, with one of the first ones being the "General Loss Metric" introduced by Iyengar [Iye02]. With the term "in-data", we mean that the metric operates only on the anonymized data, ignoring the original data. We consolidate the various approaches in the literature with this metric. A normalization term is introduced by Nergiz and Clifton [NC06], which we also adopt. This makes the metric expressive when comparing different anonymized datasets. Let $t' \in T'$ be the generalization of tuple $t \in T$.

$$M_{IPL}(T, T') = \frac{1}{|T| \cdot m} \sum_{t \in T} \sum_{i=1}^{m} IPL_{A_i}(t')$$

It is possible to assign different weights to the losses, depending on how important an attribute is deemed. One would then multiply $IPL_{A_i}(t')$ by w_i, where w_i is the weight that is assigned to the respective attribute A_i. It should be noted, however, that the normalization term needs to be adjusted should any weights be $\neq 1$. The precision loss is calculated over the attributes in the quasi-identifier, and $IPL_{A_i}(t')$ is assigned the maximum loss of 1 if t' is suppressed. A distinction is made between calculating the loss for numerical and categorical attributes.

Numerical Attributes Given a generalized tuple t', an attribute A_i, and the interval $t[A_i] = [y, z]$, the loss is

$$IPL_{A_i}(t') = \frac{z - y}{|A_i|},$$

where $|A_i| = max_{t \in T}\ t[A_i] - min_{t \in T}\ t[A_i]$ is the range of attribute A_i over all tuples in the original table.

Categorical Attributes There are different approaches to calculating the loss for categorical attributes. A straightforward approach is mapping the categorical attribute values to integers and proceeding with the formula for numerical attributes [AMCM14]. Another option is defining the distance between any two different categorical values to be equal to 1. However, both of these approaches disregard any semantic relations between categorical values. *Iyengar [Iye02]* propose a

mapping that captures more aspects of the semantics of the values. Given a taxonomy tree like the one in Figure 3.2 and $t'[A_i]$ taking the value of node P in the tree, the loss is

$$IPL_{A_i}(t') = \frac{M_P - 1}{M - 1} \in [0, 1] \qquad \text{[Iye02]}$$

Here, M is the total number of leaf nodes in the taxonomy tree, and M_P is the number of leaf nodes rooted at node P. This approach is similar to the one introduced by Loukides and Shao [LS07], who count the number of distinct values covered by $t'[A_i]$ and divide this value by the total number of distinct values in the domain of attribute A_i. The distinct values of an attribute are equivalent to the leaf nodes of the taxonomy tree. The only difference is that Iyengar [Iye02] subtract one from the numerator and denominator such that the value 0 indicating no loss at all is theoretically possible. Another variant is proposed by *Sweeney [Swe02a]* and also found in *Byun et al. [BKBL07]*. They calculate the loss of categorical values using the height of the taxonomy tree G.

$$IPL_{A_i}(t') = \frac{H(\hat{G})}{H(G)} \in (0, 1] \qquad \text{[BKBL07; Swe02a]}$$

$H(G)$ denotes the height of the full taxonomy tree, and $H(\hat{G})$ denotes the height of the subtree that is rooted at $t'[A_i]$. Considering the attribute taxonomy shown in Figure 3.2, the previous approaches assume that values are generalized to non-leaf nodes in the tree. For example, "Divorced" is generalized to "Not Married". *Xu et al. [XWP+06]* extend this taxonomy-based approach by allowing a generalization to an arbitrary set of the tree's leaf nodes. This would allow the generalization of "Divorced" to the set of values {"Divorced", "Single"}. They define the notion of a *closest common ancestor*. Let $V = \{v_1, \ldots, v_l\}$ be a set of nodes in the hierarchy. A node u is the closest common ancestor of V if u is an ancestor of every node in V and no descendant of u is also an ancestor of all nodes in V. In Figure 3.2, the closest common ancestor of *Divorced* and *Single* is *Not Married* and the closest common ancestor of *Married* and *Widowed* is *. A different loss $IPL_{A_i}(t')$ of tuple t' under attribute A_i can be specified using this definition.

$$IPL_{A_i}(t') = \frac{size(u)}{|A_i|} \in [0, 1] \qquad \text{[XWP+06]}$$

Here, $|A_i|$ is the number of distinct values of the attribute A_i, u is the closest common ancestor of $t'[A_i]$, and $size(u)$ is the number of leaf nodes that are descendants of u.

Summary The In-Data Loss is a distance-based, cell-based metric that measures the information loss resulting from generalization. It only applies to data that were transformed using generalization and suppression. The output value of the metric lies either in $[0, 1]$ or $(0, 1]$, depending on the chosen loss calculation. A low value means there is little data transformation, while a value of 1 indicates that the values are maximally imprecise. On the other hand, imprecise quasi-identifier values provide better privacy. The in-data precision loss is adopted frequently in the literature [AMCM14; NAC07; XWP+06], and Poulis et al. [PGL+15] use it as a metric to evaluate the anonymizations done by their tool SECRETA [USD23].

4.2.8 Cross-Data Precision Loss

In contrast to in-data, "cross-data" loss measures the decrease in precision compared to the original data. The value is computed by summing up the loss for every record $t \in T$ and its transformed counterpart $t' \in T'$ and then normalizing over the number of attributes and the table size.

$$M_{CPL}(T, T') = \frac{1}{|T| \cdot m} \sum_{t \in T} \sum_{i=1}^{m} CPL_{A_i}(t, t')$$

Again, the losses can be multiplied by weights to assign a stronger impact on important attributes. Similar to the in-data metrics, the exact way of calculating $CPL_{A_i}(t, t')$ varies in the literature. *Soria-Comas et al. [SDSM15]* use the squared distance between the transformed and the original value.

$$CPL_{A_i}(t, t') = (t[A_i] - t'[A_i])^2 \qquad \text{[SDSM15]}$$

Kikuchi et al. [KYH+16] measure the absolute distance between the original and transformed values.

$$CPL_{A_i}(t, t') = |t[A_i] - t'[A_i]| \qquad \text{[KYH+16]}$$

The difference is taken for sensitive attribute values in the paper, but the metric can also be evaluated on the quasi-identifier attributes. The loss that suppressed tuples contribute is not specified explicitly in either approach. It is also not apparent how the distance between categorical values is defined. Again, a basic mapping of categorical values to integers can solve the issue but loses the possible semantic meaning of the values. There is the possibility of using the definition of the closest common ancestor explained in Section 4.2.7. Let u be the closest common ancestor of $t[A_i]$ and $t'[A_i]$, then we can again apply the formula provided by *Xu et al. [XWP+06]*.

$$CPL_{A_i}(t, t') = \frac{size(u)}{|A_i|} \qquad \text{[XWP+06]}$$

It should be noted that this approach requires an attribute taxonomy. Also, note that t' does not necessarily need to be a tuple transformed by generalization in this case. For example, the values in t' may be generated using noise or microaggregation.

Summary The Cross-Data Precision Loss is a distance-based, cell-based metric that measures the information loss resulting from a dataset transformation. The output is normalized to lie between 0 and 1. A requirement for the metric is the existence of a mapping between the original and the transformed records.

4.2.9 Earth Mover's Distance (t-closeness)

The EMD is a measure used by Li et al. [LLV07] to calculate the value of t in the t-closeness privacy model. As a reminder, a table that provides t-closeness guarantees that the distance between the distribution of sensitive values in an EC and the distribution of sensitive values in the entire table is no more than t for all ECs in the table.

$$M_{EMD}(T') = \max_{e \in EC(T')} EMD(P_e, Q_{T'})$$

In the above equation, P_e is the distribution of sensitive values in EC e, and $Q_{T'}$ is the distribution of sensitive values over the entire table T'. Since sensitive attributes can be numerical (like a salary) or categorical (like a medical condition), there is one explicit formula to calculate the EMD for each case.

Numerical Attributes Given a sensitive numerical attribute with domain $\{v_1, v_2, \ldots, v_m\}$ where v_i is the ith smallest value, the Earth Mover's Distance between two distributions P and Q on this domain is

$$EMD(P,Q) = \frac{1}{m-1} \sum_{i=1}^{m} \left| \sum_{j=1}^{i} p_j - q_j \right|$$

Here, p_j is the probability of the jth element in the domain according to the distribution P, and q_j is the probability of the jth element according to Q. As an example, let the values in distribution P_1 be $\{3k, 4k, 5k\}$ and those in Q be $\{3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k\}$ and all values in one distribution independent and identically distributed (i.i.d.), we have that $C_{EMD}(P_1, Q) = 0.375$. On the other hand, if P_2 contains the values $\{6k, 8k, 11k\}$ (i.i.d.), then $C_{EMD}(P_2, Q) = 0.167$. The latter score is lower because EMD does not simply see the values as different points in the space, but assigns them a semantic meaning. Intuitively, P_2 is closer to Q because its values are more evenly distributed in the domain than P_1.

Categorical Attributes The authors propose two distance measures for categorical attributes. The *equal distance* measure is oblivious to any semantic relationships on attribute values and defines the distance between any two values to be 1. For two distributions P and Q of a sensitive attribute with m distinct values, we have

$$EMD(P,Q) = \frac{1}{2} \sum_{i=1}^{m} |p_i - q_i|$$

The *hierarchical distance* considers the attribute domain hierarchies by defining the distance of any two values in the hierarchy to be based on their closest common ancestor, a concept that was also employed in the in-data precision loss metric introduced in Section 4.2.7. As a short reminder, the closest common ancestor of two nodes is the lowest node in the hierarchy to which these two nodes generalize. Now, the distance between any two nodes v_1 and v_2 in the hierarchy is defined to be $level(cca(v_1, v_2))/H$, where $cca(\cdot, \cdot)$ returns the closest common ancestor of the two input nodes, $level(\cdot)$ returns the height of an input node, and H is the total height of the domain generalization hierarchy. The distance between two distributions P and Q is then

$$EMD(P,Q) = \sum_{N} \frac{height(N)}{H} \min(pos\_extra(N), neg\_extra(N))$$

where

$$pos\_extra(N) = \sum_{C \in child(N) \wedge extra(C) > 0} |extra(C)|$$

$$neg\_extra(N) = \sum_{C \in child(N) \wedge extra(C) < 0} |extra(C)|$$

$$extra(N) = \begin{cases} p_i - q_i & \text{N is leaf} \\ \sum_{C \in child(N)} extra(C) & \text{otherwise} \end{cases}$$

N denotes a non-leaf node in the hierarchy, and $child(N)$ returns the set of all leaf nodes for the subtree rooted at N.

Summary The Earth Mover's Distance is a distribution-based, cell-based metric that evaluates the difference between the distribution of sensitive values in an EC and the distribution of sensitive values across the entire table. The lowest value, 0, indicates that both distributions are equal. If the difference between the distributions is large, there is a large information gain regarding the sensitive value of a tuple in such an EC. Therefore, a metric that evaluates the privacy of any anonymization that groups tuples into Equivalence Classes. The authors mention that EMD is not a perfect distance measure. The EMD between the distributions $(0.01, 0.99)$ and $(0.11, 0.89)$ is $|0.01 - 0.11| + |(0.01 - 0.11) + (0.99 - 0.89)| = 0.1$, and the EMD between $(0.4, 0.6)$ and $(0.5, 0.5)$ is also $|0.4 - 0.5| + |(0.4 - 0.5) + (0.6 - 0.5)| = 0.1$. One could argue that the difference between the first pair is bigger because the probability of the first value increases more than tenfold, while it only increases by 25% in the second pair. Therefore, the authors note that the relationship between the value t and information gain is unclear.

4.2.10 g-balance

University of Illinois at Springfield, USA et al. [UKLU20] present a measure suited specifically for data with multiple records per individual. The g-balance is a generalization of k-anonymity and measures the balance of the individuals in one Equivalence Class EC. Let n_t be the number of individuals in the EC, which is not equal to the size of the EC if individuals have multiple records. Let c_i be the number of occurrences of the ith individual in the EC.

$$M_{GB}(T') = \min_{e \in EC(T')} GB(e)$$

where

$$GB(e) = 1 - \sum_{i=1}^{n_t} \left(\frac{c_i}{\sum_{j=1}^{n_t} c_j}\right)^2$$

Summary g-balance is a metric suited to evaluate the distribution of individuals in an EC if the data contains multiple records for each individual. Similar to k-anonymity, g acts as a lower bound for all ECs in the data. The value is 0 if all records in an EC belong to the same individual, i.e., $n_t = 1$. Higher values indicate a better distribution of individuals in the EC, reducing the possibility of identifying individuals in the group.

4.2.11 h-affiliation

The second metric presented by University of Illinois at Springfield, USA et al. [UKLU20] for data with multiple records for an individual is h-affiliation. Again, let EC be an Equivalence Class, n_t the number of individuals in EC, and n_j the number of individuals in EC associated with the jth sensitive value.

$$M_{HA}(T') = \max_{e \in EC(T')} HA(e)$$

where

$$HA(e) = \max_j \frac{n_j}{n_t}$$

Summary h-affiliation measures how well the sensitive values in an EC are distributed over the individuals in the EC. The maximum value 1 indicates that one specific sensitive value occurs for every group member, making a homogeneity attack possible. In general, higher values of h indicate a higher disclosure risk of the sensitive value. The minimum value of $1/n_t$ is reached if no individuals in the EC share a common sensitive value. The maximum value over all ECs in the table acts as an upper bound on the disclosure risk for the complete table.

4.2.12 Adversarial Knowledge Gain

Brickell and Shmatikov [BS08] introduce a metric that measures the information gain of an adversary due to the disclosure of the anonymized data. An adversary's base knowledge is the distribution of sensitive values in the original table, regardless of the degree of generalization applied. This is because even a trivial anonymization that would suppress the values of all attributes in the quasi-identifier would reveal the distribution of sensitive values in the table. The adversary's knowledge gain is the amount they learn from the generalized table compared to his base knowledge. Let EC be the set of all Equivalence Classes, then

$$M_{A\_gain} = \frac{1}{|T'|} \sum_{e \in EC(T')} |e| \cdot \mathcal{A}_{diff}(e)$$

So, the adversarial knowledge gain is the normalized knowledge difference of the adversary over all tuples in the table due to them knowing the Equivalence Classes of tuples. The authors give two ways to calculate the value of $\mathcal{A}_{diff}$, one measured additively, the other measured multiplicatively. For every EC e and sensitive attribute $S \in A$, the knowledge gain of the adversary is

$$\mathcal{A}_{diff\_add}(e) = \frac{1}{2} \sum_{v \in S} |p(T, v) - p(e, v)|$$

$$\mathcal{A}_{diff\_mult}(e) = \sum_{v \in S} \left| \log \frac{p(e, v)}{p(T, v)} \right|$$

Using the multiplicative calculation $\mathcal{A}_{diff\_mult}$ requires that there are no missing sensitive values in an equivalence class. Otherwise, the log and, therefore, the entire metric value becomes undefined. Alternatively, $\log 0 = 0$ can be fixed to calculate this metric.

Summary The Adversarial Knowledge Gain metric measures the amount of information an adversary gains by observing the anonymized dataset compared to the trivially anonymized dataset. This gain emerges through different distributions of sensitive values in the ECs compared to the overall table. For an EC with a distribution differing greatly, the adversary learns new information about the individuals in this Equivalence Class. The Adversarial Knowledge Gain is 0 if there is no difference in the distribution of sensitive values. Greater values indicate higher information gain. As with the Earth Mover's Distance, assigning an intuitive semantic meaning to the metric values is difficult. This measurement is also applied by Cormode et al. [CPE+13], where it is used to empirically measure privacy to compare different privacy models with each other.

4.3 Experiment Design

This section presents the setup for the experiments. We first introduce the vehicle dataset that is subject to anonymization. Then, we explain the Privacy and Utility Demonstrator we use to evaluate the metrics. Lastly, we present the anonymization approaches used in the experiments.

4.3.1 Data

In this work, we apply a selection of anonymization methods and evaluate their results on vehicle sensor data. The raw dataset consists of signals transferred by the car to the cloud. The data have been collected from 15 different vehicles. Each vehicle contributes a different number of trips to the data, such that the data is comprised of around 5,500 trips in total. The distribution of trips over the drivers is shown in Figure 4.2. Each trip contains many sensor events, totaling over 60 million individual records in the data. The data is given in a tabular form, where each record corresponds to one signal collected by a vehicle on a trip. Table 4.3 lists all relevant attributes in the dataset along with their descriptions.

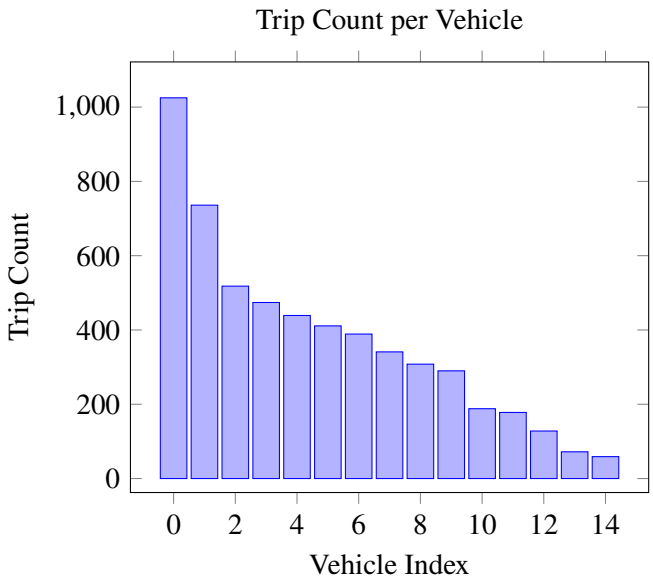


Figure 4.2: The number of trips per vehicle in the raw data.

Attribute	Description
VIN	A unique vehicle identifier
TripID	A unique trip identifier
Timestamp	Timestamp of the signal event
Event Name	Name of the signal event
Event Payload	Any additional data associated with the event

Table 4.3: The attributes of each record in the data.

The Vehicle Identification Number (VIN) is a standardized unique identifier for a vehicle made up of 17 digits and letters [37709]. Among other information, it encodes the vehicle's specifications, model, and manufacturer and includes a serial number allowing unique vehicle identification. A short explanation of an example VIN is displayed in Figure 4.3. The data contains hundreds of different events, covering a wide variety of incidents in vehicles. Some examples include the position information of the vehicle, speed information of the vehicle, the currently activated driving mode, the switching from radio to Bluetooth, and many more. The specific payloads for these events should be classified as sensitive information about an individual.

$$\underbrace{\text{WDCGG8HB0}}_{\text{Information on manufacturer and vehicle specification}} \qquad \underbrace{\text{AF462890}}_{\text{Serial number assigned by the manufacturer}}$$

Figure 4.3: Deconstruction of the Vehicle Identification Number WDCGG8HB0AF462890

Considerations for Privacy Preservation

The goal of anonymization for this dataset is to protect the privacy of the individuals who contributed their data. Although there is no direct information about any individual in the raw data, meaning no names, addresses, ages, or gender values, an individual can be identified using the VIN. A VIN technically identifies vehicles and not drivers, but it can identify individuals using external data. For example, combined with a car sales database of the manufacturer, a VIN can be linked to the individual who bought the vehicle. Therefore, the VIN is an identifying attribute, and data protection becomes necessary. We use the terms driver and individual interchangeably in the remainder of this work. Using the VIN directly is not the only way of identifying an individual in the data. If Eve knows that Bob drives to work regularly, she could search for the GPS information of Bob's trip to work in the data. If the search is successful, Eve also finds all of Bob's other trips because they are linked by the VIN. Another attack could consider the timestamps of the events. If Eve knows at which time Bob usually drives to work, she can look for trips in the data that start at the time Bob gets in his car and end when he arrives at work. Again, finding one (or more) of Bob's trips means that all other trips are found too. These kinds of attacks need to be studied and prevented but are out of scope for this work. We focus on reducing the risk of identification through direct usage of the VIN.

The raw data show some peculiarities of the automotive field discussed in Section 4.1. Table 4.4 presents exemplary data points illustrating the existence of multiple records per trip and individual. The methods and models introduced in the previous chapters assume that only one record in the dataset corresponds to any individual data giver. Another property of the data is that there is no single sensitive attribute but events carrying sensitive information in their payloads.

VIN	TripID	Timestamp	Event Name
XYZ	1	123	average energy consumption changed
XYZ	1	124	**speed changed**
XYZ	1	128	average energy consumption changed
XYZ	2	234	reverse gear active
XYZ	2	245	**location changed**

Table 4.4: Exemplary records for two trips of the same driver show the time series nature of the raw data. Events that can be classified as sensitive are displayed in bold font.

4.3.2 Privacy & Utility Demonstrators

We implement two demonstrators that aim to provide a realistic assessment of the selected anonymization approaches. The demonstrators are specifically designed with the automotive domain and the data at hand in mind. The Utility Demonstrator (UD) exemplifies a data use case for product improvement in the automotive domain. The Privacy Demonstrator (PD) simulates an adversary using the anonymized data to obtain sensitive information about the individuals in the dataset. We apply the demonstrators to the outputs of the anonymization approaches and investigate the results. We then assess how well the evaluations of the metrics on the anonymization approaches and the evaluations of the demonstrators on these approaches match.

Utility Demonstrator

The automotive industry shows a wide variety of domain-specific data use cases. For this demonstrator, we implement the calculation of KPIs. We choose a KPI that addresses the following question: "What is the percentage of trips where the touchscreen in the car was used?". Calculating the percentage of trips where a certain event occurred is a frequently used measure in the automotive domain. Among other uses, this kind of KPI can be used to evaluate design decisions in the car. For example, consider the removal of a hard key from one position in the cockpit to place it in a different position. Comparing the percentage of trips where the key was pressed before and after transferring it could indicate whether it is used less because its position is inconvenient. We calculate the percentage of trips over all vehicles and per engine type, so the question becomes: "What is the percentage of trips of vehicles with engine Y where the touchscreen in the car was used?".

Privacy Demonstrator

The GPS locations contained in the trip data may point to the drivers' homes, workplaces, or other places that should not be disclosed. This demonstrator simulates an adversary trying to extract sensitive position info from the anonymized data using a homogeneity attack as introduced in Section 3.2.1. As a short recap, the homogeneity attack on anonymized data is possible if a group of records in the anonymized data share the same sensitive information. If an attacker knows that a target individual is contained in such a group, they learn sensitive information about this individual without the need for exact identification of the individual in the anonymized data. In this demonstrator, we assume that the adversary knows all individuals who participate in the raw data. For each driver in the raw data, the adversary looks for fitting candidate drivers in the anonymized data. Then, the adversary intersects the GPS information contained in the trips of the candidate drivers. The result of this intersection is a set of GPS locations shared by each driver. These locations constitute the adversary's guess. We investigate how successful the adversary is in extracting GPS information with this strategy.

4.3.3 Selection of Approaches

We select 4 different approaches for anonymization, chosen to cover different strategies. The approaches are introduced in the following. The detailed implementation is explained in Section 4.4.

k-anonymity through Generalization Achieving k-anonymity through generalization is the only truthful approach that we use. Truthful means that any statement about the anonymized output is also true about the input data, so there is no perturbation or synthesis of values. This approach is often found in the literature and is a popular pick for the experiments in the papers. We perform generalization using local recoding and suppression.

k-anonymity through Microaggregation Another possibility to achieve k-anonymity is microaggregation. This approach clusters the raw data points together and reduces their uniqueness in the clusters. Different functions exist that achieve this reduction of uniqueness. We choose the mode as the aggregation function. This means that, for each attribute, the values in one cluster are replaced by the most common value in that cluster.

Data Synthesis Data Synthesis differs greatly from approaches that group records into ECs. We choose an approach proposed by Bowen and Liu [BL20] for synthesizing categorical data. The synthetic data is created by sampling from the cross-tabulation of the dataset. The cross-tabulation counts the number of occurrences in the data for each combination of quasi-identifier values. For our data, we create the cross-tabulation for the quasi-identifier and sample new values according to these counts. This data synthesis can satisfy DP if the counts in the cross-tabulation are perturbed with noise before sampling [BL20]. We implement the approach without DP and with DP.

Differential Privacy on Queries This last approach does not quite fit the topic of PPDP. Instead of creating a privacy-preserving dataset from the raw data, this approach allows arbitrary queries to be run on the raw data and reports the query results privately. We implement differentially private query calculation using the Laplace mechanism as explained in [DR13]. The true query result is calculated on the raw data and then randomized by adding noise sampled from a Laplace distribution. This approach is use case-bound because it requires that a query function on the data is defined beforehand. We apply this approach only to the Utility Demonstrator (UD) because its output is a single, randomized value, not a transformed table. This does not allow the evaluation using the selected metrics or the Privacy Demonstrator (PD).

4.4 Implementation

This section describes implementation details and decisions made during the implementation process. We first introduce the software and tools used in the experiments. Then, we explain the preparations and prerequisites for applying the anonymization approaches. We continue with the implementation details of the anonymization approaches and, lastly, describe the implementation of the Utility and Privacy Demonstrator.

4.4.1 Software & Tools

Due to the many existing anonymization scenarios, use cases, and techniques, different tools provide implementations of anonymization algorithms and metrics for evaluation. These implementations range from small codebases to fully-fledged anonymization suites that support many different models and metrics. In the following, we introduce the software we use for the experiments. The full list of software and tools that were considered is given in Appendix A.2. Our criteria for selecting software include the requirement for it to be free and open-source software.

ARX Anonymizer

ARX is an open-source tool for data anonymization [PES+20; PK15]. The software provides various privacy models, quality models, and methods for data anonymization. It support k-anonymity, l-diversity, t-closeness, DP, and many more. The data quality models are split into three categories: cell-oriented, attribute-oriented, and record-oriented. Cell-oriented models measure the precision loss of attribute values, attribute-oriented models measure the distance of attribute value distributions from the original to the anonymized dataset, and record-oriented models measure syntactic properties, such as the Discernibility (Section 4.2.6) or the Average Equivalence Class (EC) Size. We choose ARX because it implements all necessary privacy and transformation models for the experiments, is actively maintained, and recent research papers explain the software's implementation details [PES+20].

ARX also implements an algorithm for differentially private data publishing, which we did not see done by many tools that we considered. The algorithm is called SafePub [BKP18]. SafePub is a truthful algorithm, which means that the output dataset will contain transformed values that are consistent with the original ones. The algorithm does not add noise to the values to satisfy DP. Instead, SafePub first randomly draws records from the original dataset. Then, the uniqueness of these samples is reduced to attain a transformed dataset that protects privacy. We mention the implementation of this algorithm by ARX here because it was a candidate for the selection of anonymization approaches in this work. However, we could not produce meaningful outputs with the software because the execution did not terminate within a reasonable time. Therefore, we leave the investigation of this anonymization approach with ARX to future work.

ARX is implemented in Java and is available as Desktop Software for Windows, Mac, and Linux, as runnable JAR files, and as a Java library [PK]. Therefore, it can easily be integrated into existing software. ARX uses a dedicated anonymization algorithm named Flash [KPE+12], which outperforms well-established existing algorithms like Mondrian [LDR06] or Incognito [LDR05]. The ARX version used for this work is release 3.9.1.

Diffprivlib

Diffprivlib is a repository hosted on GitHub and maintained by IBM that implements different mechanisms, models, and tools for Differential Privacy in Python [HBML19; IBM23]. The library provides mechanisms that act as basic building blocks to achieve DP, such as the Laplace mechanism. But there are also implementations of ready-to-use, differentially private ML models like Linear and Logistic Regression and Naive Bayes. The library is actively maintained, provides good documentation, and can easily be installed using the Python package manager `pip`. We use version 0.6 for the experiments.

4.4.2 Prerequisites for Anonymization

To apply the anonymization approaches and metrics in the experiments, some prerequisites and data preparations are required. k-anonymity requires attributes that make up a quasi-identifier in the data. l-diversity and many metrics, especially ones that measure privacy, like EMD, require that a

sensitive attribute is specified. Furthermore, we implement the differentially private data synthesis and the differentially private query calculation using the Laplace mechanism, which will also be discussed in the following.

Choosing the Quasi-Identifier and Sensitive Attribute

It is customary to remove direct identifiers from the data before proceeding with anonymization [SS98; Swe02b]. However, removing the entire VIN is a huge loss of information. Because of this, we choose to replace the VIN with the vehicle information that it encodes. Of the many encoded vehicle properties, we select three as an illustration. The engine has three possible values: EV, Gas, and Hybrid. EV indicates an electric vehicle. The body of a vehicle also has three possible values: Sedan, Long Sedan, and SUV. Lastly, the number of seats has four different values: 2, 4, 5, and 7. We set the quasi-identifier for the data to be $\{Engine, Body, Seats\}$. Table 4.5 shows the combinations of the quasi-identifier values occurring in the data and the number of vehicles and trips that match the specifications. Some combinations can identify individuals in the data because they occur only for one driver. Because the VIN is replaced by the quasi-identifier, we lose the

Engine	Body	Seats	Vehicle Count	Trip Count
EV	Sedan	2	4	2,028
EV	SUV	4	1	59
EV	SUV	5	2	828
EV	SUV	7	1	128
Gas	Sedan	2	2	826
Gas	Long Sedan	2	2	701
Gas	Long Sedan	4	1	736
Gas	Long Sedan	5	1	178
Hybrid	Long Sedan	5	1	72
			15	5,556

Table 4.5: Combinations of quasi-identifier values in the data and the number of vehicles and trips with these specifications

possibility of recognizing trips belonging to the same driver. In the original data, trips belong to exactly one driver, indicated by the VIN. Now, the trips of many drivers are combined under the same quasi-identifier values. Consider, for example, the quasi-identifier values $\{EV, Sedan, 2\}$, which combine the trips of four different drivers. The inability to link different trips to the same driver is a loss of information that is undesirable for some use cases. A data analyst might want to examine the usage of a certain functionality on a driver basis, not a trip basis. Therefore, we introduce a DriverID that indicates the trips belonging to the same driver. For simplicity in the implementation, this DriverID is an integer from 1 to 15, one for each driver. Table 4.6 shows an exemplary data snippet that includes the DriverID. We will see in the results of the experiments that this DriverID has some implications regarding the privacy of the drivers.

Lastly, a sensitive attribute is required by some of the selected metrics. We select one signal event as a representative and aggregate the event data such that each trip is assigned a sensitive value. The GPS information contained in the trips is very delicate information, but there is no apparent

way to combine the many positions in a trip into one value for each trip. Therefore, we choose the most played music artist during the trip as the sensitive attribute. Although not as delicate as GPS positions, it is arguably a piece of information that could reveal personal background or political orientation. The data contain sensor events that indicate a change in the current song for driver-controlled media input like Bluetooth or AUX input. We extract the artist from the payload of each event and set the favorite artist to be the one that occurs the most in each trip. There are several trips where no song change event is registered because these events are not fired when the current song played by a radio station changes. We assign these trips the sensitive value "Radio". As a result, the raw data contain more than 600 different artists in total. Table 4.6 shows how the sensitive attribute "Artist" is represented in the data.

Engine	Body	Seats	DriverID	TripID	Artist
EV	Sedan	2	1	M	Kid Cudi
EV	Sedan	2	1	N	Radio
EV	Sedan	2	2	O	Radio
EV	Sedan	2	3	P	Paul Kalkbrenner
EV	Sedan	2	4	Q	BLACKPINK
EV	SUV	4	5	R	Taylor Swift
EV	SUV	4	5	S	Radio
EV	SUV	4	6	T	Taylor Swift

Table 4.6: An exemplary snippet of trips belonging to different drivers. The DriverID allows associating different trips with the same driver. Each trip has a sensitive value associated with it.

The Laplace mechanism

Both Data Synthesis and differentially private queries use Laplace noise to satisfy Differential Privacy. We use this paragraph to give an introduction to the concept and the implementation that we use. The Laplace mechanism adds noise to the true result of a function $f : X \rightarrow \mathbb{R}^k$ to achieve DP [DMNS06; DR13]. The noise is sampled from a Laplace distribution, which has two parameters. μ is a location parameter and b is a scale parameter. The probability density functions of some Laplace distributions for different parameters are shown in Figure 4.4. The noise applied to the true result is unbiased, so $\mu = 0$. The scale parameter b depends on the privacy parameter ϵ and the l_1-sensitivity Δf of the function f. In simple terms, the l_1-sensitivity Δf of a function f is the maximum amount of change in the output that the presence of an individual entry in the input can have. More precisely,

$$\Delta f = \max_{D,\, D'} ||f(D) - f(D')||_1$$

for two databases D and D' where $|D| - |D'| = 1$, meaning they differ in one element [DR13]. The second parameter is now fixed to $b = \Delta f / \epsilon$. For function f and input x, the Laplace mechanism calculates $f(x) + (Y_1, \ldots, Y_k)$, where all Y_i are drawn i.i.d. from Laplace$(0, \Delta f / \epsilon)$. As shown in Figure 4.4, larger values of the parameter b tend to flatten out the probability density function, hence increasing the standard deviation of the added noise. Increasing ϵ (softening the privacy

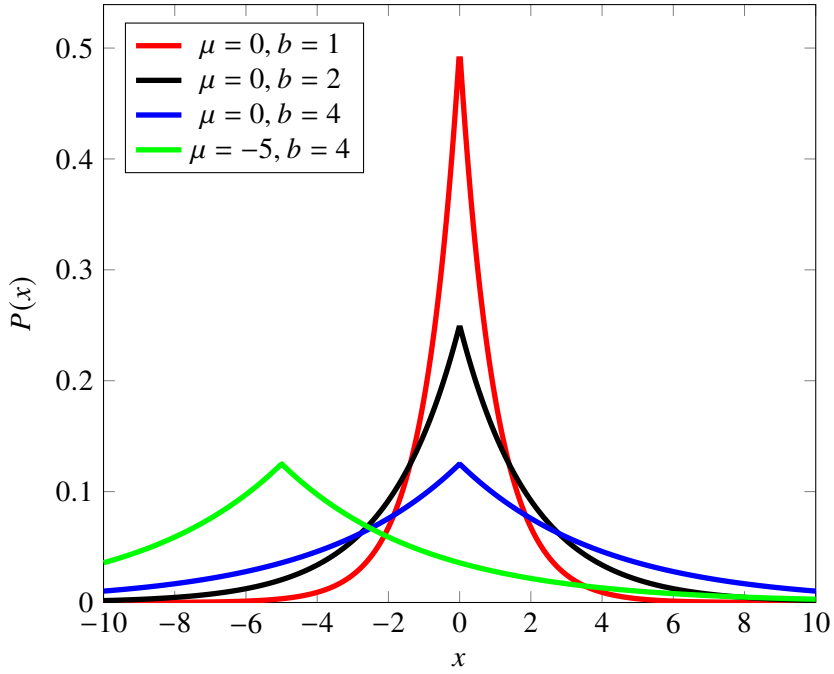


Figure 4.4: Probability density functions for Laplace distributions of different values for parameters μ and b.

claim) decreases the magnitude of the added noise. On the other hand, increased sensitivity of the function f results in a higher magnitude of noise values. In practice, the only adjustable parameter is ϵ, as the function f fixes the sensitivity. Diffprivlib provides the implementation that we use for randomizing the true results of function f with the Laplace mechanism.

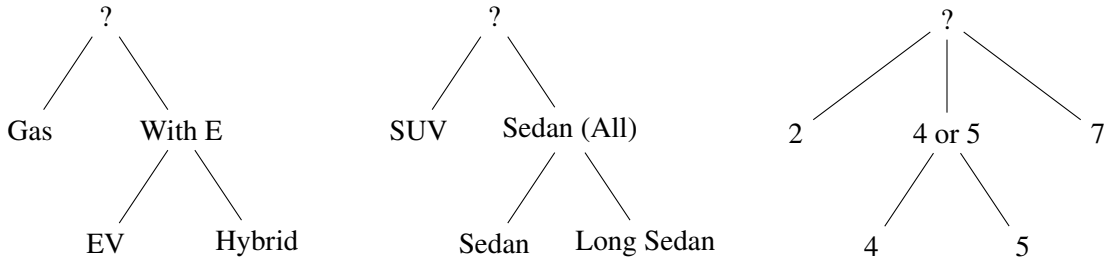
4.4.3 Implementation of Approaches

With the prerequisites settled, the following sections describe the implementation details of the anonymization approaches.

k-anonymity through Generalization

We perform generalization using the Desktop version of the ARX anonymizer. There are two variants of applying k-anonymity to our dataset. The parameter k can refer to the trips or the drivers in the data. Since k-anonymity is designed to guarantee that the information for one individual can not be distinguished from the information of at least $k - 1$ other individuals in the data, we apply k-anonymity on the drivers. Therefore, the input dataset only contains the 15 drivers with their respective quasi-identifier values. The anonymization procedure then generalizes the attribute values or suppresses entire drivers. To receive the full anonymized dataset, we remove the trips of all drivers that have been suppressed and replace the quasi-identifier values of the other drivers with their generalizations. The anonymized dataset will also satisfy k-anonymity on the drivers, as any drivers in the same EC in the output will also be in the same EC in the anonymized dataset.

In the GUI of ARX, we set the types of the "Engine", "Body", and "Seats" attributes to be "Quasi-identifying" and define their Value Generalization Hierarchies in accordance with the ones shown in Figure 4.5. The transformation operation for all quasi-identifier attributes is set to "Generalization". We set k-anonymity as the privacy model and perform anonymization with $k = 2, 5, 10$. The suppression limit is set to 100%, which would theoretically allow the tool to suppress every tuple. In practice, however, we find that ARX only suppresses outlier values that would have forced a heavy generalization on all other records. For each anonymization, we set the search strategy to "Optimal" and set the transformation model to be local transformation with 100 iterations. We use the preset configuration for all other options. The output table is exported to a CSV file for further processing.



(a) Value Generalization Hierarchy of the attribute "Engine".

(b) Value Generalization Hierarchy of the attribute "Body".

(c) Value Generalization Hierarchy of the attribute "Seats".

Figure 4.5: The Value Generalization Hierarchies for the attributes in the quasi-identifier.

k-anonymity through Microaggregation

For this approach, we also define k-anonymity on the drivers for the same reasoning as in the generalization approach. Therefore, the input dataset again only contains the 15 drivers and their quasi-identifier values. We use the same procedure and set the same configurations as for generalization in the ARX GUI, with the following exceptions. The transformation type for the attributes in the quasi-identifier is set to "Clustering and microaggregation", and the microaggregation function is set to "Mode". To receive the full anonymized dataset, we again discard the trips of every driver that has been suppressed by the anonymization procedure and replace the quasi-identifier values of the remaining drivers with those generated by the procedure. Again, k-anonymity on the output of the procedure implies k-anonymity on the full anonymized dataset.

Data Synthesis

This approach operates by replacing the quasi-identifier values of the records by sampling new values. The sampling is done according to the counts in a contingency table over the quasi-identifier. Table 4.5 is a contingency table that is missing all other possible combinations of quasi-identifier values, which have a driver and trip count of 0. Now, it is possible to replace the quasi-identifier values on a driver or a trip basis. Sampling on a driver basis means that new values are sampled according to the driver counts and assigned to each driver, such that all trips of one driver receive the same new quasi-identifier values. Sampling on a trip basis, on the other hand, means that each

trip receives new quasi-identifier values according to the trip counts in the contingency table. This destroys the connection between trips belonging to the same driver. We choose to sample on a driver basis to keep trips done by the same driver still connected under the same new quasi-identifier values after sampling. The anonymization proceeds by sampling 15 quasi-identifier combinations with replacement according to the trip counts in the contingency table. To receive the anonymized dataset, the true quasi-identifier values of the drivers are replaced by the sampled values.

The data synthesis can satisfy DP if Laplace noise is applied to the counts in the contingency table [BL20]. Every count n in the table is then replaced by $n + e$, where $e \sim \text{Laplace}(0, \Delta_f/\epsilon)$. If the creation of the contingency table satisfies DP, then any sampling from this contingency table will also satisfy DP for the same ϵ due to the post-processing property of DP [DR13]. Δ_f denotes the sensitivity of the count function, which is the maximum amount of change in the counts due to a driver joining or leaving the dataset. If one driver is removed or added to the data, the trip counts change by the number of trips done by this driver. Therefore, the sensitivity of the count function is the maximum number of trips over all drivers in the data. The added Laplace noise can be significant depending on ϵ, which could lead to undesirable negative perturbed counts. Therefore, we use the truncated Laplace mechanism provided by Diffprivlib and set a lower bound of 0 for the noisy values. Table 4.7 shows the contingency table for the data and exemplary noisy counts. The table shows that the original counts are completely distorted for small ϵ.

Sampling from these perturbed counts may lead to quasi-identifier combinations that are impossible in practice. For example, an electric SUV with two seats could be sampled for $\epsilon = 1$, although no car with these specifications is built. In our implementation, we avoid the additional overhead of filtering out infeasible quasi-identifier values because the experiments do not profit from this measure.

Engine	Body	Seats	Vehicle Count	Noisy Vehicle Count		
				$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 0.1$
EV	SUV	2	0	0	0.25	0
EV	SUV	4	1	1.01	1.59	0
EV	SUV	5	2	1.73	4.63	16.85
EV	SUV	7	1	0.93	1.88	0
EV	Sedan	2	4	4.05	4.48	31.39
EV	Sedan	4	0	0	1.09	0.72
EV	Sedan	5	0	0.03	0.2	0
EV	Sedan	7	0	0.03	0	0
...						

Table 4.7: An excerpt of the contingency table that is used as a base for synthesizing new quasi-identifier values. The rightmost columns show exemplary noisy counts generated by the truncated Laplace mechanism for different ϵ. Values are rounded to two decimal places.

Differential Privacy on Queries

This approach guarantees DP on the calculation of a query function f by using the Laplace mechanism to perturb the true result of the query on the data. The sensitivity of the function f is the maximum change on the query result that the absence or addition of one individual can have. This

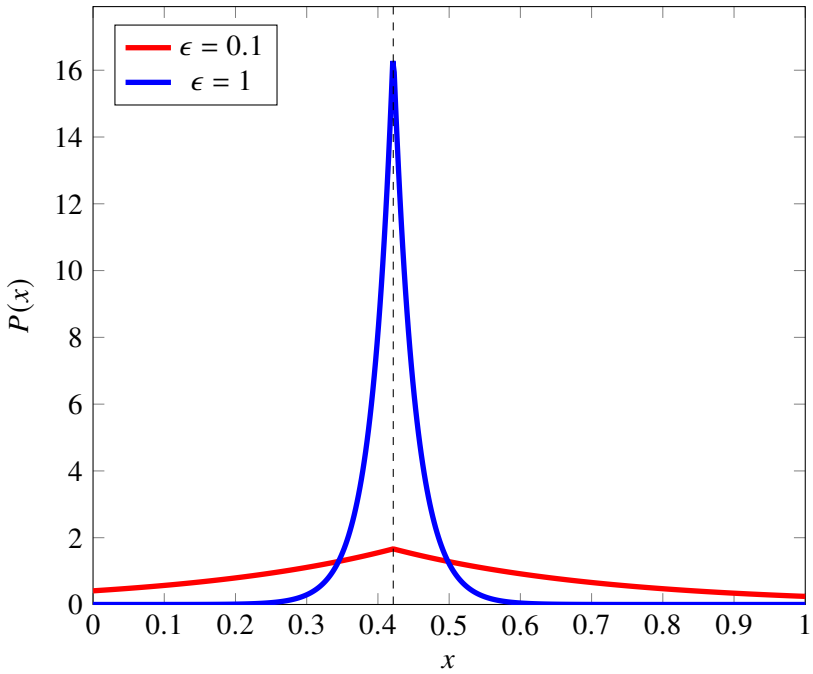


Figure 4.6: Probability density function of the Laplace mechanism on the KPI over all drivers for $\epsilon = 0.1$ and $\epsilon = 1$. The dashed black line indicates the true value.

is dependent on the query function f. Since we apply the procedure on the Utility Demonstrator, the function f is the KPI calculation. The exact query function and, therefore, the sensitivity will differ depending on whether the KPI is calculated over all vehicles or per engine type. We compute the sensitivities of a KPI by first calculating the true result over the entire dataset. Then, for every driver, we calculate the KPI again, omitting all trips of this driver from the calculation. The sensitivity of the KPI is then the maximum difference between the true result and the result with a driver omitted, over all drivers. Figure 4.6 shows the distribution of noisy values generated by the Laplace mechanism on the KPI over all drivers for two values of ϵ. The sensitivity of the query lies at around 0.031, so the scale parameter of the Laplace distributions is $b \approx 0.031/\epsilon$.

The output values of the mechanism are almost guaranteed to lie between 0 and 1 for $\epsilon = 1$, but this is not generally guaranteed. A smaller ϵ flattens out the distribution, and a different true value shifts the distribution to the left or right. The Laplace mechanism may then output values < 0 or > 1, which are invalid results for a function that measures the ratio of drivers. Therefore, we again use the truncated Laplace mechanism provided by Diffprivlib, setting the lower bound to 0 and the upper bound to 1.

4.4.4 Implementation of the Demonstrators

Utility Demonstrator

To calculate the KPI, we divide the number of trips where the event occurred by the total number of trips in the data. For a specific engine, we divide the number of trips done by cars with this engine type where the event occurred by the total number of trips done by vehicles with this engine type.

In the experiments, we report the actual result of the KPI calculations on the anonymized data and the relative error compared to the true result. Let y be the true value calculated on the raw data and y' the value calculated on the anonymized data. We report y' and the relative error $|y - y'|/y$.

Privacy Demonstrator

The PD models an adversary that tries to find Areas of Interest (AOIs) of the individuals participating in the dataset. We now explain the concrete implementation of the attack with Figure 4.7 as a running example. First, the true AOIs for each individual in the raw data are extracted. Some events in the data log the driver's position, which is given as latitude and longitude values in degrees. We crop the values after the fourth decimal place, so 12.123456 becomes 12.1234. This essentially creates GPS areas that span a 10,000th degree of latitude and longitude, such that GPS positions lying in one area are all mapped to the "root" point of that area. Cropping the values after the fourth decimal results in areas that are about as big as a standard house in the part of the globe where the data is from. We then collect the areas in one trip into a set without duplicates. In the example, the areas of the first trip of Driver 1 are $\{(0, 1), (1, 1), (2, 1)\}$. For each driver, we unify the sets of areas for all trips to receive a set of AOIs for this driver. In the example, the AOIs for Driver 1 are $\{(0, 1), (1, 0), (1, 1), (2, 0), (2, 1)\}$.

The adversary now proceeds as follows. For each driver D in the input dataset (which we assume they know), they search for candidates in the anonymized dataset according to the quasi-identifier values. In the running example, assume that Driver 1 and Driver 2 are grouped into the same EC in the anonymized data. This would mean that the adversary's candidate drivers for Driver 1 are Driver 1 and Driver 2 in the anonymized dataset. The adversary might not find any candidates for D if the trips of this driver were suppressed during anonymization or if the exact quasi-identifier values are not present in the synthetic dataset. Now, the adversary extracts the AOIs associated with each candidate driver using the approach explained above. The intersection of the AOIs for each candidate driver is then the adversary's guess. In the example, areas $(1, 0)$ and $(2, 0)$ make up the adversary's guesses because both drivers have trips going through these areas. The adversary's recall for driver D is the ratio of true AOIs of D that are part of the adversary's guess. In the example, the recall for Driver 1 is $2/3$ as 2 out of 3 AOIs of the driver are contained in the adversary's guess. The recall for Driver 2 is $2/5$. If the anonymization groups all drivers in the example into one EC, then there are no possible guesses for the adversary because no area is visited by all drivers. The adversary's recall in this case is 0%. In the experiments, we report the maximum and average recall over all drivers as well as the ratio of compromised drivers, which is the fraction of drivers for whom the adversary guesses at least one correct AOI. The fraction of compromised drivers in the example is $2/3$. Unfortunately, two drivers do not have position information associated with them on any of their trips, so we consider the remaining 13 drivers only for this Demonstrator.

It is worth noting that this kind of attack depends on the availability of a DriverID, allowing the adversary to group trips coming from the same driver. If the DriverID is omitted, n trips in an EC could come from only 1 driver or n drivers. The adversary's best guess is then the areas that appear in every trip, but intersecting all trips will return no common areas in practice with areas that have the size of a standard house. For our data, intersecting all trips of just one driver yields no common areas, which makes the attack infeasible without a DriverID.

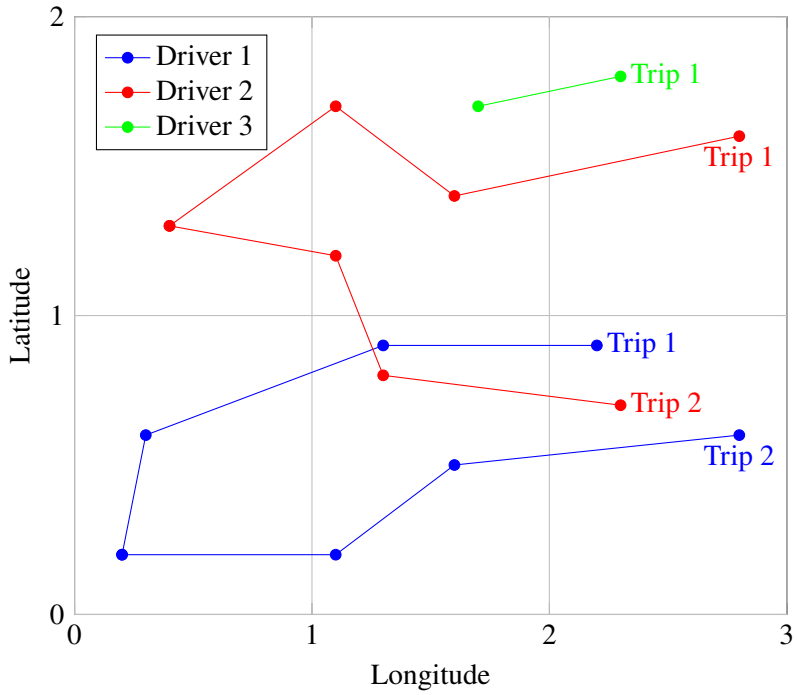


Figure 4.7: Example trips of two different drivers. Each trip has some latitude and longitude values associated with it. The grid drawn in grey represents the GPS areas.

4.4.5 Implementation of the Metrics

In this section, we explain the implementation of the metrics on the anonymized datasets. Most of the metrics introduced in Section 4.2 are defined for datasets containing only one record per individual. Because our data contains many records per individual, we specify how the metrics are applied in our case. Some metrics do not output a normalized value, which reduces their expressiveness. We scale the outputs of these metrics using a min-max scaler in the presentation of the experiment results. The min-max scaler calculates a normalized value according to the following formula.

$$\frac{m - m_{min}}{m_{max} - m_{min}}$$

Here, m denotes the actual value computed by the metric, m_{min} denotes the lower bound, and m_{max} denotes the upper bound for the metric value. The values used as m_{min} and m_{max} for the respective metrics are specified below. The effect of min-max-scaling the metric values is converting the absolute measurement to a relative one. It should be noted that this assigns the metrics an expressiveness they did not have before. We use a running example, shown in Table 4.8, to illustrate an anonymized dataset and give example results for the metrics applied on this dataset.

Suppression Ratio This metric counts the number of suppressed records in the dataset and divides the value by the number of records in the input dataset. The Suppression Ratio for the running example is $2/10$.

Engine	Body	Seats	DriverID	TripID	Artist
?	Sedan	2	1	M	Kid Cudi
?	Sedan	2	1	N	Radio
?	Sedan	2	2	O	Radio
?	Sedan	2	3	P	Paul Kalkbrenner
?	Sedan	2	4	Q	BLACKPINK
Hybrid	SUV	4 or 5	5	R	Taylor Swift
Hybrid	SUV	4 or 5	5	S	Radio
Hybrid	SUV	4 or 5	6	T	Taylor Swift
*					
*					

Table 4.8: An example anonymization of a dataset containing 10 trips. Some quasi-identifier values are generalized. A row containing a * indicates that this trip is suppressed.

Minimum k Because we define k-anonymity on the drivers, this metric reports the minimum number of drivers over all ECs in the dataset. In the example, we have a minimum k of 2 because the second EC holds two drivers while the first holds 4. We scale this value using the Minimum k of the input dataset as m_{min} and the number of drivers in the output dataset as m_{max}. Because our data contains vehicles with unique quasi-identifier values in the dataset, m_{min} is 1. The scaled metric value for the example table is $(2-1)/(6-1) = 1/5$.

Minimum l The sensitive value is set for each trip in the data. Therefore, with Minimum l, we measure the minimum number of distinct sensitive values over all ECs. The Minimum l in the example table is 2 due to the second EC only holding two distinct sensitive values. Similar to the Minimum k metric, we min-max scale the result of this metric. We set the Minimum l of the input dataset as m_{min} and the number of distinct sensitive values in the output dataset as m_{max}. In the example, we arrive at the scaled value $(2-1)/(5-1) = 1/4$.

Average Equivalence Class Size Similar to Minimum k, we calculate the Average Equivalence Class Size on the drivers. We do not include the normalization term described in Section 4.2.5. In the example, we have $(4+2)/2 = 3$. We again scale the value using the Average Equivalence Class Size of the input dataset as m_{min} and the number of drivers in the output dataset as m_{max}.

Discernibility Penalty The Discernibility Penalty measures the discernibility of records in the dataset, which is why we apply it to the trips. In the example, we have $5 \cdot 5 + 3 \cdot 3 = 34$ for the ECs and $2 \cdot 10 = 20$ for the suppressed trips, so 54 in total for the example data. It is difficult to assign the standalone value of this metric a semantic meaning, which is why we again apply a min-max scaling. We choose the Discernibility Penalty of the input dataset as m_{min}. m_{max} is set to the highest possible penalty for the output table as described in Section 4.2.6, which is the number of trips in the input squared.

In-Data Precision Loss We calculate the precision loss on a trip basis. Our data only contains categorical quasi-identifier attributes, for which we use the loss proposed by Iyengar [Iye02] described in Section 4.2.7. We have $(5 + 3/2 + 6)/(10 \cdot 3) = 12.5/30 \approx 0.42$. The normalization factor is the size of the input dataset times the number of attributes in the quasi-identifier. The penalty over the records in the first EC is 5 due to the engine values being generalized completely. The penalty incurred by the second EC is $3 \cdot 1/2$ as the penalty for generalization to the value "4 or 5" is $1/2$. The suppressed records contribute a penalty of $2 \cdot 3$.

Cross-Data Precision Loss The Cross-Data Precision Loss is also calculated on a trip level. We use the approach based on the closest common ancestor described in Section 4.2.8 to calculate the penalty for each cell. The metric value for the example data equals the In-Data Precision Loss if the anonymized dataset was generated using generalization or microaggregation. Consider the first Equivalence Class. The penalty incurred by each generalization to the value "?" is 1. In the second EC, the penalty for the value "4 or 5" is $1/2$ regardless of whether the value before generalization was 4 or 5. Thus, we have a total penalty that is equal to the In-Data Precision Loss. This is not true for Data Synthesis, where the In-Data Precision Loss is always 0, but the Cross-Data Precision Loss may be greater than 0 because of the creation of new values.

Earth Mover's Distance We use the formula for categorical attributes to calculate the Earth Mover's Distance because the sensitive attribute "Artist" is categorical. For the example table, we have probabilities $P_Q(\text{Kid Cudi}) = 1/8, P_Q(\text{Radio}) = 3/8, P_Q(\text{Paul Kalkbrenner}) = 1/8, P_Q(\text{BLACKPINK}) = 1/8, P_Q(\text{Taylor Swift}) = 2/8$. The Earth Mover's Distance for the first EC is

$$\frac{1}{2} \cdot (|\frac{1}{5} - \frac{1}{8}| + |\frac{2}{5} - \frac{3}{8}| + |\frac{1}{5} - \frac{1}{8}| + |\frac{1}{5} - \frac{1}{8}| + |0 - \frac{2}{8}|) = \frac{1}{4} = 0.25$$

The Earth Mover's Distance for the second EC is

$$\frac{1}{2} \cdot (|0 - \frac{1}{8}| + |\frac{1}{3} - \frac{3}{8}| + |0 - \frac{1}{8}| + |0 - \frac{1}{8}| + |\frac{2}{3} - \frac{2}{8}|) = \frac{5}{12} \approx 0.42$$

The metric result is the maximum Earth Mover's Distance over all ECs, which is $5/12$ for the example.

***g*-balance** The g-balance does not need to be adjusted as it already is defined on data with multiple records per individual. For the example, we have:

$$1 - ((\frac{2}{5})^2 + (\frac{1}{5})^2 + (\frac{1}{5})^2 + (\frac{1}{5})^2) = \frac{18}{25} = 0.72 \quad \text{for the first EC}$$
$$1 - ((\frac{2}{3})^2 + (\frac{1}{3})^2) = \frac{4}{9} \approx 0.44 \quad \text{for the second EC}$$

The g-balance for the data is the minimum of the two values, so $4/9$.

h-affiliation Similar to the g-balance, no adjustments are needed for this metric. For the h-affiliation, we have $2/4$ for the first EC because two out of 4 individuals have the sensitive value "Radio". For the second EC, we have a value of 1 because both drivers share a sensitive value. The h-affiliation is the maximum value over all ECs and, in this case, also the highest possible value, 1. We adjust the implementation of the h-affiliation for our data to ignore the sensitive value "Radio". This is because every driver in the data has at least one trip with this value associated with it. In the experiments, all selected approaches create at least one EC where the value "Radio" occurs for every driver. Therefore, the h-affiliation is 1 for all experiment evaluations. Note that this should not be understood as a weakness of the metric. Rather, we make this adjustment to obtain informative results in the experiments.

Adversarial Knowledge Gain To calculate the Adversarial Knowledge Gain, we use the additive difference to calculate the knowledge gain for each EC. The additive difference variant calculates the value for an EC similarly to the Earth Mover's Distance, which allows reusing the probabilities for the example dataset.

$$\frac{1}{8}(5 \cdot \frac{1}{4} + 3 \cdot \frac{5}{12}) = \frac{5}{16} = 0.3125$$

4.5 Results & Discussion

This section presents and discusses the results of the experiments. We first present an analysis of the applicability of the metrics to the approaches. Then, we present the results of applying the metrics, the Utility Demonstrator, and the Privacy Demonstrator on the data generated by the anonymization approaches. The section concludes with a discussion of these results and a consolidating evaluation of the selected metrics.

4.5.1 Applicability of Metrics

Table 4.9 shows the results of an applicability analysis of metrics on approaches. We omit the Differential Privacy on Queries approach because none of the selected metrics can be applied to it. All selected metrics can be applied to the k-anonymity through Generalization and k-anonymity through Microaggregation approaches. However, there is one detail. The In-Data Precision Loss is a metric that measures the uncertainty of attribute values in the anonymized data. Because the microaggregation function used for the approaches is the mode function, no uncertainty is present in the transformed values. The metric will only incorporate suppressed records in the calculation and, therefore, lose expressiveness when applied to this approach. However, it can still be applied, which we indicate with the parentheses in Table 4.9. The remaining metrics preserve full significance when applied on k-anonymity through Microaggregation.

Turning to Data Synthesis, we observe that the expressiveness of many metrics on this approach is limited. For example, the Suppression Ratio will always be 0 because no records are lost during anonymization. The syntactic metrics that evaluate the uniqueness of drivers in the ECs, i.e., Minimum k, Average Equivalence Class Size, and Discernibility Penalty, all lose expressiveness for the same reason. The quasi-identifier values are artificially created in this approach, so the

search for an individual in the anonymized data based on the quasi-identifier values might return the records of a different person or no records at all because the values are not present in the data. Even if there is only one driver in a given EC, their quasi-identifier values might not be accurate such that a direct identification using these values is not reliable. Metrics that measure the uniqueness of drivers in the EC as an indication of privacy are, therefore, not as expressive. A similar argument applies to the Minimum l metric. Similar values of the sensitive attribute in one EC lose their detriment to privacy if there is no reliable way of telling whether an individual is part of an EC. A similar argumentation can be applied to the Earth Mover's Distance, g-balance, h-affiliation, and the Adversarial Knowledge Gain. Because the rationale behind these metrics is that an adversary can dependably assert that an individual is part of an EC to extract meaningful information, they lose expressiveness. An argument can be made that an adversary is also not able to reliably find individuals in anonymized data generated by the k-anonymity through Microaggregation approach because raw values are replaced by the mode of the EC. However, as the mode are the values that appear most often in the EC, every group will have drivers that have their raw quasi-identifier values associated. Lastly, the In-Data Precision Loss is always 0 for the synthetic datasets because no records are suppressed, and there is no precision loss in the quasi-identifier values. The only metric that keeps its expressiveness for all approaches is the Cross-Data Precision Loss. The change in the quasi-identifier values from the raw to the anonymized data for each record in the data retains its meaning.

	k-anonymity through Generalization	k-anonymity through Microaggregation	Data Synthesis
Suppression Ratio	x	x	(x)
Minimum k	x	x	(x)
Minimum l	x	x	(x)
Average Equivalence Class Size	x	x	(x)
Discernibility Penalty	x	x	(x)
In-Data Precision Loss	x	(x)	(x)
Cross-Data Precision Loss	x	x	x
Earth Mover's Distance	x	x	(x)
g-balance	x	x	(x)
h-affiliation	x	x	(x)
Adversarial Knowledge Gain	x	x	(x)

Table 4.9: Theoretical applicability of metrics on methods. We omit the Differential Privacy on Queries approach because no metrics can be applied to it. The parentheses indicate that the metric loses expressiveness in combination with the approach.

4.5.2 Evaluation of Soft Metrics

This section presents the evaluation of the soft metrics introduced in Section 4.2.1. The results are displayed in Figure 4.8. We order the approaches relative to each other on scales from "Low" to "High" for each metric. The Suitability for on-line Data Collection is not evaluated because the approaches are not applied to data streams.

Raw Data Permanence Differential Privacy on Queries is the only approach that requires the raw data to be always available. The other three approaches need the raw data only once to create an anonymized dataset. The raw data can then be discarded.

Required Amount of Data Preprocessing Differential Privacy on Queries is an approach that requires no data preprocessing because the queries can be executed directly on the raw data. Data Synthesis, on the other hand, requires preprocessing such that a contingency table for providing the sampling probabilities can be computed. In our case, however, this preprocessing was minimal. Achieving k-anonymity requires data preparation such that each record corresponds to one individual. Furthermore, all quasi-identifiers must be explicitly present in the table structure as attributes. These approaches, therefore, require the highest amount of preprocessing.

Data Use Case Flexibility Differential Privacy on Queries outputs values that are exactly tailored to one query function. However, the approach can guarantee privacy for every real-valued query [DR13], making it highly flexible. This property is common for approaches from the PPDA field. On the other hand, the tables generated by the remaining approaches are useful for various data use cases, which is a common property of approaches in the PPDP field. However, the experiment results presented in Section 4.5.3 show that some data use cases are not realizable anymore on the anonymized data. Differential Privacy on Queries is more flexible for our specific use cases but can, for example, not be used to train a statistical classifier, which is possible using the anonymized datasets created by the other approaches. Therefore, we evaluate all approaches equally concerning the data use case flexibility. Further investigation is needed to assess this aspect more precisely.

Intuitiveness of Privacy Effect The approaches utilizing k-anonymity provide a very intuitive notion of privacy protection. The data of an individual will be hidden between at least $k - 1$ other people. This definition of anonymity is easy to grasp and easy to trust, especially for high k. For the Data Synthesis approach, it is understood that the quasi-identifier values are artificially generated, but the anonymized data still contain very specific information, although it is likely, not true information. The promise that DP makes is that any potential bad outcome for an individual is not influenced by the fact that this individual was part of the data. This notion of privacy is strong, but processing the meaning of this guarantee is more complex than the guarantees of the syntactic models. Furthermore, the privacy that the parameter ϵ provides is not easily comprehended because its impact on the privacy guarantee is abstract and technical.

Required Amount of Domain Knowledge We assign Differential Privacy on Queries the lowest score of the approaches because it is very generic for any given query function. Calculating the sensitivity of the function and applying noise to the true result does not require domain knowledge. However, the mechanism may output infeasible values as described in Section 4.4.3. Truncating out-of-bounds values requires some domain knowledge. The Data Synthesis approach is scored slightly higher, as some domain knowledge may be required to filter out infeasible combinations of synthetic quasi-identifier values sampled from the perturbed counts. For k-anonymity through Microaggregation, identification of quasi-identifiers requires some domain knowledge. Furthermore, choosing a fitting aggregation function depends on the specific data at hand. k-anonymity through Generalization requires the highest amount of domain knowledge. Additionally to defining quasi-identifiers, the data engineer needs to specify Value Generalization Hierarchies, which require a significant understanding of the underlying data and the semantics of the attribute values.

4.5.3 Experiment Results

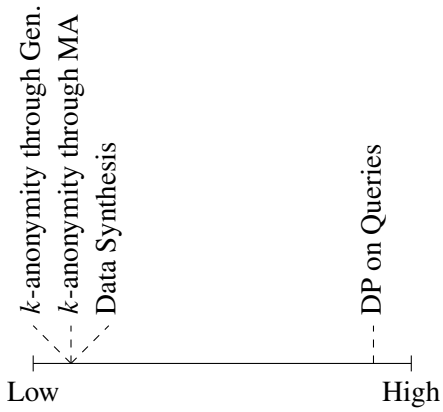
This section shows the evaluation of the metrics and demonstrators on the anonymization approaches. Table 4.10 shows the scores of the raw data, and Tables 4.11 to 4.14 show the evaluation of the anonymization approaches. For metric values that receive a scaling, we report the min-max scaled metric value first and the unscaled value in parentheses. Values are rounded to two decimal places, with exceptions for values that would round down to 0 or round up to one.

Raw Data

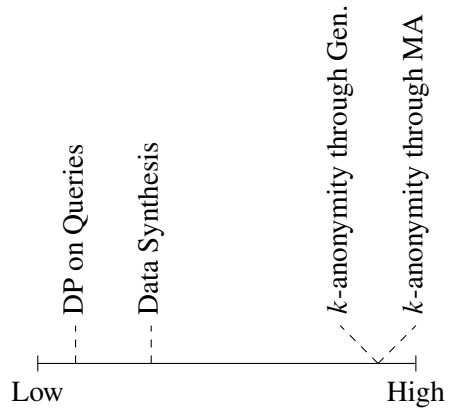
Table 4.10 shows the results of the metrics and demonstrators on the raw data. All min-max scaled metric values are naturally 0 because the respective m_{min} values used for scaling are also calculated on the raw data. The Suppression Ratio, In-Data Precision Loss, and Cross-Data Precision Loss are also trivially 0 as no records are suppressed and no quasi-identifier values transformed. We have a g-balance of 0 because there are ECs in the raw data that only contain the trips of a single driver. The h-affiliation is 1 as there again are drivers with unique quasi-identifier values that make up their own EC. All drivers in these ECs trivially share one sensitive value. The Earth Mover's Distance and Adversarial Knowledge Gain are both values that incorporate the difference in the distribution of sensitive values between the ECs and the entire table. Their values are not easily assigned an intuitive meaning but allow comparisons between the approaches.

The performance values on the UD are the true values of the KPI calculations.

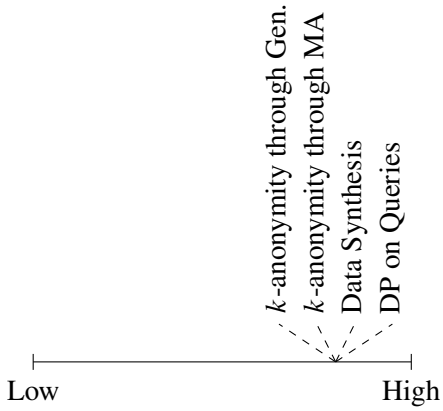
The PD shows that the raw data is heavily prone to homogeneity attacks. The attack compromises all drivers, although some drivers are not alone in an EC due to their quasi-identifier values not being unique. Furthermore, the maximum recall over all drivers is 100%, which means that the adversary could extract all existing location information of at least one driver. This is not astonishing, as single-driver ECs automatically allow an adversary to have full recall. On average, the adversary can extract around 35% of AOIs over all drivers



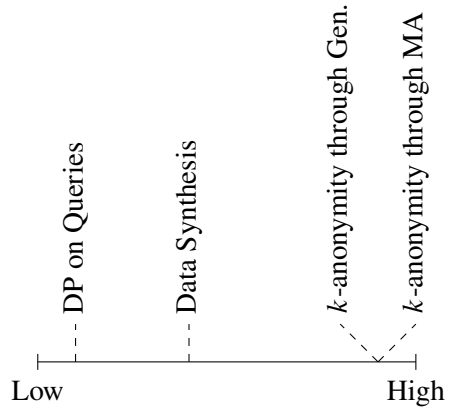
(a) Raw Data Permanence.



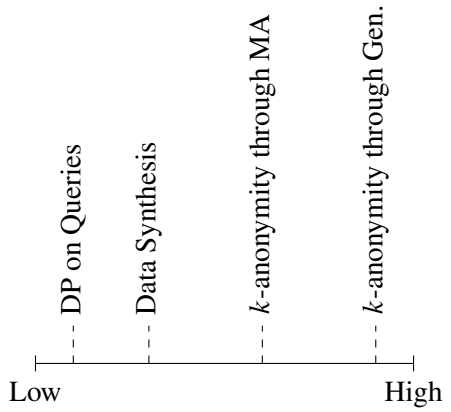
(b) Required Amount of Data Preprocessing.



(c) Data Use Case Flexibility.



(d) Intuitiveness of Privacy Effect.



(e) Required Amount of Domain Knowledge.

Figure 4.8: The evaluation of the soft metrics on the approaches. We rank the approaches on a scale from "Low" to "High" for each metric.

Raw Data		
Metric	Suppression Ratio	0
	Minimum k	0 (1)
	Minimum l	0 (1)
	Average Equivalence Class Size	0 (1.67)
	Discernibility Penalty	0 ($6.6 \cdot 10^6$)
	In-Data Precision Loss	0
	Cross-Data Precision Loss	0
	Earth Mover's Distance	0.42
	g-balance	0
	h-affiliation	1
	Adversarial Knowledge Gain	0.21
UD	KPI Overall	42.15%
	KPI Electric	38.05%
	KPI Gas	47.36%
	KPI Hybrid	38.89%
PD	Compromised Drivers	100% (13/13)
	Maximum Recall	100%
	Average Recall	34.67%

Table 4.10: Metric and demonstrator evaluation on the raw data. The values in parentheses for the metrics are the raw, unnormalized values.

k-anonymity through Generalization

Table 4.11 shows the evaluation of the metrics and demonstrators to the k-anonymity through Generalization approach for different privacy parameters k. Because this method employs suppression, we observe that the Suppression Ratio is greater than 0 and grows with rising k to fulfill the size requirement of the ECs. The Minimum k and Average Equivalence Class Size values indicate that all drivers are grouped into one EC for $k = 10$. As expected, the Discernibility Penalty grows with bigger values of k because fewer trips in the data are indistinguishable. Both distance-based metric values match for anonymized datasets created by generalization, as explained in Section 4.4.5. The values show a significant jump between the different values of k for two main reasons. Firstly, the number of suppressed trips increases from $k = 2$ to $k = 5$. Secondly, quasi-identifier attribute values are generalized heavily to group the drivers into EC of sufficient size. The Earth Mover's Distance reduces significantly for $k = 2$ and is 0 for $k = 10$ because the entire anonymized table comprises only one EC. g-balance increases for $k > 1$ as one EC is guaranteed to contain the trips of at least two drivers. The jump from $k = 2$ to $k = 5$ is more significant than the last jump to $k = 10$. The h-affiliation is 1 for $k = 2$, meaning there is at least one EC where all drivers share the same sensitive value. The value is halved for $k = 5$ and approximately halved again for $k = 10$. Lastly, the Adversarial Knowledge Gain knowledge gain does not change by much for $k = 2$. Similar to the Earth Mover's Distance, the metric value for $k = 10$ is 0 because the only EC spans the entire table.

k-anonymity through Generalization		$k = 2$	$k = 5$	$k = 10$
Metric	Suppression Ratio	0.01	0.18	0.18
	Minimum k	0.08 (2)	0.4 (5)	1 (11)
	Minimum l	0.01 (8)	0.24 (136)	1 (575)
	Average Equivalence Class Size	0.05 (2.33)	0.41 (5.5)	1 (11)
	Discernibility Penalty	0.03 ($7.2 \cdot 10^6$)	0.41 ($1.7 \cdot 10^7$)	0.81 ($2.6 \cdot 10^7$)
	In-Data Precision Loss	0.05	0.56	0.91
	Cross-Data Precision Loss	0.05	0.56	0.91
	Earth Mover's Distance	0.29	0.18	0
	g-balance	0.31	0.71	0.87
	h-affiliation	1	0.5	0.27
	Adversarial Knowledge Gain	0.2	0.13	0
UD	KPI Overall	0.001 (42.2%)	0.03 (40.89%)	0.03 (40.89%)
	KPI Electric	0	0.07 (40.89%)!	0.07 (40.89%)!
	KPI Gas	0	0.14 (40.89%)!	0.14 (40.89%)!
	KPI Hybrid	-	0.05 (40.89%)!	0.05 (40.89%)!
PD	Compromised Drivers	100% (13/13)	76.92% (10/13)	0% (0/13)
	Maximum Recall	100%	50%	0%
	Average Recall	21.92%	3.94%	0%

Table 4.11: The evaluation of the k-anonymity through Generalization approach. For the metrics, we report the raw, unnormalized value in parentheses. For the Utility Demonstrator, we report the relative error and the raw KPI output in parentheses. The entries are suffixed with a "!" if the exact value can not be calculated due to generalization.

The k-anonymity through Generalization approach shows some interesting characteristics when considering the evaluation of the UD. Firstly, there is a minimal error on the KPI calculation over all drivers in the data, although almost 20% trips are suppressed for $k = 5$ and $k = 10$. We presume that the reason for this is that the usage of the feature does not fluctuate heavily from driver to driver. However, the KPI calculation by engine type shows the impact of generalization eminently. Calculating the KPI values for $k = 2$ shows no error on engine types "Electric" and "Gas", but is not possible for type "Hybrid". The reason is that the anonymization suppresses the only driver with a hybrid vehicle. It is, therefore, not possible to calculate the KPI for this engine type. The anonymized data for $k = 5$ and $k = 10$ show an emerging problem with value generalization. We are not able to calculate the KPI values for the engine types precisely. The reason for this is the generalization of the engine types. If "EV" and "Hybrid" are generalized into the value "With E", problems arise for the calculation of the KPI for the engine types. We choose to include the trips with generalized engine values in the calculations. So, trips with engine type "?" or "With E" are included in the KPI calculation for engine type "Hybrid" or "EV". The relaxation of the calculation is indicated in the table with a "!" after the entry. For $k = 10$, all trips are included for each calculation of the KPI, which is why the percentages given are equal.

The evaluation of the Privacy Demonstrator for k-anonymity through Generalization shows that anonymization with $k = 2$ shows almost no improvement concerning an adversarial attack for our data. Although the average recall value of the adversary drops, all drivers are compromised, and the

adversary can extract all AOIs for at least one driver. The success of an adversarial attack starts to decrease significantly for $k = 5$. The number of compromised drivers drops from 13 to 10, and the maximum recall from 100% to 50%. The anonymized data satisfying 10-anonymity prevents all position disclosure because there are no AOIs shared by all 11 drivers in the dataset.

k-anonymity through Microaggregation

k-anonymity through Microaggregation		$k = 2$	$k = 5$	$k = 10$
Metric	Suppression Ratio	0.01	0.18	0.18
	Minimum k	0.08 (2)	0.4 (5)	1 (11)
	Minimum l	0.01 (8)	0.24 (136)	1 (575)
	Average Equivalence Class Size	0.05 (2.33)	0.41 (5.5)	1 (11)
	Discernibility Penalty	0.03 ($7.2 \cdot 10^6$)	0.41 ($1.7 \cdot 10^7$)	0.81 ($2.6 \cdot 10^7$)
	In-Data Precision Loss	0.01	0.18	0.18
	Cross-Data Precision Loss	0.04	0.3	0.44
	Earth Mover's Distance	0.29	0.18	0
	g-balance	0.31	0.71	0.87
	h-affiliation	1	0.5	0.27
	Adversarial Knowledge Gain	0.2	0.13	0
UD	KPI Overall	0.001 (42.2%)	0.03 (40.89%)	0.03 (40.89%)
	KPI Electric	0	0.04 (36.55%)	-
	KPI Gas	0	0.02 (48.25%)	0.14 (40.89%)
	KPI Hybrid	-	-	-
PD	Compromised Drivers	84.62% (11/13)	46.15% (6/13)	0% (0/13)
	Maximum Recall	100%	50%	0%
	Average Recall	12.81%	3.85%	0%

Table 4.12: The evaluation of the k-anonymity through Microaggregation approach. For the metrics, we report the raw, unnormalized value in parentheses. For the Utility Demonstrator, we report the relative error and the raw KPI output in parentheses.

Table 4.12 shows the experiment evaluation of the k-anonymity through Microaggregation approach. The metric values for the syntactic values match the results for k-anonymity through Generalization because the algorithm creates the same ECs for both approaches. However, the distance-based metrics now show a difference between the two approaches. For one, we observe that the In-Data Precision Loss values match those of the Suppression Ratio. Because the loss due to generalization captured by the In-Data Precision Loss is 0 for data transformed by microaggregation, the result only captures the loss for suppressed trips, which is equal to the Suppression Ratio. The Cross-Data Precision Loss, on the other hand, captures the change in attribute values due to their replacement by the mode in the EC. We observe that the metric value is relatively moderate, even for $k = 10$. All distribution-based metrics and the Adversarial Knowledge Gain show the same metric values as for k-anonymity through Generalization.

However, the evaluation of the UD shows some differences. Firstly, the overall errors are equal to the ones for k-anonymity through Generalization for all k. The table also shows that the KPI for engine type "Hybrid" can not be calculated for any k. The reason for this is that the trips with this engine type are suppressed for every k, and the absence of generalized values does not allow a relaxed calculation like the one used for k-anonymity through Generalization. Also, observe that the KPI for engine type "Electric" and $k = 10$ can not be calculated because all engine values in the only EC have been aggregated to "Gas".

k-anonymity through Microaggregation performs better on the Privacy Demonstrator because the aggregation of values hampers an attacker from linking quasi-identifier values of the drivers in the raw data to trips in the anonymized data. Therefore, the number of compromised drivers drops already for $k = 2$, which is not observed for the k-anonymity through Generalization approach. The attack is again fully thwarted for $k = 10$.

Data Synthesis

Data Synthesis		-	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 0.1$
Metric	Suppression Ratio	0	0	0	0
	Minimum k	0.001 (1.02)	0.001 (1.01)	0.0003 (1.004)	0.0006 (1.008)
	Minimum l	0.02 (13.58)	0.02 (11.65)	0.005 (4.27)	0.008 (6.4)
	Average Equivalence Class Size	0.04 (2.26)	0.03 (2.01)	-0.001 (1.65)	0.01 (1.84)
	Discernibility Penalty	0.03 ($7.2 \cdot 10^6$)	-0.001 ($6.5 \cdot 10^6$)	-0.06 ($5.2 \cdot 10^6$)	-0.02 ($6.1 \cdot 10^6$)
	In-Data Precision Loss	0	0	0	0
	Cross-Data Precision Loss	0.53	0.54	0.59	0.62
	Earth Mover's Distance	0.4	0.43	0.46	0.45
	g-balance	0.004	0.003	0.002	0.002
	h-affiliation	0.99	0.99	0.999	0.998
	Adversarial Knowledge Gain	0.21	0.22	0.24	0.23
UD	KPI Overall	0	0	0	0
	KPI Electric	0.13	0.13	0.16	0.19
	KPI Gas	0.13	0.13	0.15	0.18
	KPI Hybrid	0.53	0.47	0.25	0.19
PD	Compromised Drivers	70.29% (9.14/13)	68.65% (8.93/13)	53.04% (6.9/13)	24.37% (3.17/13)
	Maximum Recall	56.91%	52.78%	50.3%	30.67%
	Average Recall	8.26%	7.85%	7.33%	3.7%

Table 4.13: The evaluation of the Data Synthesis approach for no Differential Privacy and for Differential Privacy with $\epsilon = 10, 1, 0.1$. The calculations for the metrics were repeated over 250 runs and averaged. For the metrics, we report the raw, unnormalized value in parentheses. The evaluation for the Utility Demonstrator and Privacy Demonstrator was repeated 1000 times, and the results averaged. For the Utility Demonstrator, we report the relative error.

Table 4.13 shows the experiment results for the Data Synthesis approach. The calculations for the metrics are executed 250 times and averaged. The evaluation of the UD and PD are averaged over 1000 runs. Should a certain engine type not be present in the anonymized data, we set the full relative error of 1 for the KPI calculation of this engine type. The Suppression Ratio is 0 for all

possible synthetic datasets because no trips are lost during anonymization. The In-Data Precision Loss is also 0 for every synthetic dataset because the quasi-identifier values are not generalized, and there is no suppression of trips. For the other syntactic metrics, the values show little derivation from the raw data. Some entries are negative because the sampling process may create tables where the Average Equivalence Class Size or the Discernibility Penalty are lower than for the raw data. The numerator of the normalization term will be < 0 in these cases. The Earth Mover's Distance lingers at a relatively high value for all Data Synthesis variants. The low expected amount of drivers in an EC leads to big differences between the distribution of sensitive values of ECs and the entire table. g-balance values are close to 0 also because ECs are expected to contain few drivers. The h-affiliation is almost 1 across the board because the probability of an EC with only one driver is very high in the synthetic data. The Adversarial Knowledge Gain is at a value comparable to the raw data for each variant of the approach. The values are consistently high for the same reason that the Earth Mover's Distance values are high.

The evaluation of the UD shows 0 error for the KPI calculation over all vehicles. Again, this is because the anonymized data contains all drivers and trips of the raw data. The errors on the KPI calculation by engine type, however, are comparatively high. Especially the KPI for hybrid vehicles suffers from the fact that this engine type is underrepresented in the raw data. This results in a low sampling probability for the engine type and, thus, a high error because a non-existence of the engine type is penalized with the full error. As the values of ϵ grow, this underrepresentation gets compensated more and more through the applied noise, and the error goes down.

The results on the Privacy Demonstrator show that the attack is still possible even if finding candidate drivers in the anonymized data for a target driver is hampered due to the sampling of new quasi-identifier values. If drivers are assigned the same quasi-identifier values by chance, they become very prone to the attack. Therefore, it may be advisable to prevent assigning the same quasi-identifier values to a driver in the anonymized dataset. However, it should be noted that this impacts the data utility negatively.

Differential Privacy on Queries

Table 4.14 shows the evaluation of the Differential Privacy on Queries approach on the Utility Demonstrator. We report the sensitivity of each KPI calculation and the mean relative error of the output over $100,000$ runs. For each KPI calculation, the error for the privacy setting $\epsilon = 0.1$ is too high to output useful values. The reason for this is the high sensitivity of the KPI functions on the data. Because the data is comprised of only 15 drivers, the removal or addition of one driver impacts the KPI result noticeably. This becomes especially evident when considering the errors of the KPI on hybrid vehicles. The dataset only contains one hybrid vehicle, so the sensitivity of this KPI is exactly the KPI result.

Differential Privacy on Queries		Sensitivity	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 0.1$
UD	KPI Overall	3%	0.007	0.07	0.57
	KPI Electric	4.66%	0.01	0.12	0.79
	KPI Gas	4.96%	0.01	0.1	0.66
	KPI Hybrid	38.89%	0.1	0.71	1.2

Table 4.14: The evaluation of the Differential Privacy on Queries approach for $\epsilon = 10, 1, 0.1$. We report the KPI sensitivity and the mean relative error over $100,000$ runs.

4.5.4 Discussion

In this section, we process the results of the experiments, point out the strengths and weaknesses of the anonymization approaches and metrics, and outline the conclusions drawn. The chosen anonymization approaches employ different data transformation techniques to generate the anonymized data. Each technique has implications for the utility and privacy of the output dataset and the meaningfulness of applying metrics to the output.

k-anonymity through Generalization shows good preservation of data utility in our experiments as the generalization of values is truthful, such that queries operating on the quasi-identifier values can still be executed with some relaxations. However, outlier tuples are suppressed to avoid heavy generalization of the remaining records. This leads to utility loss that can render certain data use cases unrealizable. The utility metrics all indicate the loss of information for growing k and assign a very high penalty to $k = 10$. The In-Data Precision Loss is specifically designed to measure the information loss due to generalization.

The Privacy Demonstrator results show considerable information leakage even for moderate values of the privacy parameter k. The possibility of homogeneity attacks on data that satisfies k-anonymity for a low value of k should not be underestimated. Therefore, we suggest that this attack is considered when choosing the parameter k with this anonymization approach. A high value of k completely prevents any sensitive GPS disclosure due to the attack without an increase in the error for the Utility Demonstrator. The Earth Mover's Distance, h-affiliation, and Adversarial Knowledge Gain also indicate the decreasing information gain on the sensitive attribute for higher values of k. The value 1 of the h-affiliation for $k = 2$ indicates that at least one EC is prone to a homogeneity attack on the sensitive value. The probability of a matching artist between drivers may appear slim because hundreds of artists are in the data. However, the circumstance that drivers have multiple trips with possibly different artists associated with them is a significant booster to this probability.

k-anonymity through Microaggregation suffers from the same drawbacks as k-anonymity through Generalization concerning the suppression of tuples. Some data use cases can not be performed on the anonymized data. The approach additionally suffers from the substitution of values by their mode. For $k = 10$, there is only one EC in the entire table. We can only calculate the KPI for the specific engine value present in this EC, while the other KPI calculations are not doable. This issue becomes less significant if the size of the raw data grows as the anonymized data will likely contain more ECs. However, k-anonymity through Microaggregation performs better regarding the Privacy Demonstrator. Although both approaches group the same records together into an EC, the substitution of the true quasi-identifier values reduces an adversary's chances of finding candidate

drivers in the anonymized data. Thus, we see a lower percentage of compromised drivers, and the average recall of the adversary drops for $k = 2$. The ratio of compromised drivers is still lower for $k = 5$, and both approaches prevent the attack for $k = 10$. Despite this, we do not see a reduction in the maximum recall, which means that the maximum amount of position disclosure for one individual does not change between the approaches.

The information loss introduced by anonymization through Data Synthesis is not captured at all by the syntactic metrics. The Cross-Data Precision Loss, on the other hand, captures the loss of introducing new quasi-identifier values better. The values for this metric are high throughout all variants of the approach and even grow for larger values of ϵ. In general, the experiments back up the theoretical analysis of the applicability of the metrics in that the Cross-Data Precision Loss is the only metric that keeps its expressiveness for each of the anonymization approaches. The performance on the Utility Demonstrator matches the metric values. The errors are high even for no perturbation of the sampling probabilities. We, therefore, conclude that data synthesis incurs the highest utility loss of all approaches tested. Interestingly enough, attacks on individuals' sensitive data are still possible with considerable recall values. Even for the strongest privacy guarantee $\epsilon = 0.1$, the maximum recall lies at around 30%, a remarkable value. The approaches on k-anonymity can prevent an attack when k is strict enough. The Earth Mover's Distance and h-affiliation suggest that the data is prone to the disclosure of sensitive values, but as explained in Appendix A.1, these metrics lose expressiveness when applied to synthetic data.

Differential Privacy on Queries is the approach that reacts most drastically to privacy parameter changes. The error on the Utility Demonstrator increases rapidly with stricter privacy guarantees. The reason for this is arguably the sensitivity of the KPIs. With data that contains a small number of individuals, removing or adding an individual to the data results in a noticeable change, hence a high sensitivity. This is especially noticeable for underrepresented records, which in our case is the one individual driving a hybrid. With a higher number of drivers, sensitivity decreases, and the accuracy of the approach becomes acceptable even for low values of ϵ. Therefore, we conclude that approaches fulfilling Differential Privacy should be applied when there is a great amount of data.

One last remark added is found in the work of Brickell and Shmatikov [BS08]. They point out that it is imperative to evaluate the data utility of the trivially anonymized dataset, denoted with $\mathcal{U}_{base}$, and the utility of the raw, unsanitized dataset, denoted with $\mathcal{U}_{max}$. The trivially anonymized dataset has all its quasi-identifier values suppressed. If $\mathcal{U}_{san}$, the utility of the sanitized dataset is close to $\mathcal{U}_{base}$, it may not be worth taking the privacy risk that a release of the sanitized table would mean. Instead, a release of the trivially anonymized dataset should be considered, which would maximize privacy. Should $\mathcal{U}_{max}$, which is an upper bound on the utility of this specific dataset, be very low, there may be a need to reconsider using this specific workload to measure utility because it performs poorly even on the raw data. Their suggestion can be transferred to our experiments. Consider the results of k-anonymity through Generalization for $k = 5$ and for $k = 10$. The Utility Demonstrator shows that both approaches perform equally on the KPIs. The Privacy Demonstrator indicates that the dataset with $k = 5$ leaks sensitive information such that the majority of drivers are compromised. Considering the data use cases modeled by the UD, publishing the dataset for $k = 10$ is preferable because the utility remains unchanged while the privacy is massively improved.

Key Takeaways In general, we observe that approaches working with the suppression of tuples may pose problems regarding the feasibility of certain data use cases. The reason for the suppression of records in our experiments is an underrepresentation of the quasi-identifier values of some entries. However, these rare entries may be important for certain data analysis scenarios. The ability to investigate and work with data of seldomly sold goods is still essential for any manufacturer interested in product improvement. Next, the Cross-Data Precision Loss appears to capture the notion of information loss measured by the UD best out of the utility metrics. It is also the only metric that outputs meaningful values for the Data Synthesis approach. The most important takeaway, however, is the necessity of concrete demonstrators that evaluate the utility and privacy protection of the anonymized data. The Utility Demonstrator shows that some approaches limit the variety of data analysis tasks that can be performed on the data. This fact is only explicitly shown by the demonstrator evaluation because no metrics can capture it. The Privacy Demonstrator shows that sensitive information can be extracted from the anonymized data on an unexpected scale. These demonstrators need to incorporate the specifications and peculiarities of the data to provide meaningful results. The selected metrics are (naturally) defined in a more general-purpose manner.

5 Conclusion & Outlook

This thesis evaluates metrics on different anonymization approaches for vehicle data. The challenges and peculiarities of data in the automotive field are presented and discussed. Metrics that evaluate the utility and privacy of anonymized data in the literature are collected and introduced. We select eleven metrics to assess four anonymization approaches applied to real-world vehicle data. Furthermore, a selection of soft metrics that assess high-level characteristics of the anonymization approaches is proposed. We explain the adjustments made to the existing approaches and metrics to apply them on the given vehicle data We define a Utility Demonstrator that evaluates the utility of the anonymized data for a practical use case in the automotive field. Analogously, a Privacy Demonstrator simulates an adversary trying to extract sensitive GPS information of the individuals in the data. Experiments that apply the metrics, Utility Demonstrator, and Privacy Demonstrator to the anonymization approaches show that few metrics provide expressive results for each approach. Furthermore, the importance of assessing the data utility and privacy with domain-specific demonstrators becomes imminent when anonymized data are shown to leak sensitive information and fail to allow the realization of the chosen data analyses.

Outlook

The field of metrics for the evaluation of anonymization approaches is vast, and many metrics have been proposed. Appendix A.1 list a considerable amount of additional metrics that were encountered during research for this thesis. We do not claim that the list is comprehensive and, therefore, leave an extensive literature review of metrics for future work. Furthermore, implementing all metrics presented in the appendix was out of scope for this work. Future work can use the metrics listed in Appendix A.1 as a starting point for further investigation regarding their applicability. The experiments show that domain-specific evaluations of the anonymized data can capture more meaningful aspects of data utility and privacy, so creating these domain-specific measures is also open for research.

The data used in the experiments contain the trips of a comparatively small amount of 15 drivers. In production, the fleet of connected vehicles is magnitudes more extensive. Our experiments show that some approaches incur a loss of information because underrepresented vehicle specifications are not handled adequately.

Another topic of research is further investigation of the challenges and peculiarities of data in the vehicle domain. The anonymization of the data used in this work required some adjustments to the approaches because of the data characteristics. Future work can investigate different anonymization techniques specifically suited to the data. These techniques include anonymizing data containing sensitive GPS signals, anonymizing data containing multiple records per individual, and anonymizing data containing multiple types of sensitive values. Additionally, researching approaches operating on streaming vehicle data is a field that remains open for contribution.

Bibliography

[37709] I. 3779:2009. *Road vehicles — Vehicle identification number (VIN) — Content and structure*. en. Standard. Geneva, CH: International Organization for Standardization, Oct. 2009. URL: https://www.iso.org/standard/52200.html (cit. on p. 40).

[AA01] D. Agrawal, C. C. Aggarwal. "On the Design and Quantification of Privacy Preserving Data Mining Algorithms". In: *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '01. New York, NY, USA: Association for Computing Machinery, May 2001, pp. 247–255. ISBN: 978-1-58113-361-5. DOI: 10.1145/375551.375602. (Visited on 05/21/2023) (cit. on pp. 86, 94).

[AAC+12] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, C. Palamidessi. "Differential Privacy: On the Trade-Off between Utility and Information Leakage". In: *Formal Aspects of Security and Trust*. Ed. by D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, G. Barthe, A. Datta, S. Etalle. Vol. 7140. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 39–54. ISBN: 978-3-642-29419-8 978-3-642-29420-4. DOI: 10.1007/978-3-642-29420-4_3. (Visited on 07/18/2023) (cit. on pp. 88, 89).

[AAR21] Anco Hundepool, Aad van de Wetering, Ramya Ramaswamy. *μ-ARGUS*. https://research.cbs.nl/casc/mu.htm. 2021. (Visited on 08/08/2023) (cit. on p. 95).

[ACG+16] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 308–318. ISBN: 978-1-4503-4139-4. DOI: 10.1145/2976749.2978318. (Visited on 05/10/2023) (cit. on pp. 30, 97).

[AEP23] AEPD-EDPS. *AEPD-EDPS Joint Paper on 10 Misunderstandings Related to Anonymisation | European Data Protection Supervisor*. https://edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en. Nov. 2023. (Visited on 11/09/2023) (cit. on p. 17).

[AMCM14] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, L. Murphy. "A Systematic Comparison and Evaluation of K-Anonymization Algorithms for Practitioners". In: (2014) (cit. on pp. 15, 19, 33, 34).

[AS00] R. Agrawal, R. Srikant. "Privacy-Preserving Data Mining". In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. SIGMOD '00. New York, NY, USA: Association for Computing Machinery, May 2000, pp. 439–450. ISBN: 978-1-58113-217-5. DOI: 10.1145/342009.335438. (Visited on 05/30/2023) (cit. on p. 94).

[AS19] M. Alzantot, M. Srivastava. *Differential Privacy Synthetic Data Generation using WGANs*. Version 1.0. 2019. URL: https://github.com/nesl/nist_differential_privacy_synthetic_data_challenge/ (cit. on p. 97).

[ASMH16] M. M. Almasi, T. R. Siddiqui, N. Mohammed, H. Hemmati. "The Risk-Utility Tradeoff for Data Privacy Models". In: *2016 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*. Nov. 2016, pp. 1–5. DOI: 10.1109/NTMS.2016.7792481 (cit. on pp. 16, 86).

[Aut] T. F. Authors. *Flower: A Friendly Federated Learning Framework*. https://flower.dev/. (Visited on 08/08/2023) (cit. on p. 97).

[BA05] R. Bayardo, R. Agrawal. "Data Privacy through Optimal K-Anonymization". In: *21st International Conference on Data Engineering (ICDE'05)*. Tokyo, Japan: IEEE, 2005, pp. 217–228. ISBN: 978-0-7695-2285-2. DOI: 10.1109/ICDE.2005.42. (Visited on 04/24/2023) (cit. on pp. 32, 91).

[BDMN05] A. Blum, C. Dwork, F. McSherry, K. Nissim. "Practical Privacy: The SuLQ Framework". In: *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '05. New York, NY, USA: Association for Computing Machinery, June 2005, pp. 128–138. ISBN: 978-1-59593-062-0. DOI: 10.1145/1065167.1065184. (Visited on 06/20/2023) (cit. on p. 22).

[BF19] F. Bonchi, E. Ferrari, eds. *Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques*. Boca Raton: CRC Press, Sept. 2019. ISBN: 978-0-429-07009-9. DOI: 10.1201/b10373 (cit. on p. 19).

[BKBL07] J.-W. Byun, A. Kamra, E. Bertino, N. Li. "Efficient K-Anonymization Using Clustering Techniques". In: *Advances in Databases: Concepts, Systems and Applications*. Ed. by D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, R. Kotagiri, P. R. Krishna, M. Mohania, E. Nantajeewarawat. Vol. 4443. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 188–200. ISBN: 978-3-540-71702-7 978-3-540-71703-4. DOI: 10.1007/978-3-540-71703-4_18. (Visited on 07/06/2023) (cit. on pp. 20, 34).

[BKMR21] K. Bonawitz, P. Kairouz, B. McMahan, D. Ramage. "Federated Learning and Privacy: Building Privacy-Preserving Systems for Machine Learning and Data Science on Decentralized Data". In: *Queue* 19.5 (Oct. 2021), pp. 87–114. ISSN: 1542-7730, 1542-7749. DOI: 10.1145/3494834.3500240. (Visited on 08/17/2023) (cit. on p. 22).

[BKP18] R. Bild, K. A. Kuhn, F. Prasser. "SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees". In: *Proceedings on Privacy Enhancing Technologies* 2018.1 (Jan. 2018), pp. 67–87. ISSN: 2299-0984. DOI: 10.1515/popets-2018-0004. (Visited on 07/11/2023) (cit. on pp. 21, 44, 85).

[BL20] C. M. Bowen, F. Liu. "Comparative Study of Differentially Private Data Synthesis Methods". In: *Statistical Science* 35.2 (May 2020). ISSN: 0883-4237. DOI: 10.1214/19-STS742. (Visited on 10/03/2023) (cit. on pp. 17, 22, 43, 49).

[BS08] J. Brickell, V. Shmatikov. "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. New York, NY, USA: Association for Computing Machinery, Aug. 2008, pp. 70–78. ISBN: 978-1-60558-193-4. DOI: 10.1145/1401890.1401904. (Visited on 05/10/2023) (cit. on pp. 20, 38, 66, 85, 87, 88).

[BTM+20] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, N. Lane. "Flower: A Friendly Federated Learning Research Framework". In: July 2020. (Visited on 08/17/2023) (cit. on pp. 22, 97).

[BV] E. B.V. *Scopus*. URL: https://www.scopus.com/ (visited on 11/11/2023) (cit. on p. 15).

[CK12] J. Cao, P. Karras. "Publishing Microdata with a Robust Privacy Guarantee". In: *Proceedings of the VLDB Endowment* 5.11 (July 2012), pp. 1388–1399. ISSN: 2150-8097. DOI: 10.14778/2350229.2350255. (Visited on 07/11/2023) (cit. on p. 93).

[CKLM09] B.-C. Chen, D. Kifer, K. LeFevre, A. Machanavajjhala. "Privacy-Preserving Data Publishing". In: *Foundations and Trends® in Databases* 2.1-2 (2009), pp. 1–167. ISSN: 1931-7883, 1931-7891. DOI: 10.1561/1900000008. (Visited on 07/10/2023) (cit. on pp. 17, 22).

[CM16] R. Coppola, M. Morisio. "Connected Car: Technologies, Issues, Future Trends". In: *ACM Computing Surveys* 49.3 (Oct. 2016), 46:1–46:36. ISSN: 0360-0300. DOI: 10.1145/2971482. (Visited on 11/08/2023) (cit. on pp. 13, 15, 25).

[CPE+13] G. Cormode, C. M. Procopiuc, Entong Shen, D. Srivastava, Ting Yu. "Empirical Privacy and Empirical Utility of Anonymized Data". In: *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)* (Apr. 2013), pp. 77–82. DOI: 10.1109/ICDEW.2013.6547431. (Visited on 07/11/2023) (cit. on pp. 16, 28, 39, 88, 90).

[CPS+12] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, T. Yu. "Differentially Private Spatial Decompositions". In: *2012 IEEE 28th International Conference on Data Engineering* (Apr. 2012), pp. 20–31. DOI: 10.1109/ICDE.2012.16. (Visited on 08/23/2023) (cit. on p. 21).

[CT13] C. Clifton, T. Tassa. "On Syntactic Anonymity and Differential Privacy". In: *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)* (Apr. 2013), pp. 88–93. DOI: 10.1109/ICDEW.2013.6547433. (Visited on 07/11/2023) (cit. on p. 16).

[CWK+08] S. E. Coull, C. V. Wright, A. D. Keromytis, F. Monrose, M. K. Reiter. "Taming the Devil: Techniques for Evaluating Anonymized Network Data". In: Columbia University, 2008. DOI: 10.7916/D8BC47W0. (Visited on 05/08/2023) (cit. on p. 86).

[Dal86] T. Dalenius. "Finding a Needle In a Haystack or Identifying Anonymous Census Records". In: *Journal of Official Statistics* 2.3 (Sept. 1986). ISSN: 0282423X. (Visited on 04/24/2023) (cit. on p. 29).

[DGB+09] C. Dai, G. Ghinita, E. Bertino, J.-W. Byun, N. Li. "TIAMAT: A Tool for Interactive Analysis of Microdata Anonymization Techniques". In: *Proceedings of the VLDB Endowment* 2.2 (Aug. 2009), pp. 1618–1621. ISSN: 2150-8097. DOI: 10.14778/1687553.1687607. (Visited on 07/28/2023) (cit. on p. 95).

[Dim] M. T. Dimakopoulos Dimitris Tsitsigkos and Nikolaos. *Amnesia Anonymization Tool - Data Anonymization Made Easy*. https://amnesia.openaire.eu/. (Visited on 08/08/2023) (cit. on p. 95).

[DM02] J. Domingo-Ferrer, J. Mateo-Sanz. "Practical Data-Oriented Microaggregation for Statistical Disclosure Control". In: *IEEE Transactions on Knowledge and Data Engineering* 14.1 (Jan. 2002), pp. 189–201. ISSN: 1558-2191. DOI: 10.1109/69.979982 (cit. on pp. 20, 84).

[DMB20] J. Domingo-Ferrer, K. Muralidhar, M. Bras-Amoros. "General Confidentiality and Utility Metrics for Privacy-Preserving Data Publishing Based on the Permutation Model". In: *IEEE Transactions on Dependable and Secure Computing* (2020), pp. 1–1. ISSN: 1545-5971, 1941-0018, 2160-9209. DOI: 10.1109/TDSC.2020.2968027. (Visited on 10/07/2023) (cit. on p. 22).

[DMNS06] C. Dwork, F. McSherry, K. Nissim, A. Smith. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Theory of Cryptography*. Ed. by D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, S. Halevi, T. Rabin. Vol. 3876. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284. ISBN: 978-3-540-32731-8 978-3-540-32732-5. DOI: 10.1007/11681878_14. (Visited on 09/29/2023) (cit. on p. 46).

[DR13] C. Dwork, A. Roth. "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4 (2013), pp. 211–407. ISSN: 1551-305X, 1551-3068. DOI: 10.1561/0400000042. (Visited on 09/29/2023) (cit. on pp. 17, 21, 43, 46, 49, 57).

[DRS15] J. Domingo-Ferrer, S. Ricci, J. Soria-Comas. "Disclosure Risk Assessment via Record Linkage by a Maximum-Knowledge Attacker". In: *2015 13th Annual Conference on Privacy, Security and Trust (PST)*. July 2015, pp. 28–35. DOI: 10.1109/PST.2015.7232951 (cit. on p. 94).

[DS15] J. Domingo-Ferrer, J. Soria-Comas. "From T-Closeness to Differential Privacy and Vice Versa in Data Anonymization". In: *Knowledge-Based Systems* 74 (Jan. 2015), pp. 151–158. ISSN: 09507051. DOI: 10.1016/j.knosys.2014.11.011. (Visited on 07/24/2023) (cit. on p. 16).

[Dwo06] C. Dwork. "Differential Privacy". In: *Automata, Languages and Programming*. Springer, Berlin, Heidelberg, 2006, pp. 1–12. DOI: 10.1007/11787006_1. (Visited on 06/26/2023) (cit. on pp. 13, 17, 21).

[Dwo08] C. Dwork. "Differential Privacy: A Survey of Results". In: *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*. TAMC'08. Berlin, Heidelberg: Springer-Verlag, Apr. 2008, pp. 1–19. ISBN: 978-3-540-79227-7. (Visited on 04/19/2023) (cit. on p. 13).

[EAZS18] C. Eyupoglu, M. A. Aydin, A. H. Zaim, A. Sertbas. "An Efficient Big Data Anonymization Algorithm Based on Chaos and Perturbation Techniques". In: *Entropy* 20.5 (May 2018), p. 373. ISSN: 1099-4300. DOI: 10.3390/e20050373. (Visited on 04/19/2023) (cit. on p. 93).

[EC16] European Parliament, Council of the European Union. *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/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1–88, May 4, 2016. URL: https://data.europa.eu/eli/reg/2016/679/oj (visited on 08/02/2023) (cit. on p. 26).

[EDI+09] K. El Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, J. Bottomley. "A Globally Optimal K-Anonymity Method for the De-Identification of Health Data". In: *Journal of the American Medical Informatics Association* 16.5 (Sept. 2009), pp. 670–682. ISSN: 1067-5027, 1527-974X. DOI: 10.1197/jamia.M3144. (Visited on 08/28/2023) (cit. on p. 96).

[EGS03] A. Evfimievski, J. Gehrke, R. Srikant. "Limiting Privacy Breaches in Privacy Preserving Data Mining". In: *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '03. New York, NY, USA: Association for Computing Machinery, June 2003, pp. 211–222. ISBN: 978-1-58113-670-8. DOI: 10.1145/773153.773174. (Visited on 06/12/2023) (cit. on pp. 22, 94).

[EOL+22] E. Ekenstedt, L. Ong, Y. Liu, S. Johnson, P. L. Yeoh, J. Kliewer. "When Differential Privacy Implies Syntactic Privacy". In: *IEEE Transactions on Information Forensics and Security* 17 (2022), pp. 2110–2124. ISSN: 1556-6013, 1556-6021. DOI: 10.1109/TIFS.2022.3177953. (Visited on 07/11/2023) (cit. on pp. 16, 89).

[FHS+23] A. Fieschi, P. Hirmer, R. Sturm, M. Eisele, B. Mitschang. "Anonymization Use Cases for Data Transfer in the Automotive Domain". In: *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops)*. Mar. 2023, pp. 98–103. DOI: 10.1109/PerComWorkshops56833.2023.10150357 (cit. on pp. 13, 15, 25).

[FWFY10] B. C. Fung, K. Wang, A. W.-C. Fu, P. S. Yu. "Introduction to Privacy-Preserving Data Publishing". In: Chapman and Hall/CRC, Aug. 2010. ISBN: 978-1-4200-9150-2. DOI: 10.1201/9781420091502. (Visited on 07/05/2023) (cit. on p. 22).

[FWY05] B. Fung, K. Wang, P. Yu. "Top-down Specialization for Information and Privacy Preservation". In: *21st International Conference on Data Engineering (ICDE'05)*. Apr. 2005, pp. 205–216. DOI: 10.1109/ICDE.2005.143 (cit. on p. 92).

[Gar19] J. Gardner. *Gardn999 - Differential Privacy Synthetic Data Challenge Algorithm*. 2019. (Visited on 08/14/2023) (cit. on p. 97).

[GJK+21] S. Ghane, A. Jolfaei, L. Kulik, K. Ramamohanarao, D. Puthal. "Preserving Privacy in the Internet of Connected Vehicles". In: *IEEE Transactions on Intelligent Transportation Systems* 22.8 (Aug. 2021), pp. 5018–5027. ISSN: 1524-9050, 1558-0016. DOI: 10.1109/TITS.2020.2964410. (Visited on 11/08/2023) (cit. on p. 15).

[Goo23] Google. *Differential Privacy*. Google. Aug. 2023. (Visited on 08/14/2023) (cit. on p. 96).

[GT09a] A. Gionis, T. Tassa. "K-Anonymization with Minimal Loss of Information". In: *IEEE Transactions on Knowledge and Data Engineering* 21.2 (2009), pp. 206–219. DOI: 10.1109/TKDE.2008.129 (cit. on p. 85).

[GT09b] J. Goldberger, T. Tassa. “Efficient Anonymizations with Enhanced Utility”. In: *2009 IEEE International Conference on Data Mining Workshops* (Dec. 2009), pp. 106–113. DOI: 10.1109/ICDMW.2009.15. (Visited on 07/11/2023) (cit. on p. 93).

[guo14] guozhang. *Cornell Anonymization Toolkit*. https://sourceforge.net/projects/anony-toolkit/. July 2014. (Visited on 11/04/2023) (cit. on p. 95).

[HBML19] N. Holohan, S. Braghin, P. Mac Aonghusa, K. Levacher. *Diffprivlib: The IBM Differential Privacy Library*. July 2019. DOI: 10.48550/arXiv.1907.02444. arXiv: 1907.02444 [cs]. (Visited on 07/07/2023) (cit. on p. 44).

[HDF+12] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, P. D. Wolf. “Statistical Disclosure Control”. In: Sept. 2012. (Visited on 10/07/2023) (cit. on pp. 17, 22).

[IBM23] IBM. *Diffprivlib v0.6*. International Business Machines. Aug. 2023. (Visited on 08/08/2023) (cit. on p. 44).

[INS+19] R. Iyengar, J. P. Near, D. Song, O. Thakkar, A. Thakurta, L. Wang. “Towards Practical Differentially Private Convex Optimization”. In: *2019 IEEE Symposium on Security and Privacy (SP)* (May 2019), pp. 299–316. DOI: 10.1109/SP.2019.00001. (Visited on 08/14/2023) (cit. on p. 97).

[Iye02] V. S. Iyengar. “Transforming Data to Satisfy Privacy Constraints”. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’02. New York, NY, USA: Association for Computing Machinery, July 2002, pp. 279–288. ISBN: 978-1-58113-567-1. DOI: 10.1145/775047.775089. (Visited on 05/21/2023) (cit. on pp. 33, 34, 54, 91, 94, 95).

[JLM+15] S. Ji, W. Li, P. Mittal, X. Hu, R. Beyah. “{SecGraph}: A Uniform and Open-source Evaluation System for Graph Data Anonymization and De-anonymization”. In: *24th USENIX Security Symposium (USENIX Security 15)*. 2015, pp. 303–318. ISBN: 978-1-939133-11-3. (Visited on 06/28/2023) (cit. on p. 97).

[JMZ23] Jen Caltrider, Misha Rykov, Zoë MacDonald. **Privacy Not Included: A Buyer’s Guide for Connected Products*. https://foundation.mozilla.org/en/privacynotincluded/articles/its-official-cars-are-the-worst-product-category-we-have-ever-reviewed-for-privacy/. June 2023. (Visited on 11/11/2023) (cit. on p. 13).

[KG06] D. Kifer, J. Gehrke. “Injecting Utility into Anonymized Datasets”. In: *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’06. New York, NY, USA: Association for Computing Machinery, June 2006, pp. 217–228. ISBN: 978-1-59593-434-5. DOI: 10.1145/1142473.1142499. (Visited on 05/21/2023) (cit. on p. 86).

[Kif09] D. Kifer. “Attacks on Privacy and deFinetti’s Theorem”. In: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’09. New York, NY, USA: Association for Computing Machinery, June 2009, pp. 127–138. ISBN: 978-1-60558-551-2. DOI: 10.1145/1559845.1559861. (Visited on 08/23/2023) (cit. on p. 88).

[KMA+21] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. Nitin Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. El Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, S. Zhao. "Advances and Open Problems in Federated Learning". In: *Foundations and Trends® in Machine Learning* 14.1–2 (2021), pp. 1–210. ISSN: 1935-8237, 1935-8245. DOI: 10.1561/2200000083. (Visited on 08/17/2023) (cit. on p. 22).

[KPE+12] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper, K. A. Kuhn. "Flash: Efficient, Stable and Optimal K-Anonymity". In: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (Sept. 2012), pp. 708–717. DOI: 10.1109/SocialCom-PASSAT.2012.52. (Visited on 08/30/2023) (cit. on p. 44).

[KRG+08] D. J. Kelly, R. A. Raines, M. R. Grimaila, R. O. Baldwin, B. E. Mullins. "A Survey of State-of-the-Art in Anonymity Metrics". In: *Proceedings of the 1st ACM Workshop on Network Data Anonymization*. NDA '08. New York, NY, USA: Association for Computing Machinery, Oct. 2008, pp. 31–40. ISBN: 978-1-60558-301-3. DOI: 10.1145/1456441.1456453. (Visited on 05/04/2023) (cit. on pp. 15, 17).

[KYH+16] H. Kikuchi, T. Yamaguchi, K. Hamada, Y. Yamaoka, H. Oguri, J. Sakuma. "Ice and Fire: Quantifying the Risk of Re-identification and Utility in Data Anonymization". In: *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*. Mar. 2016, pp. 1035–1042. DOI: 10.1109/AINA.2016.151 (cit. on pp. 31, 35, 94).

[LDR05] K. LeFevre, D. J. DeWitt, R. Ramakrishnan. "Incognito: Efficient Full-Domain K-anonymity". In: *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*. Baltimore Maryland: ACM, June 2005, pp. 49–60. ISBN: 978-1-59593-060-6. DOI: 10.1145/1066157.1066164. (Visited on 04/24/2023) (cit. on pp. 44, 95).

[LDR06] K. LeFevre, D. DeWitt, R. Ramakrishnan. "Mondrian Multidimensional K-Anonymity". In: *22nd International Conference on Data Engineering (ICDE'06)*. Apr. 2006, pp. 25–25. DOI: 10.1109/ICDE.2006.101 (cit. on pp. 32, 33, 44, 90, 95, 96).

[LGM11] G. Loukides, A. Gkoulalas-Divanis, B. Malin. "COAT: COnstraint-based Anonymization of Transactions". In: *Knowledge and Information Systems* 28.2 (Aug. 2011), pp. 251–282. ISSN: 0219-3116. DOI: 10.1007/s10115-010-0354-4. (Visited on 06/28/2023) (cit. on p. 95).

[LHSM23] Y. Li, P. Hirmer, C. Stach, B. Mitschang. "Ensuring Situation-Aware Privacy for Connected Vehicles". In: *Proceedings of the 12th International Conference on the Internet of Things*. IoT '22. New York, NY, USA: Association for Computing Machinery, Jan. 2023, pp. 135–138. ISBN: 978-1-4503-9665-3. DOI: 10.1145/3567445.3569163. (Visited on 11/08/2023) (cit. on pp. 15, 25).

[Lil23] C. Lillelund. *Differentially Private Stochastic Gradient Descent*. July 2023. (Visited on 08/15/2023) (cit. on p. 97).

[LL09] T. Li, N. Li. “On the Tradeoff between Privacy and Utility in Data Publishing”. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’09. New York, NY, USA: Association for Computing Machinery, June 2009, pp. 517–526. ISBN: 978-1-60558-495-9. DOI: 10.1145/1557019.1557079. (Visited on 06/26/2023) (cit. on pp. 28, 89, 90).

[LLV07] N. Li, T. Li, S. Venkatasubramanian. “T-Closeness: Privacy Beyond k-Anonymity and l-Diversity”. In: *2007 IEEE 23rd International Conference on Data Engineering*. Apr. 2007, pp. 106–115. DOI: 10.1109/ICDE.2007.367856 (cit. on pp. 17, 20, 21, 35).

[LS07] G. Loukides, J. Shao. “Capturing Data Usefulness and Privacy Protection in K-anonymisation”. In: *Proceedings of the 2007 ACM Symposium on Applied Computing*. SAC ’07. New York, NY, USA: Association for Computing Machinery, Mar. 2007, pp. 370–374. ISBN: 978-1-59593-480-2. DOI: 10.1145/1244002.1244091. (Visited on 07/06/2023) (cit. on p. 34).

[LWFP06] J. Li, R. C.-W. Wong, A. W.-C. Fu, J. Pei. “Achieving K-Anonymity by Clustering in Attribute Hierarchical Structures”. In: *Data Warehousing and Knowledge Discovery*. Springer, Berlin, Heidelberg, 2006, pp. 405–416. DOI: 10.1007/11823728_39. (Visited on 07/06/2023) (cit. on p. 20).

[LWHC10] J.-L. Lin, T.-H. Wen, J.-C. Hsieh, P.-C. Chang. “Density-Based Microaggregation for Statistical Disclosure Control”. In: *Expert Systems with Applications* 37.4 (Apr. 2010), pp. 3256–3263. ISSN: 09574174. DOI: 10.1016/j.eswa.2009.09.054. (Visited on 06/19/2023) (cit. on p. 84).

[LZW19] N. Li, Z. Zhang, T. Wang. *DPSyn*. May 2019. (Visited on 08/14/2023) (cit. on p. 97).

[Maz23] L. Mazzone. *Crowds*. July 2023. (Visited on 08/28/2023) (cit. on p. 96).

[McK] R. McKenna. *Rmckenna - Differential Privacy Synthetic Data Challenge Algorithm*. (Visited on 08/14/2023) (cit. on pp. 96, 97).

[MGKV06] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam. “L-Diversity: Privacy beyond k-Anonymity”. In: *22nd International Conference on Data Engineering (ICDE’06)*. Apr. 2006, pp. 24–24. DOI: 10.1109/ICDE.2006.1 (cit. on pp. 20, 32, 85).

[ML21] A. Majeed, S. Lee. “Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey”. In: *IEEE Access* 9 (2021), pp. 8512–8545. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3045700 (cit. on pp. 15, 17, 27).

[Moh23] P. Mohan. *Prashmohan/GUPT*. July 2023. (Visited on 08/14/2023) (cit. on p. 96).

[MW04] A. Meyerson, R. Williams. “On the Complexity of Optimal K-anonymity”. In: *Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. Paris France: ACM, June 2004, pp. 223–228. ISBN: 978-1-58113-858-0. DOI: 10.1145/1055558.1055591. (Visited on 04/24/2023) (cit. on p. 94).

[NAC07] M. E. Nergiz, M. Atzori, C. Clifton. “Hiding the Presence of Individuals from Shared Databases”. In: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’07. New York, NY, USA: Association for Computing Machinery, June 2007, pp. 665–676. ISBN: 978-1-59593-686-8. DOI: 10.1145/1247480.1247554. (Visited on 06/19/2023) (cit. on pp. 33, 34, 93, 94).

[NC06] M. Nergiz, C. Clifton. “Thoughts on K-Anonymization”. In: *22nd International Conference on Data Engineering Workshops (ICDEW’06)*. Apr. 2006, pp. 96–96. DOI: 10.1109/ICDEW.2006.147 (cit. on pp. 33, 90).

[NCN09] M. E. Nergiz, C. Clifton, A. E. Nergiz. “Multirelational K-Anonymity”. In: *IEEE Transactions on Knowledge and Data Engineering* 21.8 (Aug. 2009), pp. 1104–1117. ISSN: 1558-2191. DOI: 10.1109/TKDE.2008.210 (cit. on p. 94).

[PAJJ20] Peter-Paul de Wolf, Anco Hundepool, Juan-José Salazar, Jordi Castro. *τ-ARGUS*. https://research.cbs.nl/casc/tau.htm. 2020. (Visited on 08/08/2023) (cit. on p. 95).

[PES+20] F. Prasser, J. Eicher, H. Spengler, R. Bild, K. A. Kuhn. “Flexible Data Anonymization Using ARX—Current Status and Challenges Ahead”. In: *Software: Practice and Experience* 50.7 (July 2020), pp. 1277–1304. ISSN: 0038-0644, 1097-024X. DOI: 10.1002/spe.2812. (Visited on 07/28/2023) (cit. on p. 44).

[PGL+13] G. Poulis, A. Gkoulalas-Divanis, G. Loukides, S. Skiadopoulos, C. Tryfonopoulos. *The SECRETA System*. https://users.uop.gr/~poulis/SECRETA/index.html. 2013. (Visited on 08/08/2023) (cit. on p. 95).

[PGL+15] G. Poulis, A. Gkoulalas-Divanis, G. Loukides, S. Skiadopoulos, C. Tryfonopoulos. “SECRETA: A Tool for Anonymizing Relational, Transaction and RT-Datasets”. In: *Medical Data Privacy Handbook*. Ed. by A. Gkoulalas-Divanis, G. Loukides. Cham: Springer International Publishing, 2015, pp. 83–109. ISBN: 978-3-319-23633-9. DOI: 10.1007/978-3-319-23633-9_5. (Visited on 06/27/2023) (cit. on pp. 34, 95).

[Pha23] M.-K. Pham. *Data Anonymization Using K-Anonymity*. July 2023. (Visited on 08/08/2023) (cit. on p. 96).

[PK] F. Prasser, F. Kohlmayer. *ARX - Data Anonymization Tool | A Comprehensive Software for Privacy-Preserving Microdata Publishing*. (Visited on 08/08/2023) (cit. on p. 44).

[PK01] A. Pfitzmann, M. Köhntopp. “Anonymity, Unobservability, and Pseudonymity — A Proposal for Terminology”. In: *Designing Privacy Enhancing Technologies*. Ed. by G. Goos, J. Hartmanis, J. Van Leeuwen, H. Federrath. Vol. 2009. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 1–9. ISBN: 978-3-540-41724-8 978-3-540-44702-3. DOI: 10.1007/3-540-44702-4_1. (Visited on 07/28/2023) (cit. on p. 17).

[PK15] F. Prasser, F. Kohlmayer. “Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool”. In: *Medical Data Privacy Handbook*. Ed. by A. Gkoulalas-Divanis, G. Loukides. Cham: Springer International Publishing, 2015, pp. 111–148. ISBN: 978-3-319-23633-9. DOI: 10.1007/978-3-319-23633-9_6. (Visited on 08/21/2023) (cit. on p. 44).

[PMG+18] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, Y. Kim. “Data Synthesis Based on Generative Adversarial Networks”. In: *Proceedings of the VLDB Endowment* 11.10 (June 2018), pp. 1071–1083. ISSN: 2150-8097. DOI: 10.14778/3231751.3231757. (Visited on 11/10/2023) (cit. on p. 23).

[Pra23] N. Prabhu. *K-Anonymity*. Aug. 2023. (Visited on 08/31/2023) (cit. on p. 96).

[PS07] H. Park, K. Shim. “Approximate Algorithms for K-anonymity”. In: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’07. New York, NY, USA: Association for Computing Machinery, June 2007, pp. 67–78. ISBN: 978-1-59593-686-8. DOI: 10.1145/1247480.1247490. (Visited on 06/20/2023) (cit. on p. 94).

[rea23] realrolfje. *GDPR Compliant Testing | Anonimatron*. https://realrolfje.github.io/anonimatron/. 2023. (Visited on 08/08/2023) (cit. on p. 97).

[RHS07] V. Rastogi, S.-K. Hong, D. Suciu. “The Boundary Between Privacy and Utility in Data Publishing”. In: *Very Large Data Bases Conference*. Sept. 2007. (Visited on 06/26/2023) (cit. on p. 90).

[Sam01] P. Samarati. “Protecting Respondents Identities in Microdata Release”. In: *IEEE Transactions on Knowledge and Data Engineering* 13.6 (Nov. 2001), pp. 1010–1027. ISSN: 1558-2191. DOI: 10.1109/69.971193 (cit. on pp. 19, 93, 94).

[SC22] D. Sadhya, B. Chakraborty. “Quantifying the Effects of Anonymization Techniques Over Micro-Databases”. In: *IEEE Transactions on Emerging Topics in Computing* 10.4 (Oct. 2022), pp. 1979–1992. ISSN: 2168-6750. DOI: 10.1109/TETC.2022.3141754 (cit. on pp. 16, 94).

[SDSM14] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, S. Martínez. “Enhancing Data Utility in Differential Privacy via Microaggregation-Based $$k$$-Anonymity”. In: *The VLDB Journal* 23.5 (Oct. 2014), pp. 771–794. ISSN: 0949-877X. DOI: 10.1007/s00778-014-0351-4. (Visited on 05/11/2023) (cit. on pp. 16, 20).

[SDSM15] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, S. Martínez. “T-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation”. In: *IEEE Transactions on Knowledge and Data Engineering* 27.11 (Nov. 2015), pp. 3098–3110. ISSN: 1558-2191. DOI: 10.1109/TKDE.2015.2435777 (cit. on p. 35).

[SLZ20] J. Shan, Y. Lin, X. Zhu. “A New Range Noise Perturbation Method Based on Privacy Preserving Data Mining”. In: *2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS)*. Mar. 2020, pp. 131–136. DOI: 10.1109/ICAIIS49377.2020.9194850 (cit. on p. 92).

[SMS21] P. Silva, E. Monteiro, P. Simões. “Privacy in the Cloud: A Survey of Existing Solutions and Research Challenges”. In: *IEEE Access* 9 (2021), pp. 10473–10497. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3049599 (cit. on pp. 15, 95).

[SS98] P. Samarati, L. Sweeney. “Protecting Privacy When Disclosing Information: K-Anonymity and Its Enforcement through Generalization and Suppression”. In: (Aug. 1998) (cit. on pp. 19, 29, 31, 45, 93).

[SSSS17] R. Shokri, M. Stronati, C. Song, V. Shmatikov. “Membership Inference Attacks Against Machine Learning Models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. May 2017, pp. 3–18. DOI: 10.1109/SP.2017.41 (cit. on p. 22).

[Sta15] K. Stark. *Open Anonymizer*. May 2015. (Visited on 08/15/2023) (cit. on p. 97).

[sun23] sunblaze-ucb. *Differentially Private Convex Optimization Benchmark*. sunblaze-ucb. July 2023. (Visited on 08/14/2023) (cit. on p. 97).

[Swe00] L. Sweeney. “Simple Demographics Often Identify People Uniquely”. In: . *Pittsburgh* (2000) (cit. on p. 13).

[Swe02a] L. Sweeney. "ACHIEVING K-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (Oct. 2002), pp. 571–588. ISSN: 0218-4885, 1793-6411. DOI: 10.1142/S021848850200165X. (Visited on 05/10/2023) (cit. on p. 34).

[Swe02b] L. Sweeney. "K-Anonymity: A Model for Protecting Privacy". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.5 (Oct. 2002), pp. 557–570. ISSN: 0218-4885. DOI: 10.1142/S0218488502001648. (Visited on 04/17/2023) (cit. on pp. 13, 18, 31, 45).

[Tem17] M. Templ. *Statistical Disclosure Control for Microdata*. Cham: Springer International Publishing, 2017. ISBN: 978-3-319-50270-0 978-3-319-50272-4. DOI: 10.1007/978-3-319-50272-4. (Visited on 08/08/2023) (cit. on pp. 17, 97).

[TMK+23] M. Templ, B. Meindl, A. Kowarik, J. Gussenbauer, O. F. E. C.-O. code for rank swapping, D. (published c(++) code (under LGPL), mdav-microaggregation, s. a. o. (risk measures), S. N. (c. c. (E. v1.1)), P. H. (m. t. c. c. (LGPL)). *sdcMicro: Statistical Disclosure Control Methods for Anonymization of Data and Risk Estimation*. Jan. 2023. (Visited on 08/08/2023) (cit. on p. 97).

[UKLU20] University of Illinois at Springfield, USA, H. Kartal, X.-B. Li, University of Massachusetts Lowell, USA. "Protecting Privacy When Sharing and Releasing Data with Multiple Records per Person". In: *Journal of the Association for Information Systems* 21 (Nov. 2020), pp. 1461–1485. ISSN: 15369323. DOI: 10.17705/1jais.00643. (Visited on 11/01/2023) (cit. on pp. 26, 37, 38).

[USD23] USDOT Office of the Secretary of Transportation. *Privacy Protection Application (PPA)*. usdot-its-jpo-data-portal. Feb. 2023. (Visited on 08/14/2023) (cit. on pp. 34, 97).

[VFLS23] S. Vimercati, S. Foresti, G. Livraga, P. Samarati. "K-Anonymity: From Theory to Applications". In: *Trans. Data Priv.* (2023). (Visited on 07/11/2023) (cit. on p. 20).

[WBWA19] Widodo, E. K. Budiardjo, W. C. Wibowo, H. T. Achsan. "An Approach for Distributing Sensitive Values in K-Anonymity". In: *2019 International Workshop on Big Data and Information Security (IWBIS)*. Oct. 2019, pp. 109–114. DOI: 10.1109/IWBIS.2019.8935849 (cit. on p. 85).

[WD01] L. Willenborg, T. De Waal. "Elements of Statistical Disclosure Control". In: ed. by P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg, I. Olkin, N. Wermuth, S. Zeger. Vol. 155. Lecture Notes in Statistics. New York, NY: Springer New York, 2001. ISBN: 978-0-387-95121-8 978-1-4613-0121-9. DOI: 10.1007/978-1-4613-0121-9. (Visited on 06/14/2023) (cit. on p. 17).

[WFY05] K. Wang, B. Fung, P. Yu. "Template-Based Privacy Preservation in Classification Problems". In: *Fifth IEEE International Conference on Data Mining (ICDM'05)*. Nov. 2005, 8 pp.-. DOI: 10.1109/ICDM.2005.142 (cit. on p. 94).

[WVX+15] Z. Wan, Y. Vorobeychik, W. Xia, E. W. Clayton, M. Kantarcioglu, R. Ganta, R. Heatherly, B. A. Malin. "A Game Theoretic Framework for Analyzing Re-Identification Risk". In: *PLOS ONE* 10.3 (Mar. 2015). Ed. by C.-Y. Xia, e0120592. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0120592. (Visited on 07/18/2023) (cit. on p. 85).

[WWH16] Y. Wang, X. Wu, D. Hu. "Using Randomized Response for Differential Privacy Preserving Data Collection". In: *EDBT/ICDT Workshops*. 2016. (Visited on 07/24/2023) (cit. on p. 21).

[XT06] X. Xiao, Y. Tao. "Anatomy: Simple and Effective Privacy Preservation". In: *Proceedings of the 32nd International Conference on Very Large Data Bases*. VLDB '06. Seoul, Korea: VLDB Endowment, Sept. 2006, pp. 139–150. (Visited on 06/20/2023) (cit. on pp. 90, 94).

[XT07] X. Xiao, Y. Tao. "M-Invariance: Towards Privacy Preserving Re-Publication of Dynamic Datasets". In: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. SIGMOD '07. New York, NY, USA: Association for Computing Machinery, June 2007, pp. 689–700. ISBN: 978-1-59593-686-8. DOI: 10.1145/1247480.1247556. (Visited on 05/09/2023) (cit. on p. 27).

[XWP+06] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A. W.-C. Fu. "Utility-Based Anonymization for Privacy Preservation with Less Information Loss". In: *ACM SIGKDD Explorations Newsletter* 8.2 (Dec. 2006), pp. 21–30. ISSN: 1931-0145. DOI: 10.1145/1233321.1233324. (Visited on 05/21/2023) (cit. on pp. 33–35, 95, 96).

[XXY10] Y. Xiao, L. Xiong, C. Yuan. "Differentially Private Data Release through Multidimensional Partitioning". In: *Secure Data Management*. Ed. by D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, W. Jonker, M. Petković. Vol. 6358. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 150–168. ISBN: 978-3-642-15545-1 978-3-642-15546-8. DOI: 10.1007/978-3-642-15546-8_11. (Visited on 08/23/2023) (cit. on p. 21).

[YQ10] W. Yang, S. Qiao. "A Novel Anonymization Algorithm: Privacy Protection and Knowledge Preservation". In: *Expert Systems with Applications* 37.1 (Jan. 2010), pp. 756–766. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2009.05.097. (Visited on 06/30/2023) (cit. on p. 93).

[ZKSY07] Q. Zhang, N. Koudas, D. Srivastava, T. Yu. "Aggregate Query Answering on Anonymized Tables". In: *2007 IEEE 23rd International Conference on Data Engineering*. Apr. 2007, pp. 116–125. DOI: 10.1109/ICDE.2007.367857 (cit. on p. 90).

[ZLZY17] T. Zhu, G. Li, W. Zhou, P. S. Yu. "Differentially Private Data Publishing and Analysis: A Survey". In: *IEEE Transactions on Knowledge and Data Engineering* 29.8 (Aug. 2017), pp. 1619–1638. ISSN: 1041-4347. DOI: 10.1109/TKDE.2017.2697856. (Visited on 08/28/2023) (cit. on pp. 21, 27).

[ZMK+18] D. Zhang, R. McKenna, I. Kotsogiannis, M. Hay, A. Machanavajjhala, G. Miklau. "EKTELO: A Framework for Defining Differentially-Private Computations". In: *Proceedings of the 2018 International Conference on Management of Data*. SIGMOD '18. New York, NY, USA: Association for Computing Machinery, May 2018, pp. 115–130. ISBN: 978-1-4503-4703-7. DOI: 10.1145/3183713.3196921. (Visited on 08/14/2023) (cit. on p. 96).

A Appendix

A.1 Pool of Metrics

This section lists all metrics encountered during the making of this work. It includes the metrics that were selected and implemented for the experiments and those that were considered but not selected. The metrics are grouped by their types for a better overview.

A.1.1 Syntactic Metrics

Suppression Ratio

See Section 4.2.2.

k-anonymity

See Section 4.2.3.

l-diversity

See Section 4.2.4.

Average Equivalence Class Size

See Section 4.2.5.

Discernibility Penalty

See Section 4.2.6.

A.1.2 Metrics on Attribute Value Distance

In-Data Precision Loss

See Section 4.2.7.

Cross-Data Precision Loss

See Section 4.2.8.

Sum of Squares

This metric is common to measure homogeneity in clustering approaches and is used in the literature to evaluate anonymization using microaggregation [DM02; LWHC10]. Let T be an input table grouped into g groups where group i contains n_i entries. With t_{ij}, we denote the j-th vector in the i-th group. The normalized information loss induced by microaggregation is the within-groups sum of squares loss SSE divided by the total sum of squares loss SST.

$$M_{SS} = \frac{SSE}{SST}$$
$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (t_{ij} - \bar{t}_i)^2$$
$$SST = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (t_{ij} - \bar{t})^2$$

Here, $\bar{t}_i$ denotes the average tuple in the group i, and $\bar{t}$ denotes the average tuple over the entire table. Note that SST is constant for a fixed table regardless of the groupings chosen. Although the authors define the metric using the average as the aggregation function, it can be extended to work for different aggregation functions. To this end, $\bar{t}_i$ denotes the output of the aggregation function over group i, and $\bar{t}$ denotes the output of the aggregation function over the entire table.

A.1.3 Metrics on Attribute Value Distributions

Earth Mover's Distance

See Section 4.2.9.

g-balance

See Section 4.2.10.

h-affiliation

See Section 4.2.11.

Non-Uniform Entropy

The non-uniform entropy measures information loss due to generalization [GT09a]. Given a table $T = \{t_1, \ldots, t_n\}$ and its generalization $T' = \{t'_1, \ldots, t'_n\}$, the loss of information according to the non-uniform entropy over attributes $A = \{A_1, \ldots, A_m\}$ is:

$$M_{NUE}(T, T') = \sum_{i=1}^{n} \sum_{j=1}^{m} -\log P(t_i[A_j] \mid t'_i[A_j])$$

For the value v of attribute A_j and its generalization v', the probability $P(v|v')$ is:

$$P(v|v') = \frac{|\{t \in T \mid t[A_j] = v\}|}{|\{t \in T \mid t[A_j] \in v'\}|}$$

The metric is also used by Bild et al. [BKP18].

Entropy

One basic measure of the distribution of sensitive values is Entropy. Simply put, Entropy measures the amount of information inherent in a distribution. If the distribution heavily favors one specific outcome, the information content is not as high compared to a scenario where all outcomes are equally likely. The entropy over a distribution $\mathcal{X}$ is defined as:

$$M_H(P) = -\sum_{x \in \mathcal{X}} p(x) \cdot \log p(x)$$

Entropy is a basic way to measure the diversity of sensitive values in an EC. The value for an entire data set could be the minimum value over all ECs, the average value, or another kind of value aggregation over ECs. The higher the entropy, the higher the diversity of sensitive values. Therefore, the harder it is for an adversary to assign a value to an individual in an EC with high probability. A distribution that always returns the same value when sampled will have an Entropy of 0, the lowest possible value. Entropy in this form is only applied to the anonymized table, so there is no way to capture the change in the entropy value due to the anonymization process. Works in the literature usually assume that an adversary knows the distribution of sensitive values over the original table and, therefore, also its Entropy [BS08]. Because of this, an adversary could gain information from lower Entropy values in the anonymized table compared to the Entropy over all sensitive values, while the data holder that releases the anonymized data stays oblivious to the disclosure of information. Of course, a simple comparison between the Entropy before and after anonymization will reveal the information gained by an adversary to the data holder. Entropy is employed throughout the literature, ranging from early works like the introduction of l-diversity by Machanavajjhala et al. [MGKV06] to more recent ones [WBWA19].

Information Loss based on Entropy

Introduced by Wan et al. [WVX+15], this metric measures information loss using Entropy over the attribute values. The metric is applied to the generalized values of the quasi-identifier attributes in the paper. Given a tuple t and a generalized entry $t[A_i]$ under attribute A_i, the authors assume that

the original value is uniformly distributed among the values covered by $t[A_i]$. They define the loss on the anonymized table to be:

$$M_{ILE}(T) = \sum_{t \in T} \sum_{i=1}^{m} -\log \frac{1}{range(t[A_i])}$$

Here, $range(t[A_i])$ returns the number of distinct values covered by the tuple t under attribute A_i in the transformed table. The authors also give a way to normalize the value by dividing it by its maximum value $ILE_{max}(T)$.

$$ILE_{max}(T) = \sum_{t \in T} \sum_{i=1}^{m} -\log \frac{1}{range(A_i)}$$

Here, $range(A_i)$ returns the total number of distinct values of attribute A_i.

KL-divergence

Kifer and Gehrke [KG06] use the Kullback-Leibler divergence (KL-divergence) between the original data and the anonymized data to measure the loss of information due to anonymization. Their approach involves releasing anonymized marginals, which are summaries of the original table providing frequency counts for fixed attribute combinations. The authors compute a maximum entropy distribution from the marginals because the two distributions that are compared with the KL-divergence need to be defined on the same sample space. The KL-divergence between two distributions P and Q defined over the same sample space $\mathcal{X}$ is

$$M_{KL}(P, Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$

A higher cost means that the two distributions are more different, which means that information from the original data is lost. There are many possibilities to apply this metric to measure information loss. Almasi et al. [ASMH16] use KL-divergence to measure the difference between the distribution of sensitive values in an EC in the sanitized dataset and the total distribution in the original dataset. The metric can also be seen as one for privacy: the higher the value, the more different the anonymized data and, therefore, safer.

L_1-norm

The L_1-norm provides another measurement to quantify the difference between two distributions. Coull et al. [CWK+08] use it to quantify the anonymity of network data, while Agrawal and Aggarwal [AA01] use it as a measure of information loss in an adjusted form. They assume anonymization by perturbation and define a metric that measures the difference between the distribution of values in the original data and the estimated distribution of the original values by a reconstruction algorithm. The reconstructed distribution is the output of an Expectation Maximization algorithm applied to the perturbed dataset. We consider the simple definition of the L_1-norm between two distributions X and Y over domains $\mathcal{X}$ and $\mathcal{Y}$, respectively.

$$M_{L_1}(X, Y) = \sum_{z \in \mathcal{X} \cup \mathcal{Y}} |P(X = z) - P(Y = z)|$$

The values range from 0, which indicates identical distributions, to 1, indicating completely disjoint distribution spaces. Whether the measure quantifies anonymity or information loss depends on the distributions to which it is applied. One could apply the metric to the distributions of sensitive values to measure how they change due to anonymization.

δ-disclosure privacy

In addition to the total value of the adversarial knowledge gain over the entire table, Brickell and Shmatikov [BS08] propose the notion of δ-disclosure privacy. The concept is similar to t-closeness in that there is a threshold δ on the difference between the sensitive value distribution in Equivalence Classes vs. the entire table. The authors define that a table is δ-disclosure private if for every equivalence class e and sensitive value $v \in S$ we have:

$$\left| \log \frac{p(e,v)}{p(T,v)} \right| < \delta$$

Intuitively, this privacy definition places an upper bound on the adversarial knowledge gain. The value of δ can be used as a metric value, where lower values of δ indicate better privacy. In Section 4.2.9, it was noted that the relationship between the parameter t in the t-closeness concept, which is calculated using EMD, and information gain is unclear. This was shown using two pairs of distributions assigned an equal distance by EMD but intuitively showed an unequal information gain. If we want to find bounds δ for these distributions, we find that for the first pair, $(0.01, 0.99)$ and $(0.11, 0.89)$, we have $|\log \frac{0.01}{0.11}| \approx 1.04$ and $|\log \frac{0.99}{0.89}| \approx 0.05$. For the second pair, $(0.4, 0.6)$ and $(0.5, 0.5)$, we have $|\log \frac{0.4}{0.5}| \approx 0.1$ and $|\log \frac{0.6}{0.5}| \approx 0.08$. The second pair easily satisfies 0.2-disclosure privacy, while the first pair is far from it due to the relatively large information gained on the first sensitive value. We can, therefore, conclude that the parameter δ captures the notion of information gain better than the parameter t in t-closeness. Another difference between the two parameters is how missing sensitive values in ECs are evaluated. The authors do not specify how to calculate δ for missing sensitive values in the ECs. A missing sensitive value in an EC would require $|\log \frac{0}{p(T,v)}|$ to be calculated, but because $\log 0$ is undefined, there can be no possible bound δ set on the value. Missing sensitive values are a significant information gain for an adversary, so they should not be ignored and are therefore not allowed in the context of δ-disclosure privacy. The EMD does allow missing sensitive values. They simply increase the value of t. We omit this metric from our experiments, as the number of sensitive values in the data is very high, which results in many ECs in the anonymized data not containing every sensitive value.

A.1.4 Metrics that Consider Adversarial Knowledge

Adversarial Prediction of Sensitive Values

See Section 4.2.12.

Adversarial Accuracy Gain

Many works in the literature model an adversary that uses the quasi-identifier values to predict the sensitive value of an individual. The guessing strategy of an adversary can be implemented in several ways. A naïve Bayes classifier is one option that can be found in the literature [CPE+13; Kif09]. A simpler way is guessing the sensitive value corresponding to the majority in the respective EC, which is what Brickell and Shmatikov [BS08] use to measure the privacy that an anonymized table provides. Let $s_{max}(\mathcal{T})$ be the most common sensitive value in a given set of tuples $\mathcal{T}$, and let $p(\mathcal{T}, v)$ return the fraction of tuples in $\mathcal{T}$ that have sensitive value v.

$$M_{A\_pred}(T', T^*) = \left(\frac{1}{|T'|} \sum_{e \in EC(T')} |e| \cdot p(e, s_{max}(e))\right) - p(T^*, s_{max}(T^*))$$

The metric measures the difference between the adversary's base prediction accuracy and the adversary's accuracy on the anonymized dataset. The dataset T^* is the table T with all quasi-identifier values suppressed, creating one EC containing all tuples in the table. Therefore, an attacker's base accuracy on guessing the correct sensitive value is simply the probability of the most common sensitive value in the original data. The authors argue that this base knowledge will always be available to the attacker, even if all quasi-identifier values are suppressed. A value of 0 indicates that the adversary learns nothing when observing the anonymized table compared to his base knowledge. Higher values indicate distributions of the sensitive values in ECs that deviate from the overall distribution. The authors note that the metric underestimates the amount of information leaked by the anonymized data because the metric ignores shifts in the distribution of sensitive values that are not the majority.

Min-Entropy Leakage

Alvim et al. [AAC+12] propose the Min-Entropy Leakage as a measure of the information that an attacker can learn about a dataset by observing the output of a query. We consider a differentially private function from $\mathcal{X}$ to $\mathcal{Z}$, where $\mathcal{X}$ is the set of all databases, and $\mathcal{Z}$ is the image of the function. The input and output of the function are modeled by the probability distributions X and Z, respectively.

$$\begin{aligned} M_{MEL}(X, Z) &= H_\infty(X) - H_\infty(X|Z) \\ H_\infty(X) &= -\log_2 \max_{x \in X} p(x) \\ H_\infty(X|Z) &= -\log_2 \sum_{z \in Z} p(z) \max_{x \in X} p(x|z) \end{aligned}$$

The authors impose a bound on the leakage for an algorithm that provides ϵ-differential privacy. The algorithm operates on a dataset that contains u individuals that participate in the dataset. Furthermore, they fix the number of sensitive attribute values of each individual to be v, where $v \geq 2$.

$$M_{MEL}(X, Z) \leq u \log_2 \frac{ve^\epsilon}{v - 1 + e^\epsilon}$$

The right side of the inequality becomes 0 if $\epsilon = 0$. It is also shown in the paper that the bound is tight.

Global Utility based on information gain

Introduced by Alvim et al. [AAC+12] and used by Ekenstedt et al. [EOL+22]. Consider a privacy mechanism $p_{Z|Y}(\cdot|\cdot)$ that is applied to a random variable (our original dataset) Y to create the random variable (our transformed dataset) Z. We assume a user observes Z and tries to make a guess $\hat{Y}$ that is as close as possible to the original data Y.

$$M_{GU}(Y, Z) = \sum_{y} p_y(y) \sum_{\hat{y}} p(\hat{y}|y) g(y, \hat{y})$$

where $g(y, \hat{y})$ is a gain function that measures how useful it is to guess $\hat{y}$ when the actual value is y. $p(y)$ is the prior probability of real answer y, and $p(\hat{y}|y)$ is the probability of the user guessing $\hat{y}$ when the real answer is y. There is a special case where the gain function is binary, meaning all guesses are useless except the correct one. The formula then simplifies to:

$$M_{GU}(Y, Z) = \sum_{z} \max_{y} p(y, z)$$

In [EOL+22], they furthermore define the metric for their use-case, which is the ϵ-LDP mechanism. Alvim et al. [AAC+12] also derive a bound on the utility of a privacy mechanism that satisfies ϵ-differential privacy.

Worst-Case Privacy Loss

Li and Li [LL09] propose a metric that measures privacy loss in an anonymized table by comparing an adversary's prior knowledge about the sensitive attribute of a tuple to the adversary's posterior knowledge. Let Q be the distribution of the sensitive attributes of the overall population in the entire table. This distribution Q is the adversary's prior knowledge of the data because it will always be available after the anonymized table is published, even if all quasi-identifier attributes are suppressed. Of course, the values of the sensitive attributes must not be changed during anonymization. Observing the anonymized table, the adversary receives an update on their knowledge about the distribution of sensitive values for tuple t', which is now reduced to the EC that contains t'. This posterior knowledge will be denoted with $P(t')$ and describes the distribution of sensitive values in the EC that contains t'. The worst-case privacy loss is then given by

$$\begin{aligned}
M_{P_{loss}} &= \max_{t' \in T'} P_{loss}(t') \quad \text{where} \\
P_{loss}(t') &= \frac{1}{2}[KL(Q, M) + KL(P(t'), M)] \quad \text{and} \\
M &= \frac{1}{2}(Q + P(t')) \quad \text{and} \\
KL(Q, P) &= \sum_{i} q_i \log \frac{q_i}{p_i} \quad \text{is the Kullback-Leibler divergence}
\end{aligned}$$

As the KL-divergence measures the difference between two probability distributions, a higher value means a bigger disagreement between the distributions and therefore higher privacy loss. The lowest value is 0 and occurs when the published table has all the quasi-identifier attributes suppressed such that there is only one EC, which is the entire table and therefore $Q = P(t)$ for any tuple t.

A.1.5 Metrics of Other Types

The metrics listed in this section do not fit any of the types that we set, so we collect them under this title.

Ambiguity Metric

Nergiz and Clifton [NC06] use a metric that, for each tuple in the generalized table, sums up the number of theoretically possible combinations of input tuples it can represent. The penalty for a transformed table T' is calculated as follows.

$$M_{AM}(T') = \frac{1}{|T'|} \sum_{t' \in T'} \prod_{i=1}^{m} range(t'[A_i]),$$

where $range(t'[A_i])$ returns the number of distinct values covered by the value of tuple t' under attribute A_i. For example, consider the generalized attribute values of the attribute age. Then $range([3,5]) = 3$, as the interval covers three age values.

Evaluation of Query-answering accuracy

Many anonymization approaches in the literature are evaluated by running aggregation queries on the anonymized data and evaluating the accuracy of the results [CPE+13; LDR06; LL09; RHS07; XT06; ZKSY07]. These works usually do not introduce this evaluation as a specific metric but use it to assess the data quality that their proposed anonymization algorithms provide in experiments, therefore using different approaches from paper to paper. The aggregate queries often follow a schema like "SELECT <operator> FROM <table> WHERE <pred(A_1)> AND ... <pred(A_m)>", where "operator" is an aggregation like SUM, COUNT, AVG, etc., and "pred(A_i)" is a predicate on attribute A_i, for example "Age > 30". There are different ways of evaluating the accuracy of the query results. One of the most popular ones is reporting the relative error between running the query on the original data and on the anonymized data. LeFevre et al. [LDR06] report the mean and standard deviation of the absolute error over their set of counting queries. The relative error of a query q, the original table T and the anonymized table T' is:

$$M_{E_{rel}}(T, T', q) = \frac{|q(T) - q(T')|}{q(T)}$$

Cormode et al. [CPE+13] run a set of `COUNT` queries on the anonymized data and measure the relative error compared to running the queries on the original data. The authors report the median relative error for a given query workload in the paper. The reciprocal of this value then corresponds to the empirical utility of the anonymized dataset.

The accuracy of aggregate queries can be a useful measure to evaluate an intuitive notion of the utility of anonymized data. In combination with a comparison of the accuracy of the queries on the original data, it is a suitable metric to measure information loss due to the anonymization process. In the literature, many works use this metric, but there is no fixed definition of what these aggregate queries look like and how the accuracies of the anonymized and original data are compared. The query performances allow useful comparisons between anonymizations of different tables if the values are normalized by using the relative error, for example.

Classification Metric

Iyengar [Iye02] introduce a usage-based metric, which evaluates the anonymization by assuming that the data are used for a classification task. The Classification Metric supposes that one of the attributes in the anonymized table is a class label that is the target of some classification model. Since k-anonymity requires that the records in one EC are indistinguishable from each other with regards to the quasi-identifier, a classification model would not be able to discriminate them if there were more than one class label in the group. Therefore, the metric penalizes ECs that contain rows with different labels. Note that the data may have non-identifying attributes that could have predictive capabilities. The author leaves the consideration of these attributes in the transformation process to future work. Given the anonymized table T', the function $class(t)$ that returns the class label of tuple t, and function $majority$, which outputs the majority class label of the EC that is given as argument, the penalty is calculated as follows:

$$M_{CM} = \frac{1}{|T'|} \sum_{t' \in T'} penalty(t'), \quad \text{where}$$

$$penalty(t') = \begin{cases} 1 & \text{if } t' \text{ is suppressed} \\ 1 & \text{if } class(t') \neq majority(EC(t')) \\ 0 & \text{otherwise} \end{cases}$$

The Classification Metric penalizes ECs that contain multiple class labels. It is also stated that the metric can be extended to incorporate different costs for the $penalty$ function. This can, for example, be achieved by defining a cost matrix that specifies the cost of misclassifying a row from its true class to the majority class of its EC. The loss is normalized over the entire table, which makes this metric suitable for a comparison of anonymizations of different tables. Bayardo and Agrawal [BA05] use the metric to evaluate the performance of their k-optimize procedure. The author notes that using this metric in combination with k-anonymity may dramatically increase the risk of attribute disclosure through homogeneity attacks. Imagine choosing a sensitive attribute to be the class label for this metric, then "a good" transformation of the table would contain ECs whose sensitive values are mostly homogenous, making the data prone to homogeneity attacks. It is, therefore, advisable not to use a sensitive attribute as the class label or use a more advanced privacy requirement like l-diversity, which prevents homogeneity attacks.

Classifier Accuracy

Due to the increasing interest in using Machine Learning models to extract meaningful information from data, training classifiers is a huge part of (big) data analysis. Comparing the accuracy that a classifier achieves on the anonymized data vs. what the classifier achieves on the original data is an intuitive way of measuring one aspect of information loss due to anonymization. The specific classification algorithm used for the evaluation is not set in stone, but Support Vector Machines are among the most used classifiers in the literature. Shan et al. [SLZ20] use Bayes, Support Vector Machines, and Decision Trees as classifiers to assess the data utility. Fung et al. [FWY05] use the C4.5 classifier and the Naive Bayes classifier. Different measures can evaluate the performance of a classifier.

Accuracy The accuracy of a classifier is given by

$$M_{ACC} = \frac{TP + TN}{P + N},$$

where TP is the number of correctly labeled true positives, TN is the number of correctly labeled true negatives, P is the total number of positive tuples, and N is the number of negative tuples.

F_1-score The F_1-score, or F-measure is another way to evaluate the performance of a classifier.

$$M_{F1} = \frac{2 \cdot precision \cdot recall}{precision + recall}$$
$$precision = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{TP + FN}$$

Although both of these measures are here defined on a binary classification task, they can be extended to work on multiple classes by calculating a score for each class in a "one vs. rest"-manner.

Training classifiers on the anonymized data gives a measure of the usefulness of the data for a specific data use case. However, if the base performance of a classifier on the original data is low, it might be pointless to release an anonymized version of this data, as the risk of privacy breaches might be too high compared to the low use for classification of the anonymized data. It might also be valuable to evaluate a classifier on the trivially sanitized data (data where either all quasi-identifiers or all sensitive attributes are suppressed). If the performance of classifiers on the trivially sanitized data is approximately equal to that of the anonymized data, then non-trivial sanitation might be pointless.

A.1.6 Honorable Mentions

In this section, we mention metrics that were encountered during the research for this work but not investigated thoroughly.

β-likeness Similar to the parameter t in the t-closeness privacy model, β-likeness places an upper bound on the distribution of sensitive values in each EC in the anonymized table [CK12]. In contrast to t-closeness, this metric places a bound on the frequency of each sensitive value in an EC, and not on the overall distribution of sensitive values.

δ-presence Nergiz et al. [NAC07] introduce δ-presence as a metric to evaluate the risk of an individual in a table being identified using an external public table.

Probabilistic Anonymity This metric is proposed by Yang and Qiao [YQ10] and also used by Eyupoglu et al. [EAZS18] and measures the inability of an attacker to link quasi-identifier values to sensitive attributes when the quasi-identifier values have been randomly changed to achieve anonymity. For each tuple, the anonymization approach randomly picks an attribute of the quasi-identifier and exchanges the value of this attribute and tuple by sampling a new value from the distribution of values of the selected quasi-identifier attribute.

Mutual Information Goldberger and Tassa [GT09b] present a metric that enhances the correlation between the generalized public data and the private data. It measures utility content in the anonymized data over utility loss.

Minimum absolute distance Samarati [Sam01] provide one of the first algorithms achieving k-anonymity that minimizes a metric score. The paper outlines a handful of measures that can be used to guide the algorithm in finding an optimal transformation. Although not explicitly denoted as metrics in the paper, these measures can be used as such. There are some caveats, however, in that some measures require a Value Generalization Hierarchy and therefore assume that data have been generalized using global recoding as introduced in Section 3.2.1. The original algorithm minimizes the "minimum absolute distance". The value of this metric is the total number of generalization steps across the Value Generalization Hierarchy of all attributes. All steps contribute a value of 1 to the total amount.

Minimum relative distance Another metric proposed by Samarati [Sam01] sums up the ratios of generalizations done across the hierarchies. It can already be found in an earlier work by Samarati and Sweeney [SS98] on k-anonymity. The amount of generalization steps is divided by the height of the respective VGH to produce a relative value.

Maximum distribution This is the last metric proposed by Samarati [Sam01] that we discuss in this section. The metric sums up the number of distinct tuples in the anonymized dataset. Since the tuples in one EC are indistinguishable from each other, the value of this metric can be simplified to the number of ECs in the anonymized table. These straightforward metrics were initially meant to be used as preference policies for the anonymization algorithm, and some of them are, therefore, very specific to one anonymization approach. It may, therefore, be sensible to combine some of these measures into a more powerful metric. The measure of minimum absolute distance is trivially

optimized by not generalizing any values and suppressing all records (the same applies to the minimum relative distance). To prevent suppression of the entire table, Samarati [Sam01] impose a hard limit on the total number of suppressed tuples considered acceptable.

Sadhya and Chakraborty [SC22] propose the Sanitization Degree, consolidating the data privacy and utility aspect into one value in $[0, 1]$.

Kikuchi et al. [KYH+16] hold a competition of two players, one trying to anonymize a dataset as securely as possible, the other trying to estimate the anonymization method and re-identify records in the anonymized data. They assume a worst-case attacker who knows the original dataset and tries to map the anonymized records to the original ones. This attacker model is introduced by Domingo-Ferrer et al. [DRS15] with the idea that anonymization that protects against the worst-case attacker also protects against all other attackers. The model only considers the risk of Identity Disclosure. Still, the authors note that it is also possible to assess the risk of attribute disclosure by excluding one attribute of the original dataset from the attacker's knowledge.

Wang et al. [WFY05] propose an anonymization approach whose aim is to preserve the classification value of the data while eliminating possible sensitive inferences. To eliminate these inferences, the data owner defines a set of inference templates. Each template specifies an inference from a set of attributes to some sensitive attribute and a maximum allowed confidence for this inference. Starting with the original table and every attribute value suppressed, their algorithm progressively discloses single attribute values according to a priority metric while ensuring no inference templates are violated. This metric calculates the ratio of information gain divided by privacy loss due to the disclosure of the attribute value.

Evfimievski et al. [EGS03] give metrics for the analysis of privacy preserved by randomization. They show that using the measure based on mutual information introduced in Agrawal and Aggarwal [AA01] has limitations. Therefore, the authors define "worst case information" as a bound on "upward privacy breaches".

Nergiz et al. [NCN09] consider anonymization of data with multiple relations. This means that information about an entity is contained in multiple tables and not easily represented in a single table. The authors adjust the General Loss Metric [Iye02] and Discernibility Penalty (see Section 4.2.6) for this scenario.

Park and Shim [PS07] use the number of suppressions, which was allegedly proposed by Meyerson and Williams [MW04].

Xiao and Tao [XT06] propose a reconstruction error based on the actual pdf and the approximate pdf of the data.

Nergiz et al. [NAC07] give a metric that focuses on the risk of membership disclosure of an individual in a data set.

Agrawal and Srikant [AS00] provide a privacy metric that measures how well the original values of a perturbed table can be estimated. They use the confidence that the original value of a perturbed value lies in a certain interval as a privacy measure. Agrawal and Aggarwal [AA01] show weaknesses of this metric and additionally introduce a general-purpose metric for information loss based on the L1-norm.

A.2 Existing Software for Anonymization

In the following, we introduce all existing software investigated for usage in the experiments for this work. The software mentioned has been found using web searches for data anonymization tools, investigating papers in the field that use existing software to conduct their experiments, and searching for collections of anonymization tools presented in the literature like the work by Silva et al. [SMS21].

Desktop Applications

Amnesia is a desktop software for anonymization but also provides source code allowing users to set up their own instances [Dim]. The software allows users to upload their own data and have it anonymized. It provides five different privacy models, two of which are achieved using pseudonymization. Another privacy model is k-anonymity, and the two remaining ones are weaker forms of k-anonymity. Amnesia takes an arbitrary dataset as its input and produces an anonymized output that can be exported from the tool. The user can define their own Value Generalization Hierarchies or have them automatically generated by the tool. After creating all the necessary hierarchies, the user can select the parameter k and have the tool produce a new dataset that satisfies k-anonymity according to the given k and the specified VGHs.

TIAMAT is a tool for analyzing anonymization techniques [DGB+09]. The software compares anonymization quality and runtime efficiency between various methods that implement k-anonymity. The metrics it uses to compare the quality of the anonymized data are the Certainty Penalty Xu et al. [XWP+06] and the Classification Metric [Iye02].

SECRETA is an anonymization tool that also allows the evaluation and comparison of different anonymization approaches [PGL+13; PGL+15]. The tool supports anonymization of relational, transactional, and relational-transactional data. The metrics used to evaluate the anonymization methods for relational data are the Certainty Penalty [XWP+06] and the average relative error that a workload of `COUNT` queries shows on the anonymized data compared to the original data. For transactional data, the software uses the Utility Loss metric as introduced by Loukides et al. [LGM11], and the aforementioned average relative error, as it works on both data types. This relative error metric is also used by LeFevre et al. [LDR06].

The Cornell Anonymization Toolkit (CAT) provides many different methods for PPDP [guo14]. CAT implements the Incognito algorithm [LDR05] and the l-diversity privacy model. The anonymized data is then analyzed in terms of privacy based on user-specified assumptions about the background knowledge of an adversary.

ARX is the tool selected for our experiments. It is introduced in Section 4.4.1.

μ-ARGUS and τ-ARGUS are software programs developed partly under funding from the European Union [AAR21; PAJJ20]. They provide multiple ways to anonymize microdata using global recoding, suppression, micro-aggregation, and also synthetic data generation. τ-ARGUS directs its focus on privacy preservation for tables and μ-ARGUS on general microdata. The two tools are similar in many ways, but their main difference is that μ-ARGUS will only process microdata given as Action Script files (.asc) while τ-ARGUS will take Action Script files, Comma Separated Values (.csv) or Tab Separated Values (.tab).

Free and open-source software libraries

The implementations listed in this section are free and openly available libraries or programming frameworks.

Implementations of Grouping-based approaches There exists a GitHub repository that implements five different algorithms to achieve k-anonymity on the anonymized dataset [Pha23]. The anonymized datasets can be evaluated using the EC size metric (Section 4.2.5), the Discernibility Metric (Section 4.2.6), and the Certainty Penalty [XWP+06]. Furthermore, the authors implement three different classification models for further evaluation of data utility. However, the software is not designed to work on arbitrary datasets. Instead, five commonly used benchmark datasets are included in the repository, and the instructions show only how to apply the algorithms to the included datasets.

`crowds` is a Python module that implements k-anonymity using the Optimal Lattice Anonymization (OLA) algorithm [Maz23]. OLA was introduced by El Emam et al. [EDI+09] and transforms a dataset to satisfy k-anonymity via single-dimensional global recoding. Apart from generalization strategies, the algorithm requires an information loss function that is minimized during the anonymization process.

[Pra23] implements k-anonymity, l-diversity, and t-closeness using the Mondrian algorithm [LDR06]. This algorithm partitions the original data set into smaller and smaller groups using the median of values along one attribute in a group. After the dataset has been partitioned, the partitions need to be recoded to reduce the uniqueness of the tuples they contain. This can be done in different ways, where one is to replace all tuples with the median tuple in that group, similar to microaggregation. The algorithm requires that the span of attribute values be defined. This is trivial for numerical values, as their span will be the difference between the maximal and minimal value for the attribute. Calculating the span of a categorical attribute is trickier, as categorical values often have semantic meaning, and distance may not be well-defined between them. One possibility would be to assign each value an integer, starting with 1 and then counting up until the number of distinct values is reached. This can potentially assign very different semantic meanings to the differences between two values. Furthermore, the algorithm only supports a single sensitive attribute.

Implementations of Differential Privacy Ektelo is a programming framework for writing differentially private computations [ZMK+18]. It provides functionality to implement privacy algorithms through the use of powerful operators. The framework is based on an article by Zhang et al. [ZMK+18].

Google maintains a repository implementing various library functions to generate differentially private statistics over datasets [Goo23].

GUPT is a data mining platform that guarantees that any output of data analyses performed by the tool satisfies Differential Privacy [Moh23].

TensorFlow Privacy is a Python library that implements differentially private training of ML models with TensorFlow optimizers [McK]. The team of TensorFlow themselves maintains it.

Lillelund [Lil23] maintain a repository implements the differentially private Stochastic Gradient Descent for ML that was proposed by Abadi et al. [ACG+16].

Diffprivlib is the second software selected for our experiments. It is introduced in Section 4.4.1.

The Differentially Private Convex Optimization Benchmark is a repository that implements a new algorithm by Iyengar et al. [INS+19] and empirical evaluations for this algorithm and several previous approaches on real data [sun23]. They provide model implementations for Logistic Regression and Support Vector Machines that are fitted using differentially private optimization algorithms.

Implementations for Federated Learning Flower is a framework that provides an infrastructure to execute Federated Learning tasks. The framework supports many different ML models and programming languages [Aut]. Simple Flower projects can be set up quickly thanks to its comprehensive documentation. The framework is presented in the work of Beutel et al. [BTM+20].

Implementations for Data Synthesis In 2018, the National Institute of Standards and Technology (NIST) of the United States of America hosted a "Differential Privacy Synthetic Data Challenge", which produced various promising solutions to differentially private data synthesis [AS19; Gar19; LZW19; McK]. The overarching aim of the challenge was to produce algorithms that retain as much useful information as possible about the original data.

Other Implementations

The "Privacy Protection Application (PPA)" is a tool that provides functionality for the anonymization of vehicle trip data [USD23]. The implemented algorithm works by suppressing sensitive location information.

OpenAnonymizer implements k-anonymity and l-diversity in Java [Sta15].

There also exist various implementations of anonymization methods that work on different data types. SecGraph [JLM+15] implements many anonymization algorithms, metrics, and de-anonymization attacks on graph data.

sdcMicro is an open-source package for the programming language R that implements methods of statistical disclosure control [TMK+23]. The theoretical basis for the implementation is the work of Templ [Tem17]. They include all methods that are also present in μ-ARGUS and additional ones. In contrast to μ-ARGUS, which is a desktop application, it is possible to use the package from other software in a flexible way.

Anonimatron is a free and extendable data anonymization tool [rea23]. The overarching goal of the developers is to provide a tool that supports workers on a software project in safeguarding sensitive production data if that data is used locally to find a bug, for example. The software works by replacing sensitive values in a dataset with synonyms generated by the tool and maintaining a mapping table that translates between original and synonymous values.

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature