Institute of Parallel and Distributed Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelor Thesis

Differential Privacy by Sampling

Olena Kuksina

Course of Study:	Data Science
Examiner:	Prof. Dr.-Ing. habil. Bernhard Mitschang
Supervisor:	Dr. rer. nat. Christoph Stach
Commenced:	December 5, 2022
Completed:	June 5, 2023

Abstract

Collection and storage of immense volumes of data has become commonplace in today's digital age, making the protection of personal data increasingly important. Private data often includes sensitive information about an individual, and is gathered by medical and financial institutions, research and social science organisations, government, etc., taking full advantage of data-driven analytics and knowledge-based decision-making to improve products and services, enterprise statistical analysis, comprehensive studies of demographic trends, and many others. The disclosure or sharing of such information among different parties could infringe on privacy. This information can be used for malicious purposes, such as identity theft, scams or targeted advertising. This work examines the field of privacy-preserving data publishing.

Quality of published data significantly affects not only understanding and processing strategy, but the accuracy of data analysis as well as consequently the interpretation and decisions derived from the data. In order to meet this challenge, synthetic anonymization techniques, such as $k-$anonymity and its enhanced algorithms, are applied. However, they are based on the background knowledge of the adversary. A semantic model, or differential privacy, is a more rigorous mathematical notion of privacy assurance that operates under no assumptions. Nevertheless, differential privacy applies to the subsequent phase, namely privacy preserving data mining, query answering and aggregate statistics.

In the scope of this work, a subsampling anonymization algorithm DP-anonym providing $k-$anonymity with integrated differential privacy mechanisms, such as Laplace mechanism and exponential mechanisms, is elaborated. The algorithm provides synthetic and semantic privacy, combining the best of the two areas of private data exploration. According to experimental results, the proposed DP-anonym algorithm provides better data utility when compared to standard anonymization algorithms among general data utility metrics. It also provides more precise answers to typical database queries as it uses multidimensional generalization approach. In contrast to standard methods, DP-anonym achieves $(\epsilon, \delta)-$differential privacy, which guarantees the privacy of published anonymized data more efficiently.

Zusammenfassung

Im heutigen digitalen Zeitalter sind das Sammeln und Speichern enormer Datenmengen alltäglich geworden, womit die Sicherung persönlicher Daten stetig an Relevanz gewinnt. Private Daten zeichnen sich dadurch aus, dass sie häufig sensible Informationen über die betroffenen Individuen enthalten, welche von medizinischen Einrichtungen und Finanzinstitutionen, Forschungseinrichtungen und Wissenschaftlichen Organisationen, der Regierung und deren Behören sowie diversen anderen Stellen gesammelt und genutzt werden, um die Vorteile auszuschöpfen welche datenbasierter Analysen und wissensbasierter Entscheidungsfindung zur Optimierung von Produkten und Dienstleistungen, statistische Unternehmensanalysen, umfassende Studien zu demografischen Trends um einige zu nennen, mit sich bringen. Die Veröffentlichung oder gemeinsame Nutzung verschiedener Parteien solch sensibler Daten kann zu Verletzungen der Privatsphäre führen. So können die Daten mutwillig missbraucht werden, beispielsweise durch Identitätsdiebstahl, Betrug oder zielgerichtete Werbung. Diese Arbeit untersucht den Bereich der datenschutzfreundlichen Datenveröffentlichung.

Die Qualität der veröffentlichten Daten hat erhebliche Auswirkungen auf das Verständnis und die Verarbeitungsstrategie sowie auf die Genauigkeit der Datenanalyse und folglich auf die aus den Daten gewonnenen Erkenntnisse und Entscheidungen. Um diese Herausforderung zu meistern, werden synthetische Anonymisierungstechniken wie $k-$anonymity und ihre optimierten Algorithmen eingesetzt. Sie basieren jedoch auf dem Hintergrundwissen des Gegners. Das semantische Modell oder die differential Privacy ist ein rigoroserer mathematischer Ansatz zur Gewährleistung der Privatsphäre, der ohne Annahmen auskommt. Allerdings lässt sich die differential Privacy auf die nachfolgende Phase anwenden, d. h. auf die datenschutzfreundliche Data Mining, die Anfragebeantwortung und die aggregierten Statistiken.

Im Rahmen dieser Arbeit wird ein Subsampling-Anonymisierungsalgorithmus DP-anonym ausgearbeitet, der $k-$anonymity mit integrierten differential privacy Mechanismen, wie dem Laplace- und dem Exponentialmechanismus, gewährleistet. Der Algorithmus bietet synthetische und semantische Privatsphäre und kombiniert damit die Vorteile beider Bereiche der privaten Datenexploration. Die experimentellen Ergebnisse zeigen, dass der vorgeschlagene Algorithmus DP-anonym im Vergleich zu Standard-Anonymisierungsalgorithmen einen höheren Datennutzen unter den allgemeinen Datennutzenmetriken bietet. Außerdem liefert er präzisere Antworten auf typische Datenbankabfragen, denn er verwendet einen mehrdimensionalen Generalisierungsansatz. Im Gegensatz zu Standardmethoden erreicht DP-anonym die $(\epsilon, \delta)-$differential privacy, was effizienter die Privatsicherheit der veröffentlichten anonymisierten Daten gewährleistet.

Contents

List of Figures

List of Tables

1 Introduction

Privacy is a fundamental right that ensures the protection of personal information and prevents its unauthorized use. In today's digital age, the collection and storage of vast amounts of data has become commonplace, making the protection of personal data increasingly important. Data is gathered by medical and financial organisations, research and social institutions, government, etc., exploiting all the potential benefits of data-driven analytics and knowledge-based decision making to enhance the quality of products and services, enterprise statistical analysis, comprehensive population trends and many others. When data is collected, it can be used to identify individuals, their preferences, and even their behavior. This information can be misused for malicious purposes, such as identity theft, fraud, or targeted advertising.

Personal data often includes sensitive information about individuals, and releasing or exchanging such information directly among various parties infringes on their privacy. The standard approach to addressing this issue is through the implementation of policies and regulations that limit the types of data being published and by agreements which govern the handling and storage of confidential data [VV17]. However, this method has its drawbacks, as it either heavily alters the data or requires an unrealistic level of trust in many data sharing situations. Internet of Things devices, for example, require specific permission models to guarantee confidentiality [SM18] when exchanging large amounts of user information among paired devices. There is also an urgent demand, even when using smartphones as part of daily routine, for methods to ensure application privacy management [Sta15].

Moreover, privacy is essential for the exercise of other fundamental rights such as freedom of speech, freedom of association, and the protection of personal security. When individuals are aware that their personal information is being collected and used, they are less likely to freely express their opinions or participate in surveys and activities that they believe may be viewed as controversial. It is imperative that individuals are able to control how their data is collected, used, and shared, and that organizations take adequate measures to secure personal data and prevent it from usage in a way that may negatively impact someone's life.

1.1 Personal Privacy Data Breaches

In the following, several data privacy breaches by means of linking attacks are discussed and revealed, why there is an urgent need for strong privacy notions to prevent individuals' sensitive information from leakage.

- The NYC Taxi & Limo Commission.

  In 2014 the NYC Taxi & Limo Commission shared visualizations of taxi usage statistics on Twitter, and claimed that this data was became available upon Freedom of Information Act (FOIA) request. FOIA allows citizens to request data from certain government organizations.

  Chris Whong filled a FOIA request and obtained a dataset of all taxi fares and trips in New York City during 2013 [Who14] - a total size of 19 GB. The table included attributes such as: medallion, hack license, pickup datetime, dropoff datetime, passenger count, trip time in secs, trip distance, pickup longitude, pickup latitude, dropoff longitude, dropoff latitude, etc. Most of these fields are comprehensible from the name, although the medallion of the taxi driver and his licence number are interesting for privacy analysis. These is a special permit in the United States that allows a taxi driver to provide a passenger transport service. The Commission attempted to obscure this information, using a form of anonymization, because the standard format for these fields was rather different than what was provided in the dataset.

  It was found that someone with one special medallion number:
  $''CFCD208495D565EF66E7DFF9F98764DA''$ had unusually profitable days, earning many times more than the other drivers, simultaneously working at different locations, avoiding sleep. Pandurangan was able to find out in his study [Pan14] that his medallion number produced a string of $''0''$ when decrypting it with hash function Message-Digest Algorithm 5 (MD5) and was used as the default value in situations where a number was missing or not entered in the record. All the medallion and license numbers were the plaintext hashed via MD5. Then the unhashed values were matched using other publicly available datasets. As a result, driver names, their incomes, addresses, as well as passenger information, their journey points and times of travel were revealed: a massive privacy violation!

- The Netflix Prize competition.

  The Netflix Prize competition serves as another example of poor data anonymization. Netflix's content is developed using user data, and its recommendation system is calibrated to maximize user engagement based on data-driven and statistical analysis.

  Netflix hosted a contest to find the ways of improving their recommendation engine in 2006 – 2009. The grand prize was a US$1.000.000. Netflix provided a training dataset of user data with attributes of an anonymized user ID, movie ID, rating, and date.

  Narayanan and Shmatikov demonstrated in their paper [NS08] that this naive form of anonymization done by Netflix was insufficient to preserve user privacy. They took the dataset provided by Netflix and cross-referenced it with the public information from the online movie database IMDb, which contained hundreds of millions of movie reviews (see example in Fig.1.1).

  Thus, they re-identified many users, giving the information on these users' movie watching history. This discovery led to a class action lawsuit filed against Netflix, and the cancellation of a sequel competition. This example shows that de-anonymization is insufficient to guarantee privacy, especially in the presence of side-information.

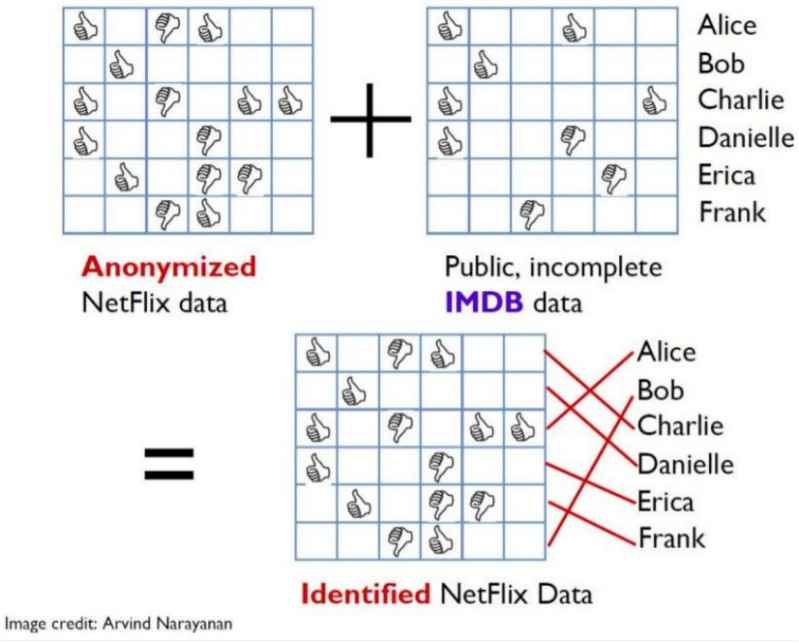


Figure 1.1: Cross-referencing of Netflix Training Dataset with IMDb Database.

- The Equifax data breach.

  Another well-known example is the Equifax data breach [ZMMS18], where the personal information of 143 million people, including Social Security numbers and birth dates, was exposed in 2017. The breach was a result of a vulnerability in Equifax's website, and it had significant consequences for individuals, as the sensitive information could be used for identity theft and other forms of fraud.

- The Cambridge Analytica scandal.

  A further famous example of a privacy breach is the Cambridge Analytica scandal [HWJ20], where the data of millions of Facebook users was harvested and used for political advertising purposes. The data was collected without users' consent, and it was later revealed that the information was used to target political advertisements during the 2016 US Presidential election. This breach of privacy led to widespread concern about the protection of personal information and sparked a global debate about data privacy.

- Reidentification of anonymized data from credit card metadata.

  De Montjoye et al. analyzed in their study [DRSP15] three months of credit card records for 1.1 million individuals and found that just four spatiotemporal points are sufficient to uniquely reidentify 90% of individuals. The study also found that knowing the price of a transaction increased the risk of reidentification by an average of 22%. In addition, it was

noted that women were more likely to be re-identified by credit card metadata than men. Thus, it was concluded that even datasets having aggregated information on several or all attributes provide little anonymity.

These examples highlight the importance of privacy in data publication and the consequences of failing to protect personal information. When data is collected and shared, it is essential that privacy is maintained.

1.2 Goals of the Work

This work studies and investigates the possibilities of releasing data to a data scientist in a private manner. A human is often involved in the data analysis and thus s/he can either maliciously or accidentally harm the privacy of the individuals involved in the dataset. In order to prevent this situation, it is worthwhile publishing the data in an anonymized and private manner already during the pre-processing stage.

Typically, $k-$anonymization methods are applied within the privacy- preserving data publication. They ensure the confidentiality of the individual with a probability of 1 to at least k, thereby preventing the possibility of precisely identifying the individual in the database. However, $k-$anonymity is grounded on the fundamentals of adversary background knowledge and can be vulnerable to a number of attacks. In addition, the methods apply generalization (obfuscation) and suppression (in worst case, deletion) of data values, which at high privacy expectations negatively affects the data utility. As the pre-processing stage is one of the first in the data analysis process, there is an urgent need to keep the utility of the data as high as possible.

A more advanced and more rigorous mathematical notation is differential privacy model. This model ensures privacy by, so to say, hiding the individual in the crowd, which means adding noise to the generated responses during the algorithm‘s operation. In doing so, the algorithms retain good usability of the data. However, mechanisms that provide differential privacy are employed in the privacy- preserving data mining, aggregated statistics and query responses. Since this step of analytics takes place after pre-processing, differential privacy mechanisms in their pure form cannot be applied to the problem posed in this study.

Considering that both $k-$anonymity (and its enhancements) and differential privacy have sufficient advantages to be applied at the stage of data publication for pre-processing, an algorithm combining the strengths of both domains has been developed. Namely, the algorithm provides both $k-$anonymity and differential privacy while preserving the usefulness of the data.

1.3 Outline

This work is structured out as follows:

Chapter 1 - Introduction highlights the extensive nature of the privacy problem, and reveals the motivation and challenge of designing this work's algorithm. It also covers unsuccessful attempts to withhold private data in practice.

Chapter 2 - Privacy Preserving Techniques introduces the reader to the main privacy protection techniques and their properties, on the basis of which the elaborated algorithm has been developed. Methods to achieve $k-$anonymity and its refinements - $l-$diversity and $t-$closeness are deeply examined, examples of their usage and the potential attacks to which they are susceptible are discussed. Then, to counter the preceding, a discussion on the conceptual idea of differential privacy is held on. An entirely different perspective on the question of privacy between anonymization techniques and differential privacy is outlined. In the following, central and local models of differential privacy are considered and the feasibility of applying them to a given algorithm is presented. A variety of implementation cases of differential privacy in leading IT companies is discussed. Further consideration is given to the investigation of possibility to combine anonymization and differential privacy methods proposed by the research community. The chapter concludes with a detailed comparison of the methods discussed and highlights their positive and negative implications to achieve the goal of this work.

Chapter 3 - Anonymization Algorithms looks in detail at four most widely used $k-$anonymization algorithms, that inspired the algorithm presented in the work. In the following, a comparison of their technical characteristics is presented, in order to identify weaknesses and avoid them in the development of a new algorithm.

Chapter 4 - Differential Privacy considers properties and basic mechanisms of differential privacy extensively, such as Randomized response, Unary encoding, the Laplace and exponential mechanism. This chapter attempts to identify opportunities to incorporate mechanisms into the new algorithm to increase the utility of the anonymized dataset. A relaxed variant, that is - approximate differential privacy and the Gaussian mechanism applicable to it are also discussed here. This chapter also examines a model that combines a central and local model of differential privacy as an option to bypass unilateral data retention. A comparison of the operation and resulting performance of the mechanisms is discussed in the concluding part of the chapter.

Chapter 5 - Privacy Consequences of Sampling examines the possibility of combining sampling with anonymization and differential privacy approaches. Because the sampling step is not only inherent in the data analysis, but also increases the privacy of the processed data, the application of this step in the new algorithm is considered. This chapter also investigates the privacy amplification effect of sampling on the anonymized data, which is widely discussed in the research community.

Chapter 6 - Data Utility Metrics in Anonymization researches the data utility metrics by which the algorithms in Experimental section can be compared. A discussion is held on which metrics are more reasonably applicable to the given algorithms and which are selected for evaluation.

Chapter 7 - Algorithm for Privacy-aware Data Publishing presents an elaborate description of the task of this work, the initial requirements and presents the concept of the developed algorithm. Here, the generalization model and the stages of implementation of differential privacy mechanisms are thoroughly investigated. Then the privacy analysis of the proposed algorithm, both at individual stages and overall, is performed.

Chapter 8 - Experimental Evaluation depicts the evaluation of the proposed algorithm in comparison to standard algorithms based on data utility metrics to analyse the quality of anonymization. The datasets that were anonymized, the library employed for the implementation and the results are presented and discussed here. The degree to which anonymized data loses utility, the dependence of data utility on the level of epsilon-differential privacy, and a comparison of the outcomes of typical query results in real-world use cases is explored in this chapter.

The concluding **Chapter 9 - Conclusion** outlines the findings and contributions of this work.

2 Privacy Preserving Techniques

Two lines of research are distinguished in the literature : Privacy preserving data mining (PPDM) [AP08; AS00] and Privacy preserving data publishing (PPDP) [FWCY10]. The concept of privacy preserving data mining involves retrieving relevant information from modified data with hidden sensitive information, while ensuring the applicability of data mining techniques. The challenge is how to modify the data and how to extract useful insights from the altered data. The solutions are often closely tied to specific data mining algorithms.

In contrast, privacy preserving data publication focuses on releasing data for research purposes while protecting individual privacy, and may not necessarily be linked to a specific data mining task. Some PPDP solutions prioritize preserving the accuracy of data at the individual level, but PPDM solutions do not always do this. In recent years, the definition of PPDM has broadened to encompass other privacy-related research topics, not just those related to classic data mining techniques.

The goal of privacy preserving data publication is to make data accessible for a third party but to preserve the privacy of individuals. Various techniques are used to achieve this, such as data perturbation, $k-$anonymization, $l-$diversity, $t-$closeness, and differential privacy. These techniques work by transforming the original data in ways that can protect the privacy of individuals, but at the same time make the data useful for further analysis and research.

The problem of preserving privacy in data releasing was awoken a long time ago by Dalenius [Dal86]. One of the main challenges of data publication is balancing the high utility of released data with the high privacy guarantee. Privacy preserving data publication addresses this challenge by enabling data to be shared for research or analysis purposes whilst ensuring that sensitive information is not disclosed. This helps to prevent the potential for data misuse, identity theft, or other privacy violations.

The anonymization methods protect data under the assumption that an adversary has some background knowledge. A more robust and mathematically rigorous method is differential privacy (DP) which explicitly specifies a privacy budget. It eliminates the limitation of anonymization methods. DP centres on PPDM, aggregated statistics and query responses. Therefore, it is rather difficult to produce high quality private data for publication. The positive effect of combining anonymization methods with DPs has been discussed in academic circles, though with regard to PPDM purposes only. This paper analyses the problem of combining anonymization and DP methods in data publication context. This chapter studies the related data privacy preservation methods in more detail, compares them and highlight some positive effects when used in combination.

The dataset is assumed to be in the form of a table (relational) that includes a collection of tuples referring to a set of individuals and defined by a set of attributes. The following attributes are distinguished in the dataset based on their privacy requirement [XJW+14]:

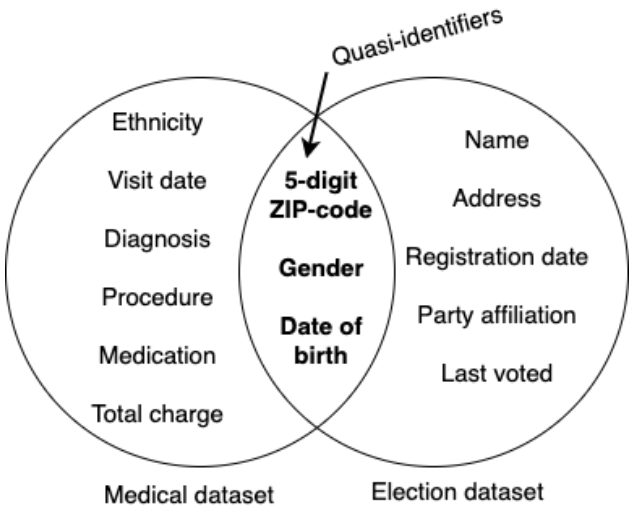


Figure 2.1: Re-identification of Massachusetts Medical Records.

- Identifiers (IID) - Attributes that explicitly and unambiguously identify a person (ID, SSN, mobile number)
- Quasi-identifiers (QI) [Dal86] - Attributes that can re-identify individual records by linking this attributes with external data (ZIP code, gender, date of birth)
- Sensitive (SA) - Attributes that have to be kept private (salary, disease)
- Non-sensitive (NSA) - other attributes.

2.1 $k-$anonymity

Latanya Sweeney shows in her study [Swe02b], that she has managed to reveal the identity and the medical sensitive information about the former governor of Massachusetts due to the linking attack on the medical records of the Group Insurance Commission and the election dataset for voter registration list for Cambridge Massachusetts (see Fig.2.1). She has found during her resarch work [Swe00] that 87% of the population of the USA can be uniquely or nearly uniquely identified by sharing only three quasi-identifiers - 5-digit ZIP-code, gender and the date of birth.

As a possible solution to the linking attack, she proposed the $k-$anonymity approach to ensure that each individual can "blend into the crowd". The idea of such an approach is to provide an individual with privacy, so that s/he can be hidden among at least other $(k-1)$ individual's records, which share same attribute value in the quasi-identifiers. The individual is indistinguishable among at least $(k-1)$ other individuals in the same dataset. The probability of identifying a person is less than $\frac{1}{k}$ and the privacy level directly depends on size of k.

Definition. $k-$**anonymity.** Let $\mathbb{D}$ be a table and $QI_{\mathbb{D}}$ be the quasi-identifier associated with it. D is said to satisfy $k-$anonymity if and only if each sequence of values in $\mathbb{D}[QI_{\mathbb{D}}]$ appears with at least k occurrences in $\mathbb{D}[QI_{\mathbb{D}}]$.

It is specified that for each individual only one record in dataset is present. For each combination of quasi-identifier values to be indistinguishably mapped to at least k individuals in the same equivalence class the $k-$anonymity technique most frequently uses data generalization and suppression [Swe02a].

The generalization operation [BA05] replaces some values with the parent value in the attribute taxonomy, f.e. 'dog' for 'Havanese' race.

A hierarchy of values is defined for attributes to be generalized. This generalization method, also called full domain generalization [LDR05], maps the entire domain of an attribute in original dataset to a more general domain from its domain generalization hierarchy. There are multiple possible variations based on the properties of each attribute to divide the values into levels. Figure 2.3b shows taxonomic tree for the attributes "Age", having three hierarchy levels, "Gender" with two levels (Fig.2.3a), and "ZIP-code" with three levels (Fig.2.2) to illustrate the following example.

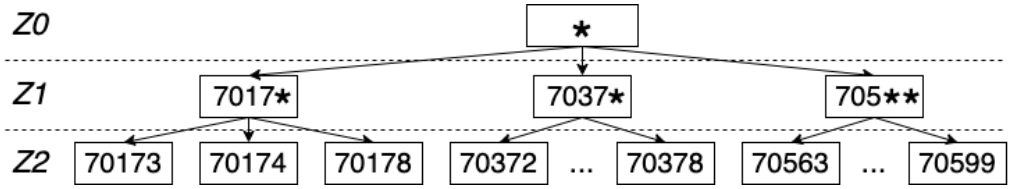
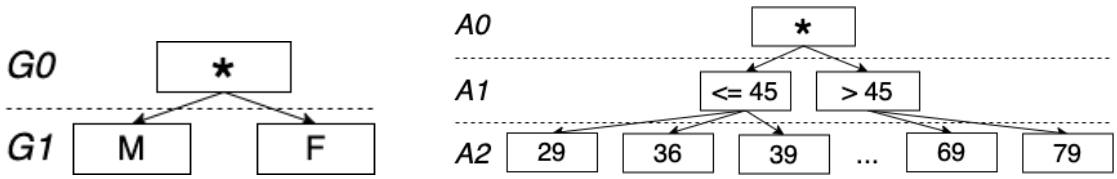


Figure 2.2: Taxonomy Tree of the Attribute "ZIP-code".



(a) Gender **(b)** Age

Figure 2.3: Taxonomy Trees of Attributes "Gender" and "Age".

Figure 2.4 shows a lattice of generalizations to visualize all possible combinations of generalized domains, when two or more attributes are generalized [Sam01]. After the generalization, the number of tuples with a frequency of the key attribute less than k can be determined. If this number is below a certain threshold, suppression is applied.

The suppression operation [Lit93] replaces some values with a special value to indicate that the substituted values are not revealed, f.e. $0123*$ for ZIP-code 01239.

Data suppression can also be expanded to the entire record, where confidential information and all inferences about the existence of confidential information are simply not published. This can be a good practice in suppressing outliers. However, suppression can severely reduce the quality of the data and the overall statistics can be significantly altered.

Table 2.2 shows an example of a dataset that satisfies $k-$anonymity, where $k = 4$ and the quasi-identifier is $QI_{\mathbb{D}} = \{\text{Age, Gender, ZIP}\}$. Thus, for each row in the table, the values of the quasi-identifier appear in the table at least four times. These row sequences are called equivalent class. There are three equivalent classes in this table:

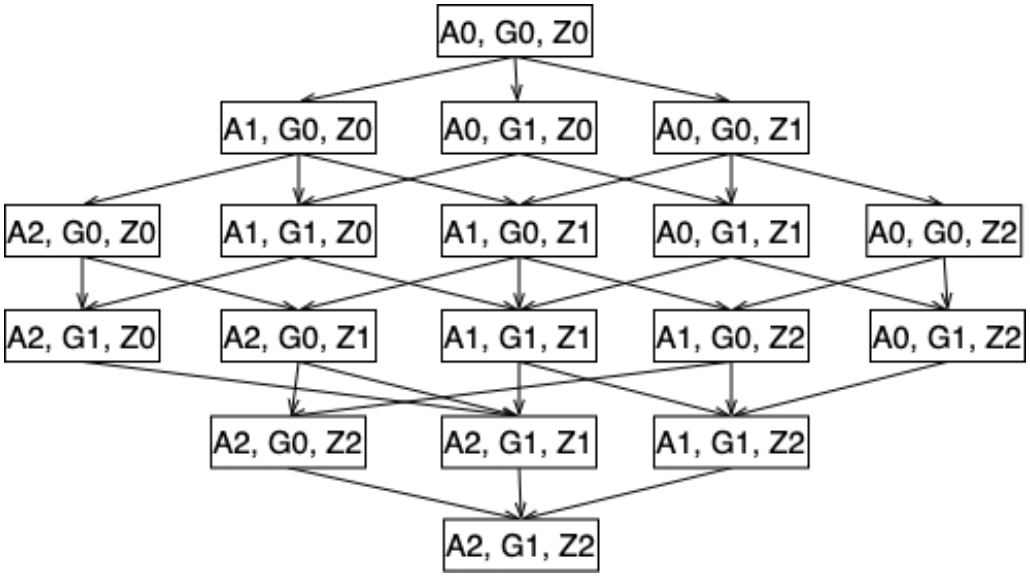


Figure 2.4: Generalization Hierarchical Lattice.

QI			SA
Age	Gender	ZIP	Diagnosis
69	M	70563	Stroke
61	M	70569	Infarction
48	F	70599	Rosacea
56	F	70597	Flu
79	M	70372	Arrhytmias
61	F	70374	Hypertension
54	M	70378	Infarction
49	F	70376	Hypertension
39	M	70173	Cirrhosis
43	F	70178	Cirrhosis
36	F	70174	Hemophilia
29	M	70173	Hemophilia

Table 2.1: Illustration of $k-$anonymity: Original Table.

QI			SA
Age	Gender	ZIP	Diagnosis
>45	$*$	705$**$	Stroke
>45	$*$	705$**$	Infarction
>45	$*$	705$**$	Rosacea
>45	$*$	705$**$	Flu
>45	$*$	7037$*$	Arrhytmias
>45	$*$	7037$*$	Hypertension
>45	$*$	7037$*$	Infarction
>45	$*$	7037$*$	Hypertension
$\leq$45	$*$	7017$*$	Cirrhosis
$\leq$45	$*$	7017$*$	Cirrhosis
$\leq$45	$*$	7017$*$	Hemophilia
$\leq$45	$*$	7017$*$	Hemophilia

Table 2.2: Example of $k-$anonymity, where $k = 4$ and QI ={Age, Gender, ZIP}.

As a result, $k-$anonymity prevents the potential linking attack. If the released table $\mathbb{D}$ satisfies $k-$anonymity with respect to the quasi-identifier $QI_{\mathbb{D}}$, then linking this data with other external sources using $QI_{\mathbb{D}}$ or its subset is only possible for k or more than k individuals, or in other words, it is impossible for less than k individuals.

Latanya Sweeney further emphasized that the $k-$anonymity approach is vulnerable to several attacks in her research paper [Swe02b], and offered some practical solutions:

1. **Unsorted matching attack.**
   Problem: The attack is based on the order in which the tuples appear when the table is released sequentially several times.
   Solution: Randomly sort the tuples in the table before releasing.

2. **Complementary release attack.**
   Problem: Sequential releases of a table that satisfies $k-$anonymity are treated as linking multiple tables. Therefore, subsequent releases of the same private information should consider all published attributes of the table as quasi-identifiable to prohibit linking by sensitive attribute.
   Solution: Use all attributes of the table as quasi-identifiable or use the previous release of the table as the basis for the next one.

3. **Temporal attack.**
   Problem: Suppose a sanitized table satisfying $k-$anonymity was released. The original table was then modified by deleting or adding entries. Thus, combining the old sanitized table and the new sanitized table (also satisfying $k-$anonymity) may reveal sensitive information and thereby violate $k-$anonymity protection.
   Solution: analogous to complementary release attack.

The finding of optimal k in this approach was proven to be NP-hard for all $k \geq 3$ by Meyerson and Williams in their research [MW04]. Instead, they proposed an approximation algorithm that is executed in polynomial time and does not suppress more than $O(k \log k)$ times the minimum number of entries that must be suppressed in order to achieve $k-$anonymity.

$k-$anonymity method preserves identity disclosure, which refers to identification of an entity. But it does not prevent the data from attribute disclosure, when the intruder finds out something new about the target entity. $k-$anonymity only prevents the association between individuals and tuples, instead of the association between individuals and sensitive values. A new concept of personalized anonymity is presented by Xiao et al. in the paper [XT06], where the degree of privacy protection for each entity's sensitive values can be specified. The authors have developed an algorithm that along with the generalization for quasi-identifiers also uses generalization for sensitive attributes. The similar approach is proposed by Wang et al. in their study [WFY07], where suppression for the second step is used.

Truta and Vinay introduced the concept of $p-$sensitive $k-$anonymity in their research paper [TV06], which brings diversity into the equivalence class to prohibit attribute disclosure, by requiring that at least p different values exist for each sensitive attribute in the records sharing the combination of quasi-identifiers. However, a big limitation of $p-$sensitive $k-$anonymity is that it is assumed that the frequencies of the different values of the sensitive attribute are the same, which is not always the case. This happens because some sensitive values naturally occur more frequently than others in the class, e.g, flu is more common than HIV in the medical database of a regular hospital.

2.2 $l-$diversity

Due to the tendency to attribute disclosure in the $k-$anonymized dataset the adversary is given the opportunity to make so-called a background attack and homogeneity attack:

- **The homogeneity attacks** may be implemented in the situation where the sensitive attribute lacks diversity.
  Example: All the individuals in the second equivalence class in Table 2.2 have heart problems. Thus, it is possible to find out the medical diagnosis of an individual who is known to belong to a given equivalence class, even without direct identity disclosure.
- **The background knowledge attacks** threaten privacy by drawing on partial knowledge about an individual, or about the distribution of sensitive attributes in a population.
  Example: Assume that the adversary's neighbor is taken to the hospital. Given that the adversary knows which equivalence class the individual is in Table 2.2, and also knows that the neighbor cannot have hemophilia based on his family history, he finds out that the neighbor has liver cirrhosis.

The $l-$diversity approach has been proposed by Machanavajjhala et al. in their paper [MKGV07] as an improvement to $k-$anonymity. The main idea is to achieve the variability, diversity of the sensitive attribute values in the equivalence class. This implies that in addition to $k-$anonymity, the sanitized table should also provide diversity in the sense that tuples having the same values for their quasi-identifiers should have different values for their sensitive attributes. The $l-$diversity requires that the values of sensitive attributes to be well-represented in every equivalent class.

The similar to $p-$sensitive $k-$anonymity [TV06] approach, distinct $l-$diversity privacy model was proposed earlier and it meets $k-$anonymity requirements automatically, where $k = l$, since each equivalent class contains at least l records with different values in sensitive attribute. This motivates the following two stronger notions of $l-$diversity.

Definition. $l-$**diversity.** An equivalent class is $l-$diverse if it contains at least l well-represented values for the sensitive attribute. A table is $l-$diverse if every equivalent class is $l-$diverse.

Machanavajjhala et al. provide several improved methods for $l-$diversity:

- The entropy $l-$diversity.

  A table is entropy $l-$diverse if for every equivalence class q holds

  $$-\sum_{s \in S} p_{(q,s)} \log(p_{(q,s)}) \geq \log(l)$$

  where $p_{(q,s)} = \dfrac{n_(q,s)}{\sum_{s' \in S} n_{(q,s)}}$ is a fraction of tuples in the equivalence class q with sensitive attribute value equal to s.
  In order to have entropy $l-$diversity for each equivalence class, the entropy of the entire table must be at least $\log(l)$. Thus, more evenly distributed sensitive values in the group result in a larger value of the entropy of the sensitive attribute. The larger the parameter l is, the less certain is the output of a particular sensitive value in the group.

- The recursive $(c, l)-$diversity.

  In a given equivalence class q, let r_i denote the number of times the i-th most frequent sensitive value appears in that class. Given a constant c, the equivalence class satisfies recursive $(c, l)-$diversity if $r_1 < c(r_l + r_{l+1} + ... + r_m)$. A table T satisfies recursive $(c, l)-$diversity if every equivalence class satisfies recursive $l-$diversity.
  According to this approach, the most frequent values do not occur too often, and the less frequent values do not appear too infrequently.

- The positive disclosure−recursive $(c, l)-$diversity and the negative/positive disclosure−recursive $(c_1, c_2, l)-$diversity,

  which serve to capture the attacker's background knowledge, as well as reflecting the type of implication knowledge [MKM+06].

As for $k-$anonymity, the $l-$diversity can not guarantee privacy protection if an individual may correspond to multiple tuples in the dataset [XT06].

One of the limitations of $l-$diversity, is the assumption that each sensitive attribute takes values evenly over its domain, which is not always the case. Achieving this leads to a large loss of data utility [FWCY10].

2.3 $t-$closeness

Distribution skewness and semantic similarity of the sensitive values in the equivalence class are possible attacks faced by the $l-$diversity technique, as it is limited in its assumption of adversarial knowledge. Li et al. showed in their research paper [LLV06] that $l-$diversity is also prone to attribute disclosure as there are two possible attacks: a skewness attack and a similarity attack.

QI			SA
Age	Gender	ZIP	Virus infection
>60	*	70535	Yes
>60	*	70535	Yes
>60	*	70535	No
>60	*	70535	No

Table 2.3: Example of $l-$diversity, where $l = 2$ and QI ={Age, Gender, ZIP}

QI			SA
Age	Gender	ZIP	Diagnosis
[20-35]	M	705**	Acne
[20-35]	M	705**	Psoriasis
[20-35]	M	705**	Alopecia
[20-35]	M	705**	Dermatitis

Table 2.4: Example of 4−diverse Table.

- **Skewness attack** occurs if the overall distribution of sensitive attributes is skewed.
  Example. Table 2.3 with 95% of records with positive sensitive attribute value and 5% of records with negative attribute value is given. Assume that there is an equivalent class of 50% positive and 50% negative and therefore the class satisfies the 2−diversity. However, the class introduces a serious privacy risk because any individual in this class has the potential to have a positive value with 50% confidence, compared to 5% in the overall table.

- **Similarity attack** is possible when the values of the sensitive attributes in the equivalence class are different, but semantically similar.
  Example. Suppose we know that an individual's record is in the given equivalent class in Table 2.4, then we can conclude that the individual has skin-related problems, because all three diseases in this class are skin disorders : "Acne", "Psoriasis", "Alopecia", "Dermatitis".

A database satisfies the $t-$closure requirement when the distance between the distributions of sensitive attributes in each equivalence class differs from the distribution of sensitive attributes in the whole table by no more than a given threshold t.

To calculate the difference between the two distributions, the Earth Movement Distance (EMD) [RTG00] is applied, where the distance describes the minimum amount of work required to convert one distribution to another by moving the mass of the distribution between them.

EMD can be formally defined using the transportation problem. Let $P = (p_1, p_2, ..., p_m)$, $Q = (q_1, q_2, ..., q_m)$, and d_{ij} be the ground distance between element i in P and element j in Q. The flow $F = [f_{ij}]$, where f_{ij} is the mass flow from element i to j for arbitrary i, j, which minimizes the overall work, has to be found. After transporting P to Q using mass flow F, the EMD is defined as the total work:

$$D[P, Q] = WORK(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}.$$

However, the EMD function is not well suited to prevent attribute linking to numerical sensitive attributes [LLV09].

A subsequent analysis of the utility of the data shows that $t-$closeness significantly limits the amount of useful information that can be extracted from the released data. Soon the authors of $t-$closeness Li et al. proposed a more flexible privacy model called $(n, t)-$closeness in their paper [LLV09], which requires that the distribution in any equivalence class must be close to the distribution in a sufficiently large equivalence class that contains at least n entries with respect to a sensitive attribute.

2.4 Semantic Privacy Guarantee

Cynthia Dwork in 2005 presented the notion of differential privacy (DP) in her research work [Dwo08]. The concept is now widely acknowledged as a strong and rigorous notion of privacy. It was awarded the 2016 TCC Test-of-Time Award, the 2017 Gödel Prize and the 2021 Paris-Kanellakis-Prize.

The anonymization models, like $k-$anonymity, guarantee privacy through syntactic conditions and although these privacy models can effectively protect privacy under certain conditions, they are vulnerable to various attacks as shown earlier. In contrast to syntactic privacy, differential privacy, or semantic privacy, ensures stronger mathematically provable privacy guarantees, regardless of the arbitrary background knowledge of adversaries.

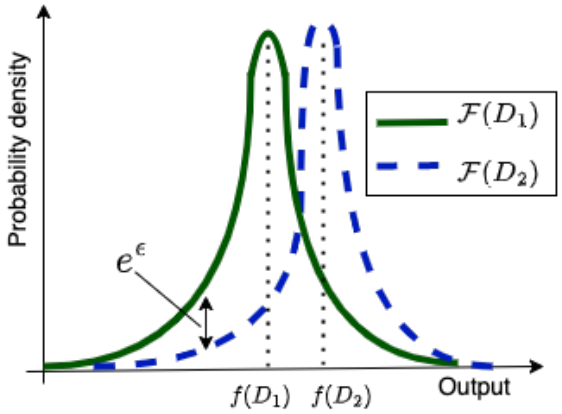


Figure 2.5: Illustration of Differential Privacy Definition.

It is important to note, that unlike $k-$anonymity, $l-$diversity or $t-$closeness, differential privacy is a property of the algorithm, not a property of the data. To show that a dataset satisfies differential privacy, it must be shown that the algorithm used to create or manipulate the data satisfies differential privacy.

Differential privacy is intended to ensure that if an individual participates in a differential private analysis of data, no additional harm will occur to the individual in consequence. DP captures the increased risk to one's privacy incurred by participating in a database. In the mathematical definition of differential privacy, this objective is achieved by requiring that the result of any differential private analysis will be indistinguishable whether or not the person participates in it (within a private budget).

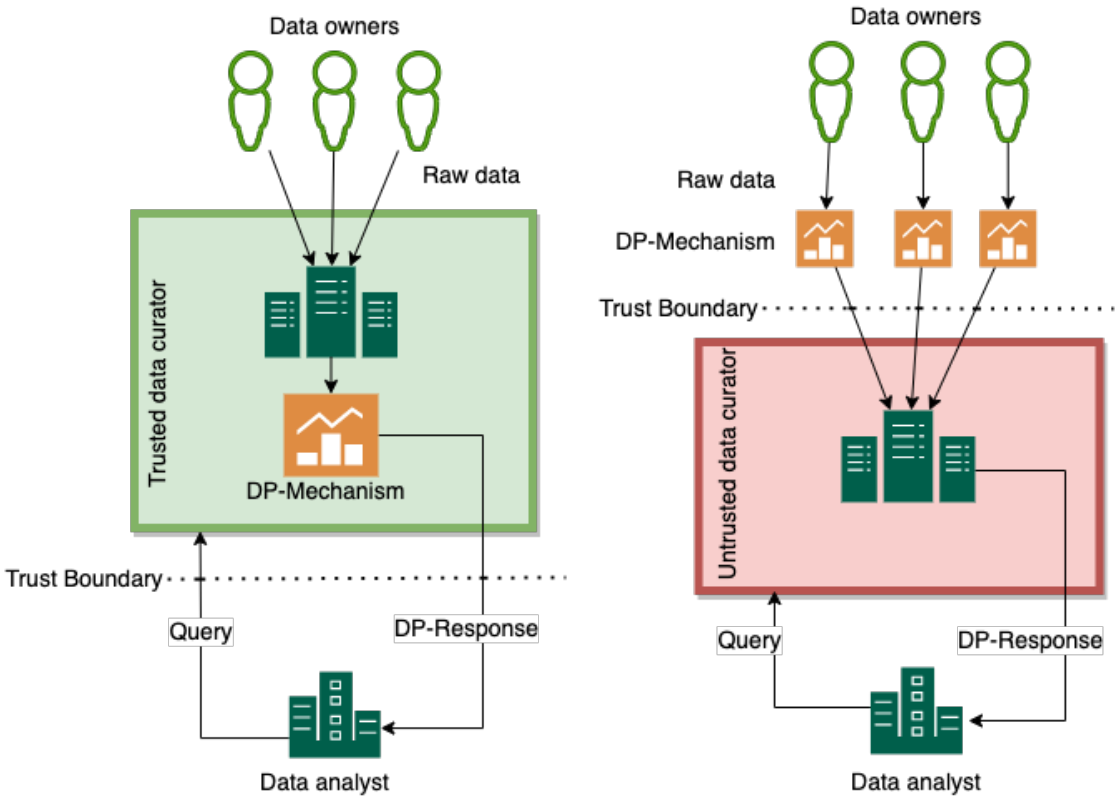
Differential privacy is conceptually designed around neighbouring databases - those databases that are formed by adding or removing a tuple in the database and are differentiated by one individual's record. Thus, neighbouring databases differ in their size by maximum one entry.

Definition. $\epsilon-$differential privacy. A randomized function $\mathcal{F}$ gives $\epsilon-$differential privacy if for all neighbouring datasets D_1 and D_2, and all subsets S of the Range($\mathcal{F}$), where Range($\mathcal{F}$) is the set of possible outputs of the randomized function $\mathcal{F}$.

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

$$\Rightarrow \log \frac{\Pr[\mathcal{F}(D_1) \in S]}{\Pr[\mathcal{F}(D_2) \in S]} \leq \epsilon$$

In this way, the randomness provided by the function $\mathcal{F}$ must be sufficient to prevent identifying in the output published dataset which of D_1 or D_2 was the input. The definition requires a mechanism to produce similar outcomes as shown in Figure 2.5 for neighbouring database inputs, making it impossible to determine whether any particular individual participated or not. Roughly speaking, differential privacy ensures that the removal or addition of a single database record does not substantially affect the outcome of any analysis. Thus, the differential private algorithm juxtaposes the privacy risk with and without the individual's record in the published data, instead of comparing the prior and posterior probability before and after having access to the released data.



(a) Central differential privacy model. **(b)** Local DP model.

Figure 2.6: Differential Privacy Models with Different Trust Boundaries.

The parameter ϵ epsilon in the definition is called the privacy budget. Smaller values of epsilon provide higher level of privacy. In this case the mechanism is required to provide very similar outputs when given similar inputs and it is different to say, whether the individual has been included in the dataset or not. Mathematical definition and properties of differential privacy are studied in detail in Chapter 4.

2.5 Central and Local Models of Differential Privacy

There are two different models of DP mechanisms: central and local. In the central model (Fig.2.6a), the confidential data is collected in a single dataset. In this situation, it is assumed that there is a trusted data curator, who stores the dataset and honestly handles requests from the (malicious) data analyst by applying DP mechanisms. However, this situation is not always realistic. In many cases the data curator is untrusted, or the curator and the (malicious) data analyst are the same entity.

An alternative to the central model is the local DP model shown in Figure 2.6b. In this model the data becomes differential private even before it leaves the control of the data owner, because each single individual adds noise to their data even before submitting it to the data curator. In this

situation, the data curator already holds only the differential private data, and does not need to be trusted. This is the huge advantage of the local model: data owners or individuals do not need to trust anyone but themselves.

The significant disadvantage of the local model is that the accuracy of the query results is usually much lower than with a central DP at the same privacy cost. The huge loss in accuracy means that only a small range of query types are feasible for local DP.

2.6 Deployment of Differential Privacy in Real-world Application

Differential privacy is widely-spread amongst the world's top companies, as it provides solid formal guarantees for the privacy protection of users' personal data. In the following, several real-world applications of DP are discussed.

- **Google** introduced the Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR) algorithm [EPK14] in 2014, which is among the first deployments of differential privacy. RAPPOR works on a more general setting than Randomized response, allowing users to map an arbitrary value from the domain to a $k-$length binary vector using a Bloom filter. Randomized response is then applied to the resulting binary vector, which entails a locally differential private representation.

  This response is memoized. It is a strategy in which a user re-uses a private response for repeated queries instead of introducing new randomness each time. This prevents an adversary from learning something new from subsequent disclosures and mitigates averaging attacks. RAPPOR was deployed in Google Chrome to collect browsing information and monitor aggregate statistics. It has now been replaced by a system from Google open-source libraries[1].

- **Apple** has also implemented local differential privacy system [Tea+17] along with employing various privacy methods such as delays in messages transmission, random subsampling of generated messages, de-identification as removal of IP addresses, TLS encryption and many more. The algorithms are mostly based on the Count Sketch technique proposed by Charikar et al. in their study [CCF02], which is used to count frequencies of items within a stream and is suitable for privacy preservation.

  The Apple deployment includes a feature that allows discovering new words entered by Apple device users that have not been previously in Apple's dictionary. The Apple DP paper presents a new approach called the Sequence Fragment Puzzle, which achieves this by breaking longer words into short substrings that can be enumerated. The user breaks a word into substrings and sends a randomly selected substring concatenated with a hash of the entire string, along with the starting index of the substring. The server can then identify common substring pairs and match up substrings that share the same puzzle pieces.

  Another curious application of Apple's local differential privacy method is to study the frequency of emoji usage, which can substantially vary across cultures. Apple has also applied this method to determine user preferences, such as whether customers want videos to auto-play on certain websites. In addition, Apple was able to determine what the most

[1] see https://github.com/google/differential-privacy

popular types of health data are on the Apple Health app. Since such data is sensitive, it is used to improve user experience without compromising privacy through the use of locally differentiated privacy.

- **Microsoft's** implementation of differential privacy is introduced in the work of Ding, Kulkarni and Yekhanin [DKY17]. It enables the collection and analysis of telemetry data from users' devices while safeguarding user privacy. The system offers new LDP mechanisms for the collection of counter data with formal privacy guarantees over an extended period. The paper presents a single-bit protocol for mean estimation and an alternative memoization strategy to RAPPOR. These algorithms have been implemented in Windows 10.

2.7 Combination of Anonymization and Differential Privacy

The data utility of protected results provided by differential privacy is limited because of the amount of noise that needs to be added to the result, and because data utility can only be guaranteed for a certain limited type of query. That is, the amount of noise that needs to be imposed can significantly distort the published data, which in practice reduces its utility [CT13].

In addition to the classical pure $\epsilon-$DP, some relaxations of it have been proposed in the research community (e.g. approximate $(\epsilon, \delta)-$differential privacy [DS10]; computational differential privacy [MPRV09]), in order to find a balance between the privacy budget and the data utility of the data. At the same time, there have been the proposed approaches that apply the methods of anonymity and DP together, in order to reduce the amount of noise needed to ensure DP. Anonymization methods make no assumptions about the use of anonymized data, focusing on preserving its utility from the general point of view, as they are designed for use in privacy preserving data publishing. Thus, it is possible to apply methods to anonymize the dataset and use DP to restrict the knowledge gain resulting from including one person in the dataset.

Soria-Comas et al. proposed the use of $k-$anonymization as a primary step before differential privacy in their paper [SDSM14] to help improve the utility of differential private responses to arbitrary queries. The paper presents that the amount of noise necessary to provide $\epsilon-$differential privacy can be reduced by adding the noise to a $k-$anonymous version of the dataset where $k-$anonymity is achieved using a specially designed microaggregation of all attributes.

Li, Qardaji and Su showed in their study [LQS12] that adding random sampling pre-processing to the differential private algorithm can improve privacy protection significantly and a larger privacy budget can be used. Thus, the authors claimed that $k-$anonymization, when performed "securely", with a preceding random sampling step, satisfies $(\epsilon, \delta)-$differential privacy conditions with reasonable parameters.

Soria-Comas and Domingo-Ferrer showed in their paper [SD13] that with $t-$closeness achieved by a bucketization technique, the $t-$closeness can produce $\epsilon-$DP with $t = \mathrm{e}^{\epsilon}$.

Later they stated in their following research [DS15] that $t-$closure and $\epsilon-$differential privacy have a close relation with each other when it comes to the anonymization of datasets. More specifically, $k-$anonymity for quasi-identifiers combined with $\epsilon-$differential privacy for sensitive attributes gives a stochastic $t-$closure, where t is a function of k and ϵ.

2.8 Comparison of Techniques

This section covers the strengths and weaknesses of the methods studied, $k-$anonymity, $l-$diversity, $t-$closeness [RJR17] and differential privacy. There are three types of disclosure risks commonly considered in data anonymization: Identity, Attribute and Inference Disclosure. All of the previously discussed attacks, such as unsorted matching, complementary release, homogeneity attacks, etc., can be consolidated to these three types of risks. Therefore, the comparison of the anonymization methods is conducted specifically verifying these types.

Identity Disclosure represents the determination of a person's identity with a high probability. That is, a person can be associated with a specific record in a data table. This is quite a serious type of attack, as shown earlier, and has both legal implications for data owners and identity threats against the individual themselves. Attribute Disclosure Threat represents a situation where it is possible to determine that an attribute value from a dataset belongs to a specific individual. It is important that even if it is not possible to determine an isolated record of this individual, an inference can be made about the value of his or her sensitive attribute. Inference Disclosure risk can be drawn about an individual, even if his or her data do not appear in the dataset, but rather on the basis of the statistical properties of the dataset. Table 2.5 gives an overview of the types of attacks to which the methods are vulnerable.

Technique	Identity Disclosure	Attribute Disclosure	Inference Disclosure
$k-$anonymity	✓		
$l-$diversity	✓	✓	
$t-$closeness	✓	✓	✓
Differential privacy	✓	✓	✓

Table 2.5: Comparison of Privacy Preserving Techniques.

$k-$anonymity:

- Pros:
  - protects against identity disclosure by ensuring that there are no less than k similar values in a dataset, preventing the adversary from connecting sensitive data with external data.
  - there are several widely used algorithms: Datafly [Swe00], Incognito [LDR05] and Mondrian [LDR06a]
- Cons:
  - computationally expensive: the naive algorithm is $O(n^2)$. Optimal generalization is extremely difficult, and outliers can make it even more challenging. Solving this problem automatically is NP-hard.
  - has several limitations including unsorted matching, complementary release and temporal attacks and is prone to homogeneity and background knowledge attacks.
  - can result in high loss of utility in high dimensional data.
  - precautions are necessary if the data has already undergone multiple rounds of anonymization.

$l-$diversity:

- Pros:
  - enhances data protection by promoting a more diverse distribution of sensitive attributes within equivalence class.
  - improves the protection against attribute disclosure compared to the $k-$anonymity technique.
- Cons:
  - is prone to skewness and similarity attacks.
  - can be redundant and labor-intensive to implement.

$t-$closeness:

- Pros:
  - provides protection against attribute disclosure.
  - improves upon the limitations of $k-$anonymity.
  - identifies the semantic proximity of attributes, which is a weakness of the $l-$diversity.
- Cons:
  - EMD application in $t-$closeness makes it difficult to determine the proximity between the $t-$value and the acquired knowledge.
  - to be effective, it requires that the distribution of the sensitive attribute in the equivalence class is similar to that in the entire dataset.

Differential privacy:

- Pros:
  - offers a solution to all the challenges of the anonymization methods by adding noise.
  - provides a formal privacy guarantee for statistical databases.
  - guarantees and quantifies how much risk an individual faces when publishing data in a statistical database.
  - does not represent a privacy issue when updating the database, as long as the attribute distribution does not change.
- Cons:
  - is strongly related to database queries, and is not suitable for publishing a generalized table.
  - privacy budget parameter epsilon must be specified.
  - more complex queries require more advanced methods of proof.
  - balance between data utility and data privacy is still challenging.

3 Anonymization Algorithms

A number of various anonymization models using the notion of $k-$anonymity in its general idea have been proposed in literature. In this section commonly used and widely cited $k-$anonymization algorithms using generalization and suppression, but various anonymization strategies are studied. The algorithm flowcharts presented below were inspired by the work [AMCM+14].

3.1 Datafly

Sweeney proposed a greedy heuristic algorithm Datafly [Swe01; Swe02a], which performs single-dimensional full-domain generalization. Figure 3.1 shows the flowchart of the algorithm. The basic idea of the algorithm is to provide a certain size of equivalence class or to estimate the frequency of tuples with respect to quasi-attributes by means of generalization and further suppression. If $k-$anonymity is not satisfied, the algorithm generalizes the attribute that has the widest domain size in the table for the next iteration. As long as $k-$anonymity is not achieved, a further process of generalization and suppression is applied again.

Although this algorithm guarantee a $k-$anonymity representation, it does not provide $k-$minimal generalization. In addition, the heuristic that selects an attribute with the largest number of different values as the next to generalize is computationally efficient, although it can be the case that this strategy produces unnecessary generalization.

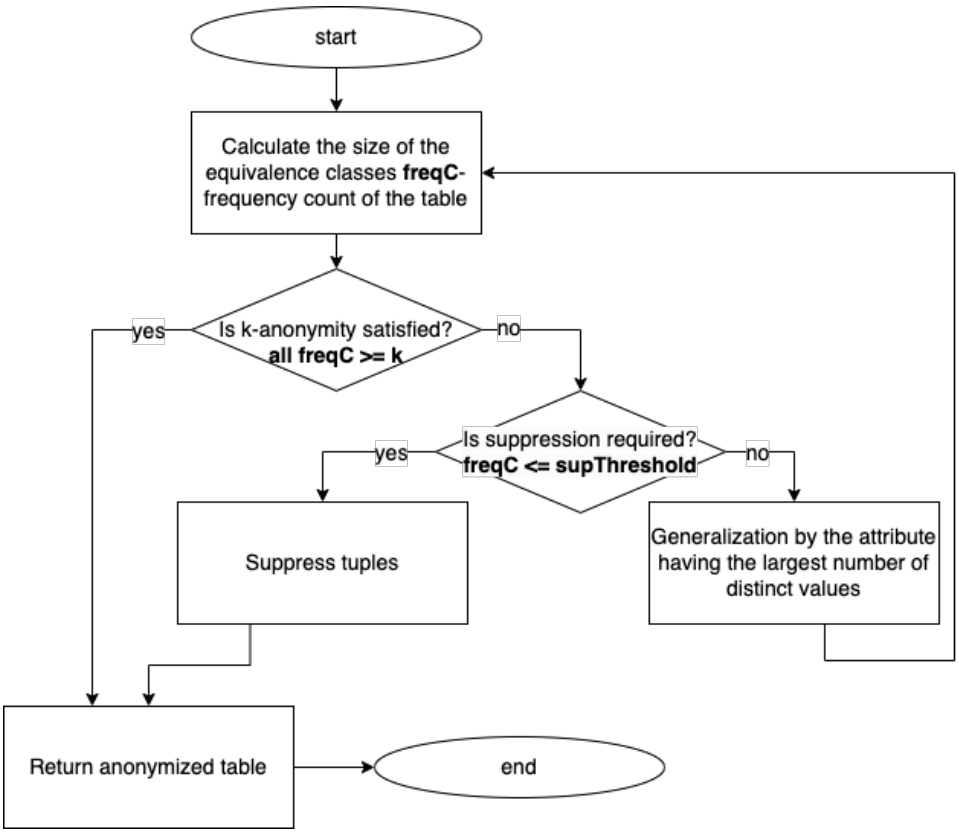


Figure 3.1: Flowchart of Datafly Algorithm.

3.2 Incognito

Incognito [LDR05] algorithm, developed by LeFevre, DeWitt and Ramakrishnan, represents a single-dimensional full-domain generalization approach. A simplified flow diagram of the algorithm is illustrated in Figure 3.2. The main challenge of the algorithm is to construct a multi-attribute lattice of generalizations, traverse it using bottom-up breadth-first search and build a generalized table for each node of the grid. The lattice is constructed for a combination of each quasi-attribute and is determined by the depth of its hierarchy tree.

Incognito prunes the generalization of some nodes to reduce the search space based on a priori property. The algorithm uses the rollup property to calculate the size of the equivalence class. This means that the size of the parent group can be directly computed from the sum of the sizes of all child classes, with the implication that the group sizes of all possible generalizations can be computed from the bottom up incrementally. Due to this property, it is not only possible to compute group sizes efficiently, but also to provide a termination condition for further generalizations. When a node satisfies $k-$anonymity, all its direct generalizations can be pruned, as it is guaranteed that they also meet $k-$anonymity property.

Unlike Datafly, Incognito outputs an optimal $k-$anonymious solution. It means that, the anonymized solution contains the maximum amount of information according to the chosen information metric. The anonymized table is drawn from a collection of solutions among the generalization lattice that satisfies a given privacy requirement.

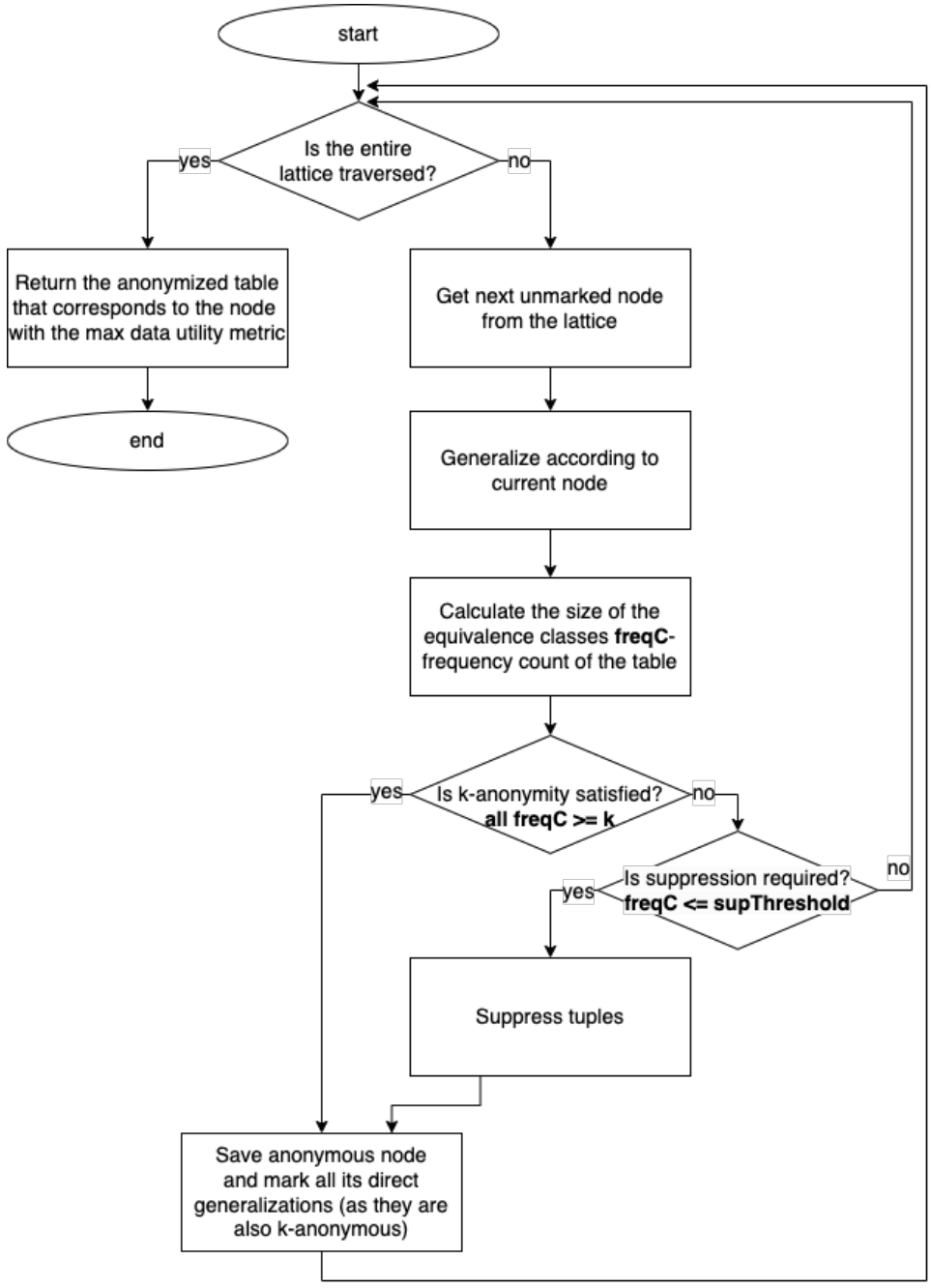


Figure 3.2: Flowchart of Incognito Algorithm.

3.3 Mondrian

Another widely used algorithm is Mondrian [LDR06a], which was later proposed by the authors of Incognito. It is named after paintings of the artist Piet Mondrian and is a greedy multidimensional algorithm. The term "multidimensional" means that a quasi-attribute in an anonymized table can have different levels of generalization in different equivalence classes. An example is depicted in

Tables 3.1, 3.2. The algorithm recursively partitions the attribute domain space into several regions, each containing at least k entries. Figure 3.3 depicts the flowchart of the algorithm.

Age	Gender	ZIP	Diagnosis
25	Male	53711	Flu
25	Female	53712	Hepatitis
26	Male	53711	Bronchitis
27	Male	53710	Stroke
27	Female	53712	AIDS
28	Male	53711	Rosacea

Table 3.1: Illustration of Generalization Models: Original Table.

Single-dimensional			Multidimensional			
Age	Gender	ZIP	Age	Gender	ZIP	Diagnosis
[25-28]	Male	[53710-53711]	[25-26]	Male	53711	Flu
[25-27]	Female	53712	[25-28]	Female	53712	Hepatitis
[25-26]	Male	53711	[25-28]	Male	[53710-53711]	Bronchitis
[27-28]	Male	[53710-53711]	[25-28]	Male	[53710-53711]	Stroke
[25-27]	Female	53712	[25-28]	Female	53712	AIDS
[27-28]	Male	[53710-53711]	[25-28]	Male	[53710-53711]	Rosacea

Table 3.2: Comparison of Single- and Multidimensional Anonymization.

The most generalized attribute value within the set of quasi-identifiers is selected initially and than refined as the data is partitioned. The algorithm finds the attribute with the broadest domain of values and performs partitioning according to selected attribute using the median partitioning approach. To calculate the median of an attribute, the frequency sets approach is used. That is, occurrences of each unique attribute value in the partition are calculated and the median is taken. The result is a split value on which the partitioning has been executed.

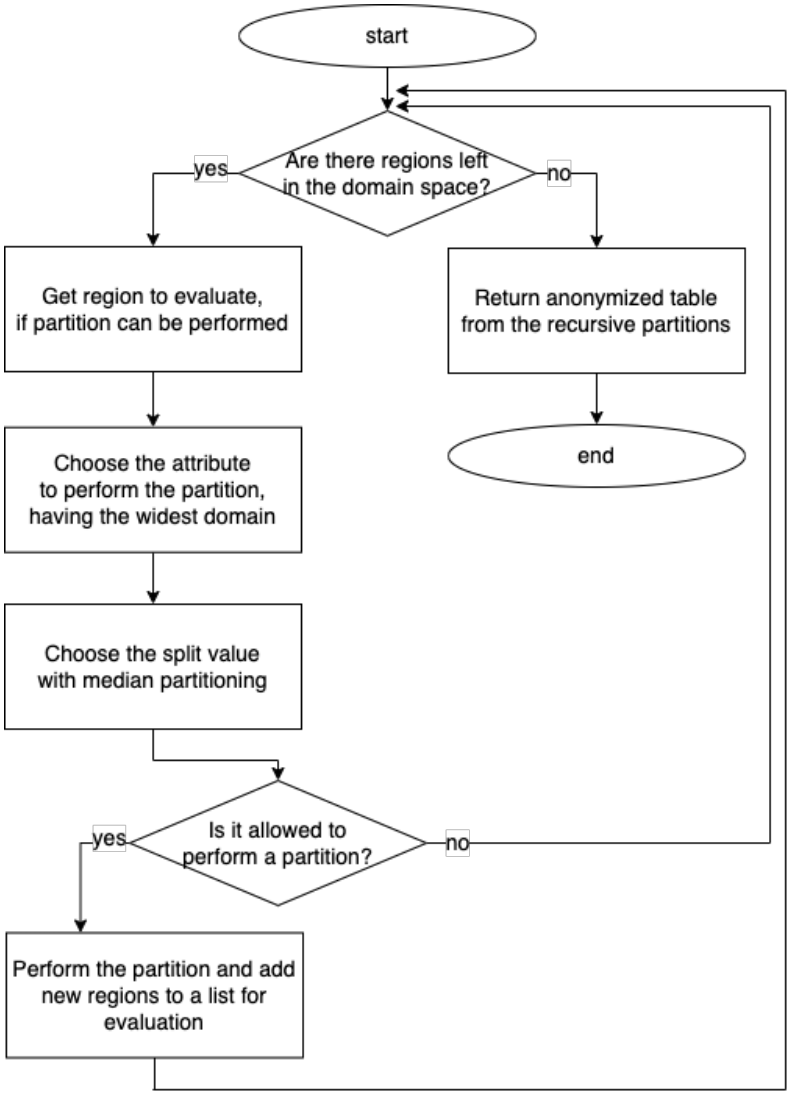


Figure 3.3: Flowchart of Mondrian Algorithm.

3.4 Top-Down Greedy

Two greedy algorithms have been proposed by Xu et al. in their research paper [XWP+06], bottom-up and top-down. The bottom-up algorithm focuses on grouping records locally according to a given utility metric. In the beginning, each record is treated as a separate group. Then iteratively, groups smaller than k are merged, assuming that the joined group now has a larger utility score. The size of the group can become larger than k. However, to avoid overgeneralization, when a group contains more than $2k$ records, such a group must be separated. Bottom-up algorithm does not partition the domain, so different groups may have overlapping ranges. Moreover, records with identical quasi-identifier can be split into different groups. In addition, the algorithm uses two successive loops to search for the neighbouring records, making it slower than its successor.

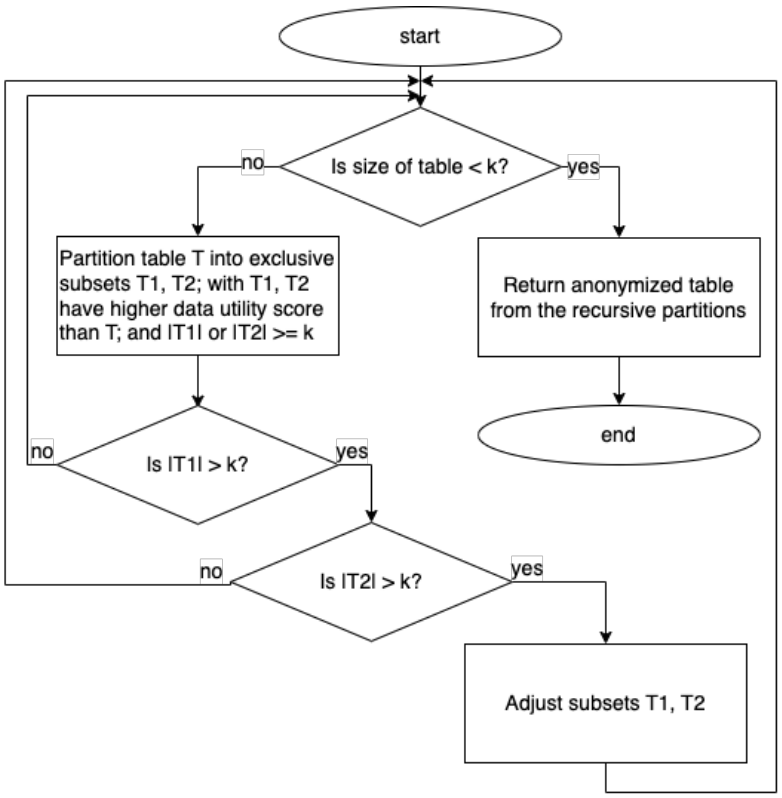


Figure 3.4: Flowchart of Top-down Greedy Algorithm.

Top-down algorithm improves these shortcomings. Figure 3.4 shows the flowchart of the algorithm. This algorithm, as the name implies, splits the initial table iteratively into subsets, when a newly formed subset is more local, that is, it has higher data utility scores. Binary partitioning is used to split the group into two subsets in each iteration. If the resulting group contains k or more tuples, the partitioning is performed recursively for this group. If the group has less than k entries, a corrective post-processing step is employed.

3.5 Comparison of Algorithms

The algorithms considered above are the most cited in research papers on $k-$anonymization of data. These methods are equally widely applied, that is why it is impossible to unambiguously answer, which one ensures data anonymity better. In fact, it depends on the initial dataset, the number of quasi-attributes, the size of the hierarchy tree for the attributes, as well as the required scenario of the anonymization system. In the following, some conclusions drawn from the usage of Datafly, Incognito and Mondrian algorithms based on the experiments discribed be Ayala-Rivera et al. in their research paper [AMCM+14] are discussed. Because the proposed algorithm is inspired by the beneficial properties of the standard algorithms, special focus in the comparison is given to characteristics such as computational complexity, generalization model, handling of different types of attributes, and requiring predefined hierarchical trees.

- The key step in Incognito and Datafly algorithms is applying full- domain generalization to all quasi-identifiers. Thus, the data curator needs prior domain knowledge to define the generalization specifications. In this case, when the generalization hierarchy for each quasi-attribute is pre-defined, the anonymized dataset conforms to the constraints defined in the generalization hierarchies.
  On the contrary, Mondrian algorithm uses the concept of median partitioning to perform generalization. This can be thought of as unsupervised anonymization, where the creation of partitions occurs dynamically without taking into account hierarchies of attribute values.

- Since Mondrian algorithm employs auto-partitioning of the attribute domain to create the value ranges, this algorithm works better with numerical attributes, which is shown in Experimental chapter 8. The reason for this is that in the case of categorical attributes, it is possible to confuse any semantic association with attribute values, since the data is combined into equivalence classes with no control over the actual semantic groups, e.g. grouping countries belonging to the same continent.

- The runtime complexity of Incognito algorithm is $O(2^{|QI|})$ - exponential to the number of quasi-identifiers (QI). As a result, the algorithm performs better when the number of quasi-identifiers is not large. In addition, Incognito requires an increased investment of time and memory when quasi-identifiers have deep hierarchies, because the search space in the generalization lattice is broader. It is related to the time it takes to traverse the lattice and verify $k-$anonymity in each individual node.
  Whereas, the time complexity of Mondrian is $O(n \log n)$ and $O(|QI| \cdot n \log n)$ for Datafly, where n is the number of tuples in the original table.

- The median partitioning method tries to get an equidistant fill. Hence, Mondrian results in a higher information loss when the data is skewed. Therefore, it works better with uniform distributions, because most QI can be used to partition the data. Incognito and Datafly algorithms work optimal for both uniformly and non-uniformly distributed data.

- Mondrian algorithm partitions the data using a single attribute instead of the multiple attributes present in the QI set. This causes the remaining attributes in the QI set to keep their least specific values and, thus, a high penalty for these attributes.

4 Differential Privacy

The aim of this chapter is to give an introduction on differential privacy (DP), study some methods of pertubation and anonymization of the initial data in more detailed manner. Special focus is on mathematical properties and composition of DP and the basic methods for local and central models. At the end of the chapter the comparison of the DP-mechanisms is given.

Differential privacy is more applicable in Privacy preserving data mining setting. DP can be used in two ways: to answer queries in interactive setting or to produce a sanitized dataset for release in non-interactive setting. In the interactive scenario, DP adds noise to query results to obscure sensitive information, but limits the analysis that can be performed, because only a small number of query types can be used. In the non-interactive scenario aggregated statistics are produced based on count queries, contingency tables [BCD+07] or histograms [XZX+13]. The properties and methods discussed below apply to both scenarios.

4.1 Properties of Differential Privacy

This chapter presents important properties of differentially private mechanisms, which are derived from the definition of differential privacy. These properties are useful in designing algorithms that satisfy differential privacy, in order to help improve qualitative parameters of these algorithms and conduct their privacy analysis. Theorems and proofs are taken and adapted from the research studies [DR+14; Vad17].

4.1.1 Post-processing

Differential private algorithms are closed under post-processing. It means that once the data is privatised, it cannot be less privatised or de-privatised, if the private mechanism is applied again to the same dataset.

Theorem 1. Post-processing. Let $\mathcal{M} : \mathbb{N}^{|\mathcal{X}|} \to Y$ be $\epsilon-$differential private algorithm and let $\mathcal{F} : Y \to Z$ be an arbitrary randomized function. Then $\mathcal{F} \circ \mathcal{M}$ is also $\epsilon-$differential private.

Proof. Let $\mathcal{F}$ be a randomized function and consider it to be a distribution over deterministic function f. The privacy proof follows for every neighbouring datasets D_1, D_2, which differ only in one record, and $T \subseteq Z$:

$$\begin{aligned}\Pr[\,\mathcal{F}(\mathcal{M}(D_1)) \in T\,] &= \mathbb{E}_{f \leftarrow \mathcal{F}}[\,\Pr[\,\mathcal{M}(D_1) \in f^{-1}(T)\,]] \\ &\leq \mathbb{E}_{f \leftarrow \mathcal{F}}[\,\exp(\epsilon)\Pr[\,\mathcal{M}(D_2) \in f^{-1}(T)\,]] \\ &= \exp(\epsilon)\Pr[\,\mathcal{F}(\mathcal{M}(D_2)) \in T\,].\end{aligned}$$

4.1.2 Group Privacy

A further useful property is, that differential privacy provides protection for small groups of individuals. For datasets, differing by multiple entries, the definition of differential privacy allows ensuring a smooth decrease, as the distance between the datasets increases. The Hemming distance $d(x, x')$, for $x, x' \in X^n$, in other words, the number of rows that need to be changed to go from x to x', can be used as an example of the distance metric.

Theorem 2. Group privacy. Let $\mathcal{M} : \mathbb{N}^{|\mathcal{X}|} \to Y$ be $\epsilon-$differential private algorithm. Suppose D_0, D_k are two datasets which differ in exactly k positions. Then for all $T \subseteq Y$, we have

$$\Pr[\mathcal{M}(D_0) \in T] \leq \exp(k\epsilon) \Pr[\mathcal{M}(D_k) \in T].$$

Proof. Let's apply a hybrid argument. Let $x_0, x_1, ..., x_k$ be datasets, in which x_0 and x_k differ in exactly k positions. Then for all $T \subseteq Y$:

$$\begin{aligned}
\Pr[\mathcal{M}(x_0) \in T] &\leq \exp(\epsilon) \Pr[\mathcal{M}(x_1) \in T] \\
&\leq \exp(\epsilon) \Big(\exp(\epsilon) \Pr[\mathcal{M}(x_2) \in T]\Big) \\
&... \\
&\leq \exp(k\epsilon) \Pr[\mathcal{M}(x_k) \in T]
\end{aligned}$$

4.1.3 Composition

Composition refers to the execution of multiple requests to the same dataset. Furthermore, it is essential for calculating the total privacy cost, when several separate analyses are performed on the data. The privacy cost limit given by composition is an upper bound, namely, the actual privacy cost of running the differentially private algorithms may be lower than this limit, but never higher [DMNS06].

The subject of the next part is to discuss how the privacy of an algorithm changes, when several differentially private algorithms are applied sequentially or in parallel to the same dataset.

Sequential composition

A sequence of differentially private mechanisms during computation is executed on the same database. If the mechanisms are run independently or the subsequent mechanism uses the results of the proceeding one, the overall algorithm is also differential private [DKM+06].

Theorem 3. Sequential Composition. Let $\mathcal{M}_i : \mathbb{N}^{|\mathcal{X}|} \to Y_i$ be ϵ_i-differential private mechanism for each $i \in \mathbb{N}$. Their sequence $\mathcal{M}_{1,...,n} : \mathbb{N}^{|\mathcal{X}|} \to Y_1 \times ... \times Y_n$, defined by the mapping: $\mathcal{M}_{1,...,n}(y) = (\mathcal{M}_1(y), ..., \mathcal{M}_n(y))$, is $\sum_i^n \epsilon_i-$differential private.

Proof. Assume two neighbouring datasets D_1, D_2 and consider some output $(y_1, y_2) \in Y_1 \times Y_2$, then:

$$\begin{aligned}\frac{\Pr[\mathcal{M}_{1,2}(D_1) = (y_1, y_2)]}{\Pr[\mathcal{M}_{1,2}(D_2) = (y_1, y_2)]} &= \frac{\Pr[\mathcal{M}_1(D_1) = y_1]\Pr[\mathcal{M}_2(D_1) = y_2]}{\Pr[\mathcal{M}_1(D_2) = y_1]\Pr[\mathcal{M}_2(D_2) = y_2]} \\ &= \left(\frac{\Pr[\mathcal{M}_1(D_1) = y_1]}{\Pr[\mathcal{M}_1(D_2) = y_1]}\right)\left(\frac{\Pr[\mathcal{M}_2(D_1) = y_2]}{\Pr[\mathcal{M}_2(D_2) = y_2]}\right) \\ &\leq \exp(\epsilon_1)\exp(\epsilon_2) \\ &= \exp(\epsilon_1 + \epsilon_2)\end{aligned}$$

Parallel Composition

A series of computations with differentially private mechanisms are applied to disjoint sets of the dataset in parallel. In this case, the total cost of privacy is not summed, but depends on the worst privacy guarantee among the all sub-mechanisms [McS09].

Theorem 4. Parallel composition. Let $\mathcal{M}_i : \mathbb{N}^{|\mathcal{X}|} \to Y_i$ be ϵ_i-differential private mechanism applied to the disjoint subset D^i from the dataset D for $i \in \mathbb{N}$. Application of any function $f(y_1, ...y_n)$, that combines the outputs of the mechanisms, is $\max_i \epsilon_i-$differential private.

Proof. Assume two neighbouring datasets D_1, D_2 are fixed and D_1^i, D_2^i be their $i-$th disjoint subset. As D_1, D_2 differ in one element, let this element be in the $j-$th subset. Then, $D_1^j \neq D_2^j$ and $D_1^i = D_2^i$ for all $i \neq j$. Suppose $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_k)$ is a sequence of ϵ_i-differential private algorithms with $\mathcal{M}_i : \mathbb{N}^{|\mathcal{X}|} \to Y_i$, where $y_i \in Y_i$ is a sequence of results and its probability is $\Pr[\mathcal{M}(D) = y] = \prod_i \Pr[\mathcal{M}_i(D^i) = y_i]$. Then:

$$\begin{aligned}\Pr[\,\mathcal{M}(D_1) = y\,] &= \prod_i \Pr[\,\mathcal{M}_i(D_1^i) = y_i\,] \\ &= \Pr[\,\mathcal{M}_j(D_1^j) = y_j\,] \prod_{i \neq j} \Pr[\,\mathcal{M}_i(D_1^i) = y_i] \\ &= \Pr[\,\mathcal{M}_j(D_1^j) = y_j\,] \prod_{i \neq j} \Pr[\,\mathcal{M}_i(D_2^i) = y_i\,] \\ &\leq \exp(\epsilon_j)\Pr[\,\mathcal{M}_j(D_2^j) = y_j] \prod_{i \neq j} \Pr[\,\mathcal{M}_i(D_2^i) = y_i\,] \\ &\leq \exp(\max_i \epsilon_i) \prod_i \Pr[\,\mathcal{M}_i(D_2^i) = y_i\,] \\ &= \exp(\max_i \epsilon_i)\Pr[\,\mathcal{M}(D_2) = y\,]\end{aligned}$$

4.2 Methods of Differential Privacy

This section describes the differential privacy mechanisms that form the basic building blocks for other DP application algorithms.

4.2.1 Randomized Response

The Randomized response [War65] was developed in order to assist in the data collection for social surveys on sensitive topics and allows the data curator to collect differentially private responses to a single binary (“yes”/“no”) question in embarrassing or illegal behavior situation. The idea behind Randomized response is to add randomness to the response so that the actual response is masked, making it more difficult for an outside party/ adversary to determine the true answer.

In this setting, each participant is given a Randomized response mechanism, such as a coin, to determine their answer. This ensures, that even if an adversary is able to observe the responses, they will not be able to determine the true response with certainty.

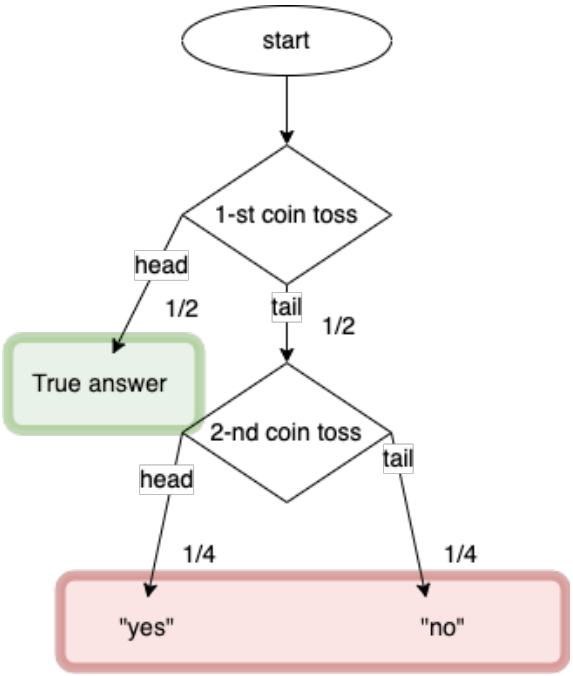


Figure 4.1: Flowchart of the Randomized Response Mechanism.

The idea of the mechanism depicted in Figure 4.1 is the following:

- Flip a coin. If the result is head, a true answer is taken.
- If the result is tail, second coin is flipped. If the result of the second coin is a head, the answer is “yes”. Otherwise, the answer is “no”.

Confidentiality in this method is provided by the second case, where the individual responds at random. The Randomized response provides the so-called “plausible deniability”.

Theorem 5. The Randomized response satisfies ϵ-differential privacy for $\epsilon = \log(3) = 1.09$ [DR+14].

Proof. In a case analysis, the true response is "Yes" and the outcome will be "Yes", if the first coin comes up heads (with probability of $\frac{1}{2}$) or the first comes up tails and second - heads (with probability of $\frac{1}{4}$):

$$\Pr[\text{ Response = Yes} \mid \text{Truth = Yes }] = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

In another case, $\Pr[\text{ Response = Yes} \mid \text{Truth = No }] = \frac{1}{4}$, when first and second coins come up tails with probability of $\frac{1}{4}$.

By applying similar reasoning to the case of a "No" answer, we get:

$$\frac{\Pr[\text{ Response = Yes} \mid \text{Truth = Yes }]}{\Pr[\text{ Response = Yes} \mid \text{Truth = No }]} = \frac{\frac{3}{4}}{\frac{1}{4}} = 3 = \frac{\Pr[\text{ Response = No} \mid \text{Truth = No }]}{\Pr[\text{ Response = No} \mid \text{Truth = Yes }]}$$

Randomized response can be applied to binary type of sensitive data. The technique has been shown to effectively preserve the privacy of participants in survey research, while still allowing for valuable insights to be obtained.

However, Randomized response does come with some limitations, such as reducing the accuracy of the data and limiting the analysis, that can be performed. Nevertheless, it remains one of the few DP methods, that is applicable to the local DP model, where each individual protects their data internally and there is no need for a trusted data curator.

4.2.2 Unary Encoding

The Randomized response method handles binary data types. In the local differential privacy model, several different algorithms have been proposed to handle different types of data. These algorithms focus on aggregate statistics, estimation of frequency as well as histograms. An Unary encoding approach [WBLJ17] is a basis for the Basic One-time RAPPOR [EPK14].

First, the domain of response values from the bins of the histogram is determined. The encoding process of this method is also called one-hot encoding in the machine learning community. The values are then perturbed. The bits in the response vector are flipped to ensure differential privacy, according to the equation:

$$\Pr[\, B'[i] = 1 \,] = \begin{cases} p & \text{if } B[i] = 1, \\ q & \text{if } B[i] = 0 \end{cases}$$

The probability parameters p and q are introduced, which together define the epsilon ϵ privacy parameter.

$$p = \frac{\exp(\epsilon/2)}{\exp(\epsilon/2) + 1} \qquad q = \frac{1}{\exp(\epsilon/2) + 1}$$

Then, for any v_1 and v_2 of the value domain, and the output B, it holds:

$$\frac{\Pr[\, B \mid v_1 \,]}{\Pr[\, B \mid v_2 \,]} = \frac{\prod_i \Pr[\, B[i] \mid v_1 \,]}{\prod_i \Pr[\, B[i] \mid v_2 \,]} \leq \frac{\Pr[\, B[v_1] = 1 \mid v_1 \,] \Pr[\, B[v_2] = 0 \mid v_1 \,]}{\Pr[\, B[v_1] = 1 \mid v_2 \,] \Pr[\, B[v_2] = 0 \mid v_2 \,]} = \frac{p(1-q)}{q(1-p)} \leq \exp(\epsilon)$$

An aggregation of the perturbed responses is then performed, which takes into account the number of "fake" responses in each category:

$$A[i] = \frac{\sum_j B'_j[i] - nq}{p - q}$$

Methods such as Randomized response or Unary encoding provide sufficiently accurate results to produce a rough frequency estimation of domain elements. However, they are still less accurate by an order of magnitude than those obtained using the Laplace mechanism in the central differential privacy model, which is discussed next.

4.2.3 The Laplace Mechanism

The Laplace mechanism [DMNS06] computes the function f and perturbs each coordinate with a noise taken from the Laplace distribution.

Definition. The Laplace distribution with center at 0, scale b and the variance $\sigma^2 = 2b^2$ is the distribution with probability density function:

$$\text{Lap}\,(\,x \mid b\,) = \frac{1}{2b} e^{\left(-\frac{|x|}{b}\right)}$$

The scale of added noise is calibrated to the sensitivity of the function f.

Definition. The l_1-sensitivity of a function f measures the largest variation between function outputs for various inputs. Δf indicates the change in the worst case, under the presence or absence of a single individual record. On the other hand, l_1-sensitivity is the amount of uncertainty in the output, which is entered to hide the participation of a given individual. That is, it is an upper bound on how much the output of the function must be perturbed in order to provide the necessary privacy level.

$$\Delta f = \max_{x,y \in \mathcal{N}, \|x-y\|_1=1} \|f(x) - f(y)\|_1$$

Definition. The Laplace mechanism computes the function $f(\cdot)$ and perturbs each coordinate with a noise taken from the Laplace distribution. The numeric functions $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$, which map database to k real numbers, are widely used in the data analysis. The mechanism is given:

$$\mathcal{M}_L(x, f(\cdot), \epsilon) = f(x) + (Y_1, ..., Y_k),$$

where $Y_i, i \in \mathcal{N}$ are independent and identically distributed random variables drawn from Lap(b), where the scale of the noise needs to be calibrated to $b = \frac{\Delta f}{\epsilon}$.

Theorem 6. The Laplace mechanism preserves $\epsilon-$differential privacy.

Proof. Let D_1 and $D_2 \in \mathbb{N}^{|\mathcal{X}|}$ be the neighbouring datasets, $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$ be a numeric function, p_D be a probability density function of $\mathcal{M}_L(D, f, \epsilon)$. We compare the two probability density functions for both datasets at some arbitrary point $z \in \mathbb{R}^k$.

$$
\begin{aligned}
\frac{p_{\mathrm{D}_1}(z)}{p_{\mathrm{D}_2}(z)} &= \frac{\prod_{i=1}^{k}\left(\frac{1}{2\frac{\Delta f}{\epsilon}}\exp\frac{-|f(\mathrm{D}_1)_i - z_i|}{\frac{\Delta f}{\epsilon}}\right)}{\prod_{i=1}^{k}\left(\frac{1}{2\frac{\Delta f}{\epsilon}}\exp\frac{-|f(\mathrm{D}_2)_i - z_i|}{\frac{\Delta f}{\epsilon}}\right)} \\
&= \prod_{i=1}^{k}\left(\frac{\exp\left(-\frac{\epsilon|f(\mathrm{D}_1)_i - z_i|}{\Delta f}\right)}{\exp\left(-\frac{\epsilon|f(\mathrm{D}_2)_i - z_i|}{\Delta f}\right)}\right) \\
&= \prod_{i=1}^{k}\left(\frac{\epsilon|f(\mathrm{D}_2)_i - z_i| - |f(\mathrm{D}_1)_i - z_i|}{\Delta f}\right) \\
&\leq \prod_{i=1}^{k}\left(\frac{\epsilon|f(\mathrm{D}_1)_i - f(\mathrm{D}_2)_i|}{\Delta f}\right) \\
&= \exp\left(\frac{\epsilon \cdot \|f(\mathrm{D}_1) - f(\mathrm{D}_2)\|_1}{\Delta f}\right) \\
&\leq \exp(\epsilon)
\end{aligned}
$$

4.2.4 Approximate Differential Privacy

Taking into account that in order to guarantee pure differential privacy it is necessary to trade-off the decreased utility of the query results or the increased complexity of the algorithm, some relaxation variations of DP have been developed. One of them is $(\epsilon, \delta)-$differential privacy [DS10].

Definition. $(\epsilon, \delta)-$**differential privacy.** A randomized function $\mathcal{M}$ gives $(\epsilon, \delta)-$differential privacy if for all neighbouring datasets D_1 and D_2, and all subsets $Y \subseteq \mathrm{Range}\,(\mathcal{M})$, where $\mathrm{Range}(\mathcal{M})$ is the set of possible outputs of the randomized function $\mathcal{M}$ holds

$$\Pr[\mathcal{M}(D_1) \in Y] \leq e^{\epsilon} \times \Pr[\mathcal{M}(D_2) \in Y] + \delta.$$

Pure differential privacy is equivalent to approximate differential privacy when $\delta = 0$. The relaxation of the method presupposes that the parameter δ can be thought of as the probability of failure. The privacy loss can be bounded by ϵ with probability at least $(1 - \delta)$. As δ is an additive term and ϵ is multiplicative, δ should typically be much smaller than ϵ. For example, when $\delta \geq 1/|D|$, the privacy can easily be breached, because, the mechanism can query single individual record in a database D and simply release it without any further randomisation. The privacy of each record in the database is assumed with probability δ.

Another positive aspect of approximate differential privacy is that it provides much better compositional properties than pure differential privacy. In particular, composing $k(\epsilon, \delta)-$differential private algorithms roughly results in $(\sqrt{k}\epsilon, k\delta)-$differential privacy.

4.2.5 The Gaussian Mechanism

The alternative to the Laplace mechanism is the Gaussian mechanism, which appends Gaussian noise instead of Laplace noise to each coordinate.

Definition. The Gaussian (normal) distribution with center at 0 and variance σ^2 has the distribution with probability density function:

$$\mathrm{N}(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Definition. The Gaussian mechanism computes the function f and perturbs each coordinate with a noise taken from the Gaussian distribution. $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$ is the numeric functions, which map database to k real numbers.

$$\mathcal{M}_G(x, f(\cdot), \epsilon) = f(x) + (Y_1, ..., Y_k),$$

where $Y_i, i \in \mathcal{N}$ are independent and identically distributed random variables drawn from $\mathcal{N}(\sigma^2)$. The Gaussian mechanism does not satisfy pure $\epsilon-$differential privacy, but it does satisfy approximate $(\epsilon, \delta)-$differential privacy.

Theorem 7. The Gaussian Mechanism with parameter $\sigma = \frac{2\delta_2(f)^2 \log(1.25/\delta)}{\epsilon^2}$ is $(\epsilon, \delta)-$differential private, where $\delta_2(f)$ is the l_2-sensitivity. Please, refer to the work of Dwork and Roth [DR+14] for proof.

4.2.6 The Exponential Mechanism

Fundamental mechanisms like Laplace and Gaussian are applicable to numerical functions and add noise directly to the response. The exponential mechanism was designed by McSherry and Talwar [MT07] to select the most appropriate response, in a situation where adding noise directly to a calculated quantity could completely destroy its value. The Laplace distribution, by the way, is a symmetric version of the exponential distribution.

The exponential mechanism determines which element is best by means of some utility function that maps the score to each element in the output set, and also determines the set of elements to be chosen from. Namely, the mechanism takes out some element from the set with the highest possible score. However, at the same time, it is worth noting that the mechanism provides differential privacy while still approximately maximizing the element's score. That is, sometimes the element, that does not have the highest score, is returned from the set to guarantee confidentiality.

Definition. The exponential mechanism. Mechanism $\mathcal{M}_E(D, u, C)$ selects and outputs an element $c \in C$ using scoring function $u : \mathbb{N}^{|\mathcal{X}|} \times C \to \mathbb{R}$ with probability proportional to $\exp\left(\frac{\epsilon u(D,c)}{2\Delta u}\right)$.

The idea of the exponential mechanism consists of three blocks:

- Select a set C of possible outputs.
- Specify a scoring function $u : \mathbb{N}^{|\mathcal{X}|} \times C \to \mathbb{R}$, which maps database and output pairs to utility scores. The value of $u(D, c)$ indicates the matching grade of output c for database D. The sensitivity of the scoring function Δu is measured with respect to the database argument.

$$\Delta u = \max_{c \in C} \max_{D_1, D_2 : \|D_1 - D_2\|_1 \leq 1} |u(D_1, c) - u(D_2, c)|.$$

- Output each possible $c \in C$ with probability proportional to $\exp\left(\frac{\epsilon\, u(D,c)}{2\,\Delta u}\right)$.

Using the exponential mechanism, it is possible to respond to queries with items from the analyzed data, where a probability is determined for each item. Items with a high score are exponentially more likely than responses with a low score. The magnitude of the probability depends on the sensitivity of the score function and the privacy budget, similar to the previous mechanisms. The privacy loss of the mechanism is :

$$\log\left(\frac{\frac{\exp(\epsilon u(D_1,c)}{2\Delta u}}{\frac{\exp(\epsilon u(D_2,c)}{2\Delta u}}\right) = \frac{\epsilon\,(u(D_1,c) - u(D_2,c))}{2\Delta u} \leq \epsilon.$$

Theorem 8. The exponential mechanism preserves $\epsilon-$differential privacy.

Proof.

$$\begin{aligned}
\frac{\Pr[\mathcal{M}_E(D_1,u,C)=c]}{\Pr[\mathcal{M}_E(D_2,u,C)=c]} &= \frac{\frac{\exp\left(\frac{\epsilon u(D_1,c)}{2\Delta u}\right)}{\sum_{c'\in C}\exp\left(\frac{\epsilon u(D_1,c')}{2\Delta u}\right)}}{\frac{\exp\left(\frac{\epsilon u(D_2,c)}{2\Delta u}\right)}{\sum_{c'\in C}\exp\left(\frac{\epsilon u(D_2,c')}{2\Delta u}\right)}} \\
&= \left(\frac{\exp\left(\frac{\epsilon u(D_1,c)}{2\Delta u}\right)}{\exp\left(\frac{\epsilon u(D_2,c)}{2\Delta u}\right)}\right)\left(\frac{\sum_{c'\in C}\exp\left(\frac{\epsilon u(D_2,c')}{2\Delta u}\right)}{\sum_{c'\in C}\exp\left(\frac{\epsilon u(D_1,c')}{2\Delta u}\right)}\right) \\
&= \exp\left(\frac{\epsilon(u(D_1,c') - u(D_2,c'))}{2\Delta u}\right)\left(\frac{\sum_{c'\in C}\exp\left(\frac{\epsilon u(D_2,c')}{2\Delta u}\right)}{\sum_{c'\in C}\exp\left(\frac{\epsilon u(D_1,c')}{2\Delta u}\right)}\right) \\
&\leq \exp\left(\frac{\epsilon}{2}\right)\exp\left(\frac{\epsilon}{2}\right)\left(\frac{\sum_{c'\in C}\exp\left(\frac{\epsilon u(D_1,c')}{2\Delta u}\right)}{\sum_{c'\in C}\exp\left(\frac{\epsilon u(D_2,c')}{2\Delta u}\right)}\right) \\
&= \exp(\epsilon)
\end{aligned}$$

4.2.7 Comparison of Basic Differentially Private Mechanisms

This chapter provides a comparison of mechanisms that guarantee differential privacy. A discussion is given on the following differential privacy mechanisms: randomised response, Laplace, Gaussian and exponential mechanism. Important parameters for analysis are the level of privacy, the accuracy of the algorithm, as well as the data types for which it is suitable.

Mechanism	Privacy level	Accuracy depends on	Data type application
Randomized response	$\log(3)-$DP	n	Binary queries
Laplace	$\epsilon-$DP	ϵ, Δf of query	Numerical queries
Gaussian	$(\epsilon,\delta)-$DP	ϵ, Δf of query	Numerical queries
Exponential	$\epsilon-$DP	ϵ, Δu of score function	Non-numerical queries

Table 4.1: Comparison of Differentially Private Mechanisms.

In summary, the Laplace and exponential mechanisms achieve $\epsilon-$differential privacy at any privacy level, while the level of privacy for the Randomized responses protocol is fixed at $\log(3)$. The Gaussian mechanism achieves weaker variation - approximate $(\epsilon, \delta)-$differential privacy.

The Laplace and Gaussian mechanisms focus on adding the right amount of noise to the answer, while the Exponential mechanism focuses on selecting the best answer from a set of candidate outputs. The accuracy of the Randomized responses increases with the number of data entries used. By other mechanisms, the accuracy of the output depends on the chosen privacy parameter ϵ and the sensitivity of the query or score function. The Randomised response method can be applied for local differential privacy setting without a trusted data curator.

4.2.8 Combining the Central and Local Differentially Private Models

The shuffle model presented by Cheu et al. in their paper [CSU+19] combines features of both local and central models while maintaining some of their benefits. It utilizes multiple shufflers to ensure the privacy of data owners before the data is sent to the data curator. The model is based on the ESA (encoder, shuffler, analyzer) architecture in the Prochlo system [BEM+17].

Encoders run on users' side and can modify monitored data by changing its encoding or adding nested encryption. For privacy purposes, encoders can control the scope, granularity, and randomness of the data released from clients, and they can add random noise to ensure user privacy without any trust assumptions.

Shufflers receive encrypted and encoded data and are responsible for masking data origin and confounding data provenance by eliminating metadata and shuffling the data to prevent traffic analysis. They collect data into batches over a lengthy period, such as, for example, one day, to eliminate timing metadata and ensure that every data item can get lost in a crowd of similar items. Once a large enough batch is formed, shufflers forward it for analysis after undoing a layer of nested encryption.

Analyzers receive shuffled data batches and are responsible for materializing and analyzing or releasing a database after undoing the innermost encryption. The specific analysis outcome and associated privacy protection are determined by the cryptographic key.

5 Privacy Consequences of Sampling

Sampling in a dataset refers to the process of selecting a subset of data points from the larger dataset in order to make statistical inferences about the entire population. One benefit of sampling is that it can help preserve the privacy of the data in the larger dataset. By selecting a smaller subset of data points, it is possible to minimize the risk of sensitive information being revealed about any individual in the dataset. This is especially important in cases where the dataset contains personal or sensitive information.

Sampling can also help improve the efficiency of data analysis by reducing the amount of data that needs to be processed. Rather than analyzing the entire dataset, analysts can work with a small and manageable subset of data to effectively capture the idea of the necessary pre-processing steps and rules to be applied to the whole dataset to gain insights and make predictions. Overall, sampling is a valuable tool for data analysis and can help protect the privacy of individuals in large datasets. However, it is important to use appropriate sampling methods and to ensure that the subset of data selected is representative of the larger population in order to ensure accurate statistical inferences.

Beyond that, the research community discusses that the sampling step assists in achieving a certain level of data privacy. Chaudhuri and Mishra quantify the level of confidentiality of data having rare values or outliers in a table in their research [CM06]. The paper quantitatively studies the relationship between the number of rare values in a table and the privacy in the released random sample. They prove that if there are no rare values in a table, it is possible to draw a direct relationship between the sample size, where each entry is selected with ϵ probability, and the evidence of $O(\epsilon)$ privacy with high probability. Otherwise, when there are t rare values, then the sample is $O(\epsilon\delta/t)-$private with probability $(1-\delta)$.

Nissim, Raskhodnikova and Smith propose a sample-and-aggregate mechanism to achieve differential privacy for releasing aggregate information of a database containing sensitive information about individuals. Their paper [NRS07] shows, that the heterogeneity of sensitivity function is a great advantage as it can significantly reduce the amount of noise that is added to the privatized statistics.

Gehrke et al. propose the idea of using private crowd-blending mechanisms for histograms and for releasing synthetic data points in combination with a pre-sampling step in their work [GHLP12]. The authors prove that the combined mechanism satisfies not only differential privacy, but also the stronger notion of zero-knowledge privacy. Indeed, they note that the method is applicable even if the pre-sampling is biased slightly and the adversary knows whether or not certain individuals appear in the sample.

Differential Privacy Amplification by Subsampling

Subsampling is a fundamental technique in the development and analysis of algorithms used not only for machine learning, but also for differentially private mechanisms. Kasiviswanathan et al. first mentioned the notion of the "privacy amplification by subsampling" principle in the paper [KLN+11], which is based on the fact that the privacy guarantees inherent in a differentially private mechanism can be strengthened by applying it to a small random subsample of records from

a given dataset. Moreover, from a technical point of view, subsampling provides a direct method for obtaining privacy amplification when the final DP-mechanism is available only as a so-called "black box".

Definition. Subsampling Lemma. The dataset D of the size n and an $(\epsilon, \delta)-$differential privacy mechanism M over a random subsample of size $m < n$ drawn using function $\text{Subsam}(D)$ without replacement are given. Then the mechanism $M \circ \text{Subsam}(D)$ is $(\epsilon', \delta')-$differential private with:

$$\epsilon' = \log(1 + \frac{m}{n}(e^{\epsilon} - 1)) \qquad \delta' = \frac{m}{n}\delta$$

Further, it was demonstrated in a number of research papers [JG17; LQS11; WBK19] that random pre-sampling applied to a particular mechanism can be used to amplify the privacy of a differential private mechanism.

Table 5.1 provides explicit privacy bounds [BBG18] for the most common subsampling methods, like Poisson subsampling, sampling with replacement and sampling without replacement. Thus, n is the size of the original dataset D, m is the size of the desired sample. γ corresponds to probability of each element from x to x' independently added.

Subsampling	Neighbouring relation	$e^{\epsilon'}$	δ'
Poisson	"remove/add-one"	$1 + \gamma(e^{\epsilon} - 1)$	$\gamma\delta$
Without repl.	"substitute-one"	$1 + \frac{m}{n}(e^{\epsilon} - 1)$	$\frac{m}{n}\delta$
With repl.	"substitute-one"	$1 + (1 - (1 - \frac{1}{n})^m)(e^{\epsilon} - 1)$	$\sum_{k=1}^{m} \binom{m}{k} (\frac{1}{n})^k (1 - \frac{1}{n})^{m-k} \delta$

Table 5.1: Privacy Amplification Bounds by Subsampling.

Two concepts of neighborhood relations are found in the differential privacy literature.
The "remove/add-one" relation $D_1 \simeq_r D_2$ occurs by removing or adding one element to D_1.
The "substitute-one" relation $D_1 \simeq_s D_2$ results from replacing an element in D_1 with another element from the domain. Thus, $D_1 \simeq_r D_2$ relates pairs of datasets of different sizes, while $D_1 \simeq_r D_2$ connects only pairs of datasets of the same size.

6 Data Utility Metrics in Anonymization

In order to determine the best approach to anonymize a given type of dataset it is necessary to evaluate anonymization algorithms. However, there are no standardised metrics which makes algorithms comparison difficult.

Various metrics exist to measure the utility of data itself, but not all of them are appropriate for use in Privacy-preserving data publication because the data curator does not know the precise usage scenarios of the data once it has been published. There has been a growing demand in the research community for general-purpose utility metrics that use syntactic properties to measure the data utility implicitly. This section discusses various general-purpose metrics applicable to the evaluation of anonymous data publishing algorithms, which can be used to calculate generalized information loss, determine the level of indistinguishability of records, and evaluate the defining of equivalence classes.

6.1 Discernibility Metric

The discernibility metric [BA05] assesses the level of indistinguishability of a record from other records within a class. Hence, it is an important measure among $k-$anonymity operation, achieved through generalization and suppression. Each record is given a penalty based on the size of the equivalence class to which it belongs. In case of a suppressed record, a penalty equals the size of the input table. The formula of the total discernibility score for the anonymized table T is defined as follows:

$$DM(T) = \sum_{\forall E \in T, |E| \geq k} |E|^2 + \sum_{\forall E \in T, |E| < k} n \cdot |E|,$$

where n is the number of records, E are equivalence classes and $|E|$ is the size of a class in the anonymized table T.

#Tuple	Equivalence class	Age	Zip-Code	Working class	Disease
1	1	[25-35]	7702∗	Non-Government	Acne
2	1	[25-35]	7702∗	Non-Government	Psoriasis
3	1	[25-35]	7702∗	Non-Government	Hemophilia
4	2	[55-65]	7701∗	Government	Hypertension
5	2	[55-65]	7701∗	Government	Cirrhosis
6	2	[55-65]	7701∗	Government	Hypertension

Table 6.1: Anonymized Table with $k = 3$ and QI={Age, Zip-Code, Working Class}.

To apply this metric, consider the example in Table 6.1. It consists of two equivalence classes of size 3 each, as specified by parameter $k = 3$. Therefore the second term of the equation is ignored. The DM score for the table is calculated as: $3^2 + 3^2 = 18$.

A key point is that larger equivalence classes introduce a greater loss of information, so they gain a greater penalty. Thus, the smaller values are better for an algorithm. Although in the concept of $k-$anonymity, a larger equivalence class size k provides a higher level of privacy.

An interesting point to note is that the DM metric does not reflect the petrurbation carried out on the records within the equivalence class. Thus, records in a class may not be generalized, but even in this case they are penalized.

6.2 Average Equivalence Class Size

The average equivalence class size metric [LDR06a] quantifies the level of approximation of the anonymization goal when each equivalence class is made up of k or more records.

$$C_{AVG}(T) = \frac{n}{q_{\forall E \in T} \cdot k},$$

where n is the number of records, q is the total number of equivalence classes in the anonymized table T and k is the privacy parameter. The result of 1 indicates a perfect anonymization, where the size of the equivalence classes is equal to the given value of k.

To apply this metric let‘s review again the example in Table 6.1. There are 6 records and two equivalence classes and privacy parameter $k = 3$. The C_{AVG} score for the table comes out as: $\frac{6}{2*3} = 1$.

6.3 Generalized Information Loss

The generalized information loss metric [Iye02] captures the penalty taken when an attribute value has been generalized. Domain values that have been generalized are scored quantitatively. Below is the normalised formula for the generalized information loss for the anonymized table T:

$$GenILoss(T) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{U_{ij} - L_{ij}}{U_i - L_i}}{n \cdot m},$$

where m is the number of attributes, n - the number of records, L_{ij} and U_{ij} are the lower and upper bounds of a node in a hierarchy level for a record j and an attribute i.

The underlying idea of the metric is that a generalized value of a data attribute representing a larger range of values, namely all children of a value in the hierarchy tree, is less accurate. For example, gender “$*$” is less accurate than “female”. According to the metric, 0 represents no generalization - original data and 1 represents the maximum level of generalization of the data.

Note that the attribute type is important for the calculation. For example, to calculate the metric for categorical attributes, it is possible to use the approach where each value is mapped to a numeric value.

For example, consider the hierarchy for the “Working class” attribute in Figure 6.1. At the $WC2$ level the value “Local-gov” coincides with 1, value “Federal-gov” with 2 and so on, until “Never-worked” is marked as 8. Thus, the working status of “Unemployed” is represented by the interval $[7 - 8]$, which covers the statuses from “Without-pay” and “Never-worked”. However, this way of handling categorical data is not necessarily the most appropriate, therefore another metric discussed in the next section is preferable.

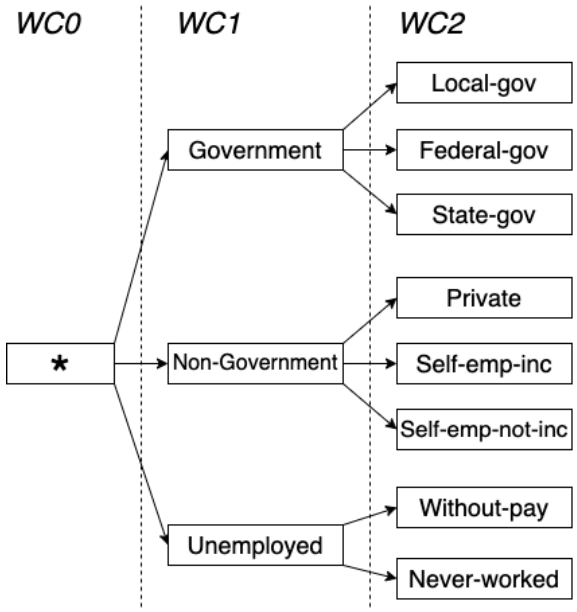


Figure 6.1: Taxonomy Tree of the Attribute "Working Class".

One thing to note is that the $GenILoss$ metric is sensitive to the attribute hierarchy depth. This means that a generalized attribute with a high hierarchy will take a smaller penalty compared to an attribute with a smaller levels in the tree.

6.4 Normalized Certainty Penalty

The normalized certainty penalty [XWP+06] quantifies the loss of information due to generalization. A hierarchy of generalizations for an attribute is given, where l denotes a node in the taxonomy tree and L denotes the range of all values for the attribute. The NCP_{value} for an original value l, when l is generalized to an l', is computed as follows:

$$NCP_{value}(l) = \begin{cases} 0 & \text{if } l' \text{ is a leaf node (the most accurate, ungeneralized value),} \\ \frac{|l'|}{|L|} & \text{otherwise,} \end{cases}$$

where $|l'|$ is the number of leaf nodes included in the generalization hierarchy of l', and $|L|$ is the total number of leaf nodes in the generalization hierarchy.

The total normalized certainty penalty for the anonymized table T is given below:

$$NCP(T) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} NCP_{value}(l_{ij})}{n},$$

where n is the number of records, m is the number of attributes to be generalized and v_{ij} is a generalized value of an attribute j and of a record i.

6.5 Applicability of Metrics in Anonymization Algorithms

Apparently the DM formula measures only the size of the equivalence class. Thus, only the number of records in the equivalence class is taken into account and it does not reflect the information loss caused by the transformations performed on the equivalence class. Based on this, multiple anonymization solutions may have the same data utility with respect to DM, even if different classes were anonymized with different QI set and with different levels of generalization in each solution. In addition, records may not be generalized, and even then, they will be penalized based on the size of their equivalence class.

$GenILoss$ metric assesses both the cardinality of the equivalence class and information loss. But it is highly sensitive to the depth of hierarchy of quasi-attributes, which means that a generalized attribute with a higher hierarchy will have a lower penalty compared to the penalty carried by other attributes with a lower hierarchy, such as the attribute of a degree of education and a person's gender.

The C_{AVG} metric fails to reflect the granularity of the equivalence classes created. That is, it cannot express the distribution of records among classes. If the C_{AVG} score is the same for two algorithms, it means that these algorithms create the same number of classes, but it does not necessarily mean that their class sizes are the same.

Since any metric does not fully capture the utility of the data behind the algorithm, metrics are often used in combination [FWCY10]. Table 6.2 presents the previously discussed algorithms with the anonymization operations used in them, and the utility metrics for the anonymized data that are most frequently applied in them.

Algorithm	Operation	Metric
Datafly	Full-domain generalization, record suppression	DM
Incognito	Full-domain generalization, record suppression	DM
Mondrian	Multi-dimensional generalization, median partitioning	DM, C_{AVG}, NCP
Top-down	Multi-dimensional generalization, binary partitioning	DM, C_{AVG}, NCP

Table 6.2: Characteristics of Anonymization Methods.

7 Algorithm for Privacy-aware Data Publishing

In this chapter, a differentially private algorithm for publishing a dataset applying anonymization techniques is presented. The algorithm has been developed taking into account the beneficial aspects of the standard anonymization algorithms covered in the previous chapters, as well as the privacy guarantees granted by differential privacy mechanisms.

7.1 Background Environment

A typical data science situation can be illustrated using the Knowledge Discovery in Databases (KDD) process [FPS96] in Figure 7.1. It summarises all the steps leading from dataset maintenance to knowledge discovery in the data. The phases include data selection, data cleaning, transformation into a data model, as well as the application of data mining techniques within the model, interpretation and evaluation of the derived patterns in the data.

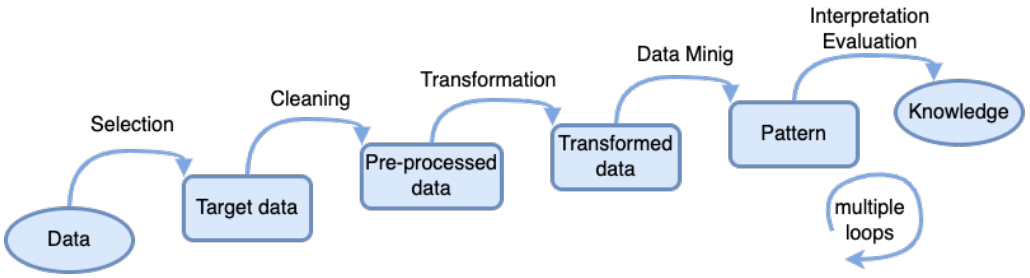


Figure 7.1: Milestones of Knowledge Discovery Process in Databases / Data Mining.

SMARTEN — A Sample-Based Approach towards Privacy-Friendly Data Refinement [SBB+22] raises a problem of data privacy in the context of data refinement. SMARTEN architecture represents the Data Processing Engine designed to handle data obtained from IoT devices that monitor goods from production to purchase [SGPM20], while providing trustworthy acquisition, secure data storage using blockchain technology, confidential monitoring and privacy-aware data delivery. In the scenario considered in the paper, the phases are divided into two stages: preparation and processing. The dataset has to be prepared by a data steward to support further processing by a data analyst. The data steward is the primary party that examines the data and establishes the rules for cleaning and transforming the raw data. However, as it is often the case in real-world applications, the data steward is an honest-but-curious party. Therefore, the data must be privately protected already at the pre-processing stage. This means that it is presented to the data steward already in a private manner. An illustrative example of this case is shown in Figure 7.2.

The research area of privacy preserving data publication is investigated in this work. Thus, various input perturbation techniques, previously discussed in Chapter 2.1, such as $k-$anonymity and its enhancements, can be used to satisfy the necessary anonymization degree. These models can

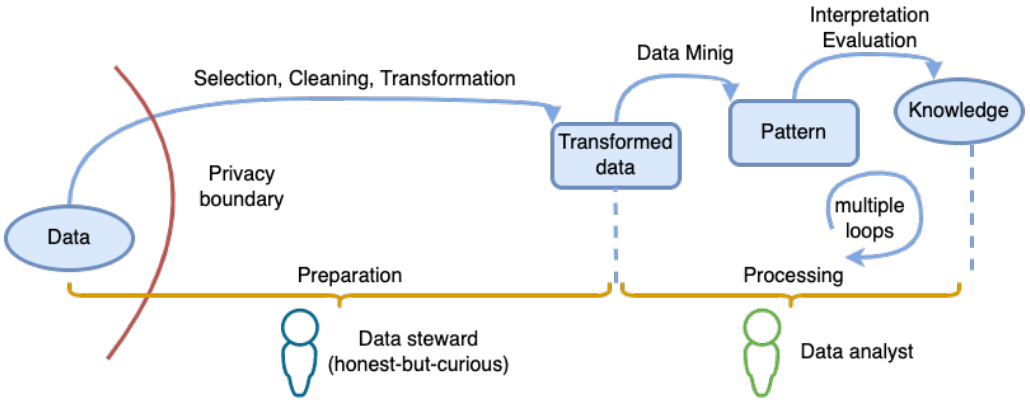


Figure 7.2: Distribution of Data Preparation and Processing Phases.

effectively secure privacy under certain syntactic conditions. That is, they make certain assumptions about the adversary's background knowledge. Therefore, they can be vulnerable to a variety of attacks, as shown earlier.
In contrast to syntactic privacy, differential privacy (DP), provides semantic privacy, i.e. more robust and mathematically provable privacy guarantees. Application of DP eliminates this drawback, due to its independence of the adversary's knowledge and immunity to underlying data. However, DP is mostly intended for the privacy preserving data mining, the query-response model or aggregated statistics. Therefore, this privacy model is not appropriate for application in a given scenario, as data analysis, data mining will take place only after the data steward has completed his/her operation.

The one $\epsilon-$DP mechanism, that can be applied in a data publishing scenario, is the modification of sensitive attribute values using the Randomized response mechanism. However, doing so is only possible if the type of attribute's values is binary. In addition, this would significantly affect the accuracy of the data at such an early stage of the data processing. The released data should preserve privacy, yet remain useful for further analysis.

For this reason, this work investigates the possibility of combining the success of the two fields, namely anonymization methods enhanced by various differentially private mechanisms.

7.2 DP-anonym - Sampling Differentially Private Anonymization Algorithm for Privacy Preserving Data Publication

A schematic illustration of the knowledge discovery process ensuring confidentiality during pre-processing is shown in Figure 7.3. Initially, a random sample is drawn from the original data according to the SMARTEN approach. Random sampling has a number of advantages in the privacy domain. Particularly, it is discussed in Chapter 5 that anonymization based on random sampling achieves $(\epsilon, \delta)-$differential privacy. The reduction in data size has a positive effect on the complexity of the anonymization algorithm. Some privacy protection is provided if some records are hidden. In addition, the data steward needs an overview of the data, not each individual record, to define the transformation rules. This sample is then anonymized with an algorithm that achieves $\epsilon-$differential privacy. If all privacy thresholds are met, this anonymized table is released for the data steward. Otherwise, a new sample is selected and privatized until it meets the specified privacy criteria.

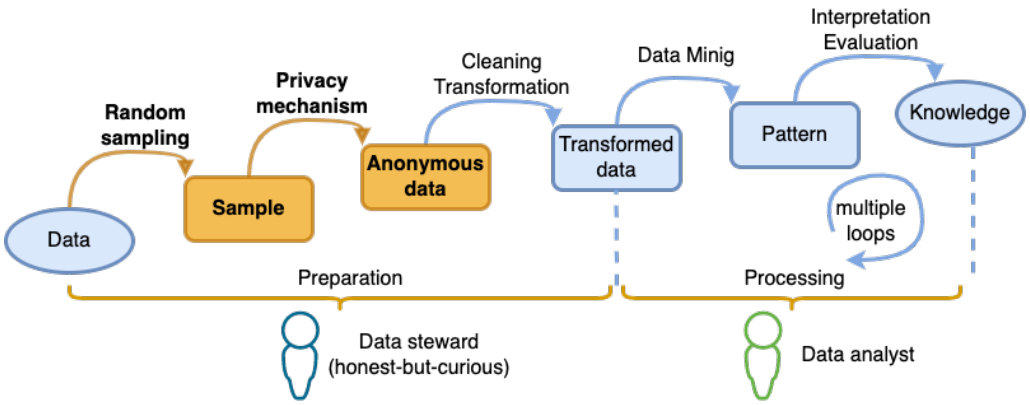


Figure 7.3: Process of Data Refinement in Data Privatisation with Sampling.

We are going to consider now a differential private algorithm for publishing anonymized data of this work. We call it "DP-anonym" for easy reference further on. This name encrypts the key objectives of the algorithm, such as anonymization and providing the properties of differential privacy. A flowchart in Figure 7.4 depicts the main stages of DP-anonym's execution.

An original dataset and metadata for each of its attribute are given. The metadata contains the information about an attribute's data type, a domain set or value space, an attribute's privacy type, whether it is QI or SA, and a hierarchical tree. This information is required for the subsequent generalization operation within anonymization. In addition, hierarchical trees are necessary to calculate the optimal value of the sensitivity function of the DP mechanisms in the algorithm. Furthermore, metadata information can be used to generate synthetic data in case where, after a given number of attempts to release anonymized data, the algorithm still fails to reach the specified privacy thresholds. A variety of algorithms have been discussed in the literature regarding the generation of synthetic data while respecting privacy. However, this topic is beyond the scope of this work. A random sample is taken from the data and the subsequent anonymization is carried out on the data sample only.

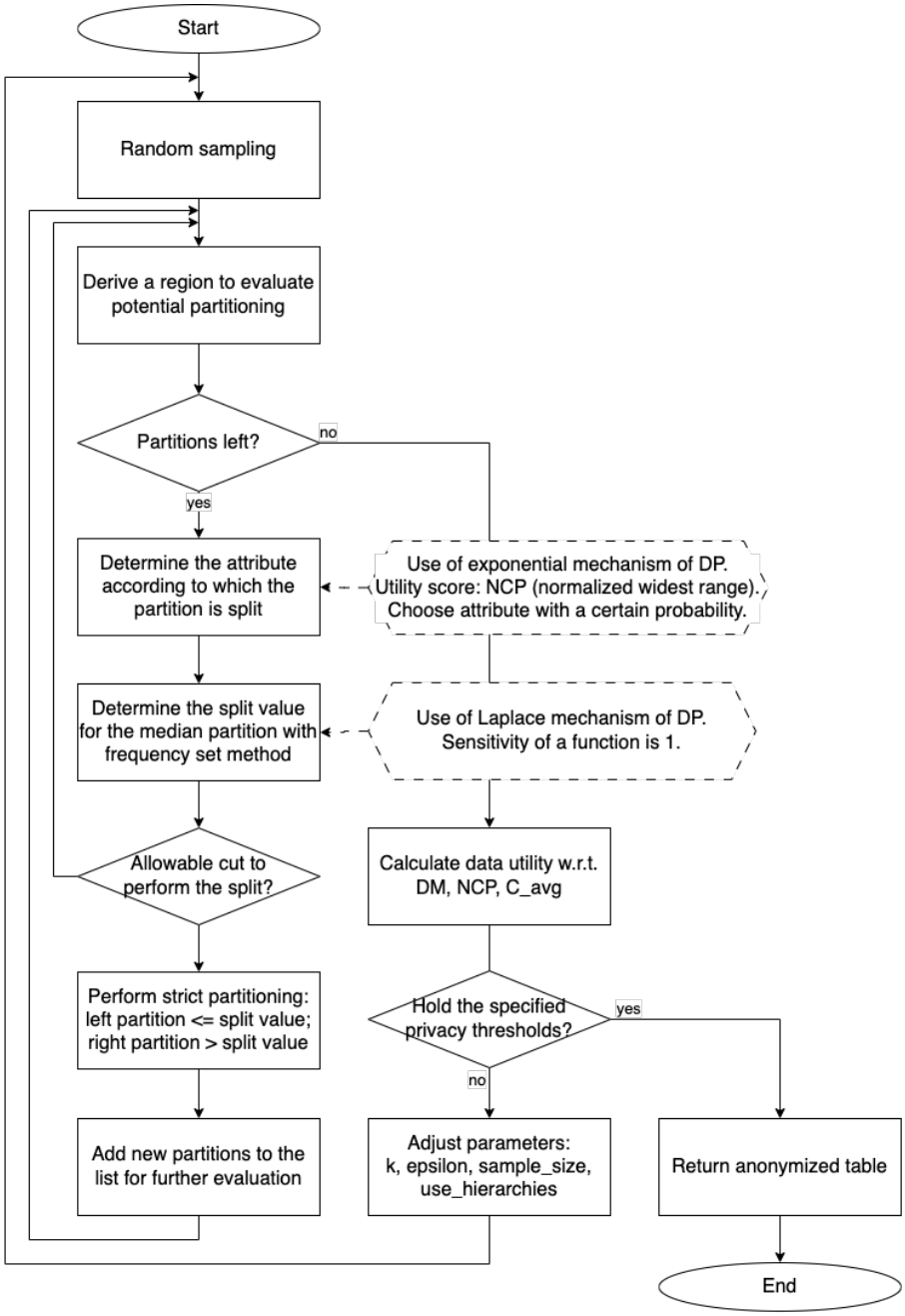


Figure 7.4: Flowchart of DP-anonym.

7.2.1 Generalization Model

Now we proceed with data anonymization and try to ensure $k-$anonymity. The main operations of data petrubation to achieve $k-$anonymity are generalization and subsequent record suppression. Within the generalization operation, a single-dimensional and a multidimensional models are distinguished. This implies that in the single-dimensional model all attribute values are converted

to the same hierarchical level after generalization, whereas in the multidimensional approach this condition does not have to be met. An example of the tables after generalization using different models is depicted in Figure 7.5.

Original table

ID	sex	age	race	marital-status	education	native-country	workclass	occupation	salary-class
1	Male	39	White	Never-married	Bachelors	United-States	State-gov	Adm-clerical	<=50K
2	Female	37	White	Married-civ-spouse	Masters	United-States	Private	Exec-managerial	<=50K
3	Female	49	Black	Married-spouse-absent	9th	Jamaica	Private	Other-service	<=50K
4	Male	52	White	Married-civ-spouse	HS-grad	United-States	Self-emp-not-inc	Exec-managerial	>50K
5	Male	32	Black	Never-married	Assoc-acdm	United-States	Private	Sales	<=50K
6	Male	34	Amer-Indian-Eskimo	Married-civ-spouse	7th-8th	Mexico	Private	Transport-moving	<=50K
7	Male	25	White	Never-married	HS-grad	United-States	Self-emp-not-inc	Farming-fishing	<=50K

Datafly: single-dimensional generalization, k=3

ID	sex	age	race	marital-status	education	native-country	workclass	occupation	salary-class
1	Male	*	White	*	Higher education	*	*	*	<=50K
2	Female	*	White	*	Higher education	*	*	*	<=50K
3	Female	*	Black	*	Secondary education	*	*	*	<=50K
4	Male	*	White	*	Secondary education	*	*	*	>50K
5	Male	*	Black	*	Higher education	*	*	*	<=50K
6	Male	*	Amer-Indian-Eskimo	*	Secondary education	*	*	*	<=50K
7	Male	*	White	*	Secondary education	*	*	*	<=50K

DP_anonym: multidimensional generalization, k=3

ID	sex	age	race	marital-status	education	native-country	workclass	occupation	salary-class
1	Male	35~39	White	spouse not present	Bachelors	North America	State-gov	Other	<=50K
2	Female	35~39	White	Married-civ-spouse	Graduate	United-States	Private	Exec-managerial	<=50K
3	Female	45~49	Black	spouse not present	High School	*	Private	Other	<=50K
4	Male	50~54	White	Married-civ-spouse	High School	United-States	Self-emp-not-inc	Exec-managerial	>50K
5	Male	30~39	Black	Never-married	Higher education	United-States	Non-Government	Sales	<=50K
6	Male	30~39	Amer-Indian-Eskimo	Married-civ-spouse	*	North America	Non-Government	Other	<=50K
7	Male	25	White	Never-married	HS-grad	United-States	Self-emp-not-inc	Other	<=50K

Figure 7.5: Comparison of Single-dimensional and Multidimensional Generalization Models on "Adult" Dataset.

These results are obtained from "Adult" dataset, which are discussed in more detail in Experimental chapter 8. The top table represents several rows from the original dataset. The middle table presents data anonymized by Datafly algorithm and the last table presents data anonymized by DP-anonym algorithm proposed in this work. From the second table with single-dimensional generalization it can be seen that among the quasi-attributes, the values are summarized to the highest level, up to the maximum data suppression, up to the asterisk. In the table with multidimensional generalization, within one column there are values as well as entries with asterisk, but in the majority of the records more detailed values are presented.

For a better understanding, let's look at the "education" attribute in Table 7.1. The hierarchical tree of this attribute can be seen in Figure 8.1. In a single-dimensional case this attribute is generalized to the penultimate level - Primary education, Secondary education, Higher education. In multidimensional one, there are unchanged values - the leaves of the

hierarchy E3: IDs 1 and 7, and the values generalized to the first level E2 - IDs 2, 3 and 4, to the second level E1 - ID 5 and to the last level E0 - ID 6. The same logic applies to all other attributes.

	Original value	Single-dimensional	Multidimensional
ID	education		
1	Bachelors	Higher education	Bachelors
2	Masters	Higher education	Graduate
3	9th	Secondary education	High School
4	HS-grad	Secondary education	High School
5	Assoc-acdm	Higher education	Higher education
6	7th-8th	Secondary education	*
7	HS-grad	Secondary education	HS-grad

Table 7.1: Results of Single- and Multidimensional Generalization Models on "Education" Attribute.

Therefore, for example, Datafly and Incognito algorithms use a single-dimensional model in their basic implementation. Mondiran algorithm, on the other hand, uses a multidimensional model. LeFevre et al. carried out an investigation on these methods in the paper [LDR06b] with respect to the data utility after generalization. It was observed that multidimensional model significantly improves the utility of data with respect to other methods under the single-dimensional model. At the same time, the value of the privacy parameter remains the same. Obviously, this kind of behavior also depends on the characteristics of a given dataset. Nevertheless, for our application, the multidimensional model is more appropriate as well. This is because an anonymized table allows the attribute values to be expressed at different hierarchical levels, thereby improving the quality of the steward's comprehension of the data.

7.2.2 Integration of Differential Privacy into Anonymization Algorithm

The anonymization algorithm presented in this work is inspired by Mondrian algorithm, as it has a number of advantages, as noted in Section 3.3, that are essential for the fulfilment of the desired scenario. Namely, it uses a multidimensional generalization model. There are variants of Mondrian algorithm in which the pre-definition of hierarchies for attribute domain values are not required. This may be relevant to situations where hierarchical trees for attributes do not exist or have not been communicated in metadata of the dataset. More details on this issue are given in Experimental chapter 8 of the work. Mondrian provides higher data utility in comparison to other algorithms when considering metrics such as discernibility metric, average equivalence class size, certainty penalty. It is also able to achieve $k-$anonymity and its improvements when the privacy parameter k is high. Moreover, it uses median partitioning as the main operation to generate equivalence classes, in contrast to the huge search space of full domain generalization in Incognito or binary search in Datafly. The complexity of Mondrian algorithm does not depend on the number of quasi-attributes or the depth of their hierarchies. Thus, Mondrain is several orders of magnitude faster and consumes less memory space.

The idea of the algorithm is straightforward, its flowchart is shown in Figure 3.1. First, a random sample is selected from the dataset. The sample size, as well as privacy parameters namely k, epsilon, maximum threshold of value of data utility metrics, and a trigger enabling hierarchies are defined by data steward. The ability to set non-honest values, such as 100% sample size, $k-$value and ϵ being equal to one, is of course restricted. This will expose all data from the set, leaving it unchanged and providing no privacy whatsoever. It is recommended to use higher values for parameter k, an epsilon ranging by $(0, 1)$ and a sample size smaller than 100 percent.

The algorithm works recursively, dividing the entire dataset into smaller partitions, each of which must be converted into a single equivalence class at the end. The partitioning is done by one quasi-attribute. The median of its values in the given partition is found and then splitting into two parts is performed. Attribute values are generalized at the partitioning of records during grouping stage as well.

Through differential privacy mechanisms, it is possible to introduce nondeterminism into the stages of generation of equivalence classes, thereby increasing the level of privacy of the resulting released table. Thus the stage of selecting a dimension, which will be subjected to a partition in an iteration of an algorithm, is performed from the size of the normalized range of dimension. This is to some extent a disadvantage of this approach. Imagine a situation where one quasi-attribute has a wide domain, like “age”, while other attributes have only a few attribute values in the database, for instance, “sex” has only two values -“female” and “male”, or the attribute “marital status” has 6 values in “Adult” dataset. Thus, “age” is selected more often than other attributes and the table is partitioned according to this attribute, thus “age” will be more generalized in the resulting table.

Exponential Mechanism during Attribute Selection for Partitioning

The exponential mechanism is applied in cases when it is necessary to determine an answer from a set of possible results depending on the utility function score. This mechanism is applied to the selection of an attribute, according to which the partition is split. Specifically, attributes for which the partitioning is allowed are taken as a set of possible answers. The normalized range of attribute values in a given partition is used as a utility score function. Thus the attribute with the highest score is then selected with a certain probability. The exponential mechanism achieves an $\epsilon-$differential privacy. Consequently, the selection of an attribute step is also differentially private, based on the post-processing property.

Laplace Mechanism for Frequency Sets when Selecting Median

After choosing an attribute for partitioning frequency sets are calculated. Frequency set for an attribute in a given partition is a set of unique values, which are mapped to the integer count of their occurrences in the partition. The median of these counts is then selected using a standard method and a partitioning is performed on the median’s value of the attribute into two parts, often called left- and right hand sides. In Mondrian algorithm, there is a division between strict partitioning, where the partitions are split into $\leq$ medians and $>$ medians, and relaxed partitioning, where rows with an attribute value equal to the median value are divided by half into both partitions.

The frequency set approach has a positive effect on the scalability of the algorithm, since there are much fewer unique attribute values in a partition than there are records. Thus, even with increasing number of records, memory space for the frequency set remains approximately the same.

The noisy counts mechanism applies the Laplace mechanism to a counting query. The frequency set approach can be represented as a regular count query. The Laplace mechanism is integrated into it, adding points from the distribution to the responses in the set. The scale of the distribution is adjusted in each partition depending on the sensitivity function, i.e. how much the data changes in the set. This improves the privacy of the algorithm. This mechanism used for frequency set leads the median selection step to $\epsilon-$differential privacy, due to the post-processing and Laplace mechanism properties.

The algorithm ends by calculating utility metrics for the anonymized data, such as discernibility metric, average equivalence class size, normalized certainty penalty. In case they satisfy the given thresholds, the anonymized table is released. Otherwise, the input parameters are adjusted and another round is performed until the table meets the requirements.

7.3 Privacy Analysis of the Algorithm

The purpose of the presented algorithm is to release the data while guaranteeing privacy. To achieve this, data anonymization is carried out by means of $k-$anonymization, in which differential private mechanisms such as Laplace and exponentiation mechanisms are integrated. In this way, semantic and synthetic data protection is ensured. In the following, privacy guarantees for each individual step of the algorithm are considered and the resulting privacy of the algorithm is derived. After that, the level of privacy is analysed when the sampling step is used in the data selection for processing phase.

7.3.1 Privacy Budget of Partitioning Attribute Selection

The partitioning attribute is determined using the exponential mechanism with the normalized width of the attribute values in a given partition.

Assume $Attr$ is a set of candidates from which a partitioning attribute v is to be chosen.
Let P be a current partition, $\text{Range}(v)_P$ be the spread of the attribute v in partition P and $\text{Range}(v)$ be the spread of the attribute v across the whole dataset.
Thus the score function is $S(P, v) = \frac{\text{Range}(v)_P}{\text{Range}(v)}$ for a given attribute v.

Based on the scores of all candidates for the partitioning attribute, the exponential mechanism selects a candidate v with the following probability

$$\frac{exp(\frac{\epsilon^{\text{selection}}}{2\Delta S} S(P, v)}{\sum_{v \in Attr} exp(\frac{\epsilon^{\text{selection}}}{2\Delta S} S(P, v)},$$

where $S(P, v)$ is a score function, ΔS is the sensitivity of function S and $\epsilon^{\text{selection}}$ is a privacy budget.

Based on the Theorem 8 in Section 4.2.6 of the exponential mechanism, choosing an partitioning attribute with the probability proportional to $exp(\frac{\epsilon^{\text{selection}}}{2\Delta S})$ satisfies $\epsilon^{\text{selection}}-$differential privacy.

7.3.2 Privacy Budget to Determine a Split Value

In order to find a split value of an attribute, a frequency set is computed. For each unique attribute value in the partition in question, the occurrence frequency of the records is calculated. Then the median is selected from these values. In order to achieve differential privacy, noise of comparable size generated by the Laplace mechanism is added to the frequency set values.

This step of the algorithm is $\epsilon-$ differential private. Let be s a sensitivity of function of frequency set approach. A function's sensitivity is the amount of change in the output of a function when its input is changed by 1. Frequency sets can be interpreted as count queries. Hence, count queries always have sensitivity 1. Let's imaging the following situation. The query counts the number of records in a dataset with a specific property. Then, exactly one record of this dataset is altered or deleted. The count query on the same property is now applied to the altered dataset. So, the query output now can change by a maximum of 1.

Therefore, the sensitivity of a function $s = 1$ and ϵ^{count} is the privacy budget.
Based on the Theorem 7 in Section 4.2.3 of the Laplace mechanism, adding noise generated from the Laplace distribution $\text{Lap}(1/\epsilon^{\text{count}})$ to the results of the counting function achieves $\epsilon^{\text{count}}-$differential privacy.

7.3.3 Overall Privacy Guarantee of an Algorithm

In summary, it has been shown that each step of the algorithm guarantees differential privacy. Since these stages operate on the same dataset and are executed sequentially one after another, then based on the Theorem 3 in Section 4.1.3 of differential privacy property of sequential composition, it can be concluded that the whole algorithm achieves a privacy level equal to the sum of the privacy budgets of all its differential privacy stages.
In this way, the DP-anonym algorithm achieves

$$(\epsilon^{\text{selection}} + \epsilon^{\text{count}}) - \text{differential privacy.}$$

7.3.4 Privacy Bounds of Sampling Step in the Algorithm

Based on the subsampling lemma in Section 5, an algorithm needs to have $(\epsilon, \delta)-$differential privacy. This is a relaxation of $\epsilon-$differential privacy, which DP-anonym algorithm achieves. That is, proposed algorithm implicitly achieves $(\epsilon, \delta)-$differential privacy, involving δ as an adjunctive additive bias factor. Thus, assume the dataset D of the size n and an $(\epsilon, \delta)-$differential privacy mechanism DP-anonym over a random subsample of size $m < n$ drawn using function $\text{Subsam}(D)$ without replacement are given. Then the mechanism (DP-anonym $\circ$ $\text{Subsam}(D)$) is $(\epsilon', \delta')-$differential private with:

$$\epsilon' = \log(1 + \frac{m}{n}(e^{\epsilon} - 1)) \qquad \delta' = \frac{m}{n}\delta$$

8 Experimental Evaluation

During the experiments, the quality of the anonymization produced via the proposed differential private multi-dimensional algorithm was evaluated by comparing these results with the results obtained by the optimal algorithms using other models. It is important to note that computationally intensive algorithms, like Incognito and Datafly, are exponential in the worst case. Therefore, they are much slower. However, the quality of the results obtained by these algorithms is still compelling to compare. In addition, since Datafly and Incognito algorithms, according to research in the literature, reach comparable results, therefore it makes sense to use only one algorithm in the experiments, namely Datafly. This algorithm has a better compexity and thus can accelerate the experiments. The quality was compared using common data utility metrics as well as the results of the basic queries.

8.1 Datasets

Considering that it is extremely difficult to specify universal methods for anonymization as well as data utility metrics for evaluation, it is necessary to perform experiments on different datasets. In the following, the algorithms is empirically evaluated on four datasets given below. These datasets differs in the set size of their quasi-identifiers, the depth of their hierarchies, and the types of data (categorical and numerical). In addition, the datasets have different cardinality, ranging from 830 records to more than 30 thousand rows. The hierarchies for quasi-identifiers presented below have been additionally developed, while varying the hierarchy levels for semantically similar attribute, like “age”, among different datasets for better testing experience.

- **Adult dataset**[1] or also called Census Income dataset from the UCI Machine Learning Repository is already a gold standard in applied algorithm analysis. It contains information about the population census 1994 in the USA. In order to further anonymize the dataset, the following sample has been targeted. It consists of 30162 records and contains the attributes depicted in Table 8.1. “Salary-class” is defined as a sensitive attribute with two classes: “>50K” and “<=50K”. Hierarchical trees for quasi-attributes are shown in the following illustrations.

[1] see https://www.kaggle.com/uciml/adult-census-income

Privacy type	Attribute	Type	Unique values	Example value
QI	sex (Fig.2.3a)	categorical	2	“Female”
	age	numerical	[17,91]	27
	race (Fig.8.2)	categorical	5	“White”
	marital-status (Fig.8.3)	categorical	7	“Married-civ-spouse”
	education (Fig.8.1)	categorical	16	“Bachelors”
	native-country (Fig.8.5)	categorical	41	“Germany”
	workclass (Fig.6.1)	categorical	7	“Local-gov”
	occupation (Fig.8.4)	categorical	14	“Prof-specialty”
SA	salary-class	categorical	2	“<=50K”

Table 8.1: Attributes of “Adult” Dataset.

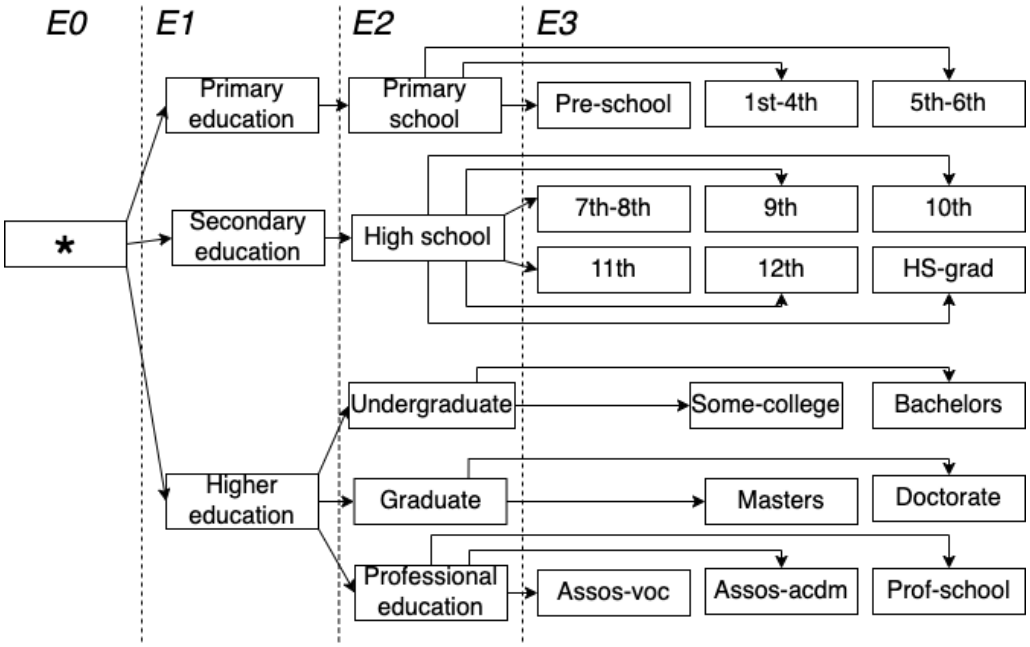


Figure 8.1: Hierarchical Tree of the Attribute “Education” in “Adult” Dataset.

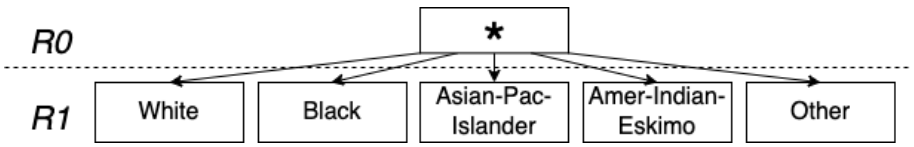


Figure 8.2: Hierarchical Tree of the Attribute “Race” in “Adult” Dataset.

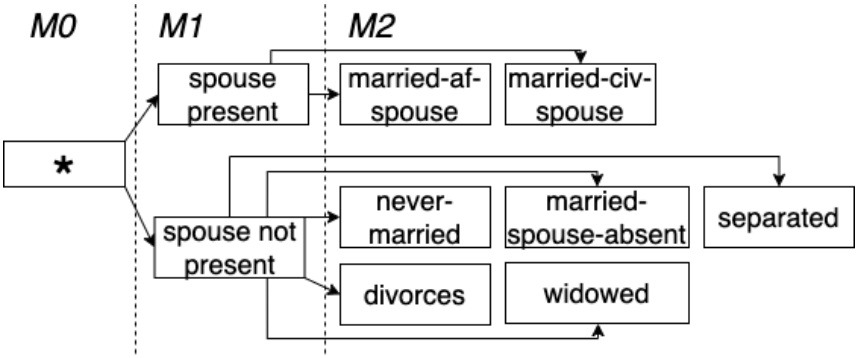


Figure 8.3: Hierarchical Tree of the Attribute "Marital-status" in "Adult" Dataset.

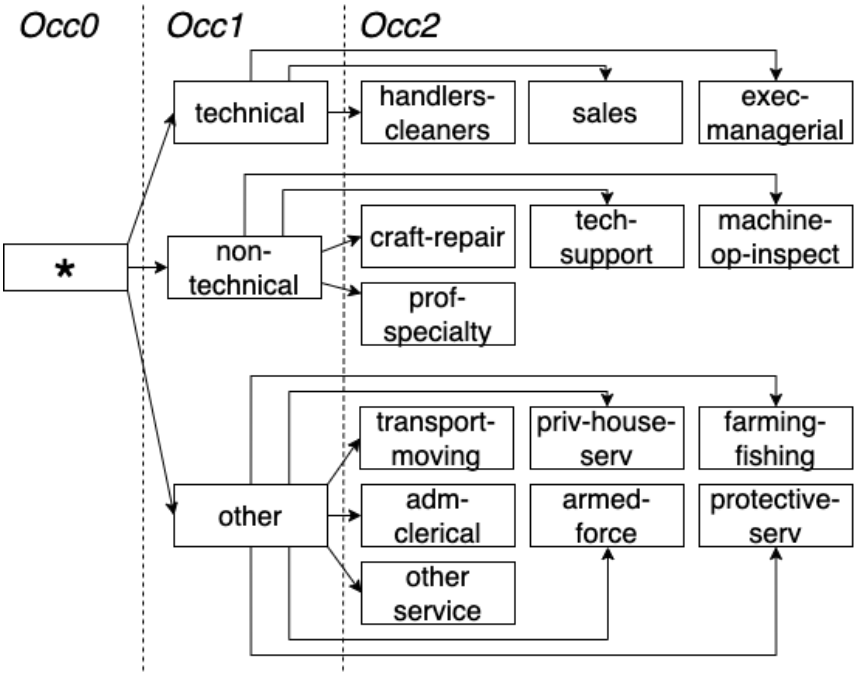


Figure 8.4: Hierarchical Tree of the Attribute "Occupation" in "Adult" Dataset.

- **California Housing dataset**[2] is presented in the StatLib repository. It was derived from the 1990 US census. This set is most often used in machine learning for regression tasks. It provides data on characteristics of properties on the Californian coast, such as the room numbers, house's age, income of tenants, etc. There are 20640 records in the dataset and it contains the attributes, which are marked as quasi- and sensitive attributes, given in Table 8.2. The level of "ocean proximity" is qualified as a sensitive attribute with 3 classes: 0- low, 1- middle, 2- high.

[2] see https://www.kaggle.com/camnugent/california-housing-prices

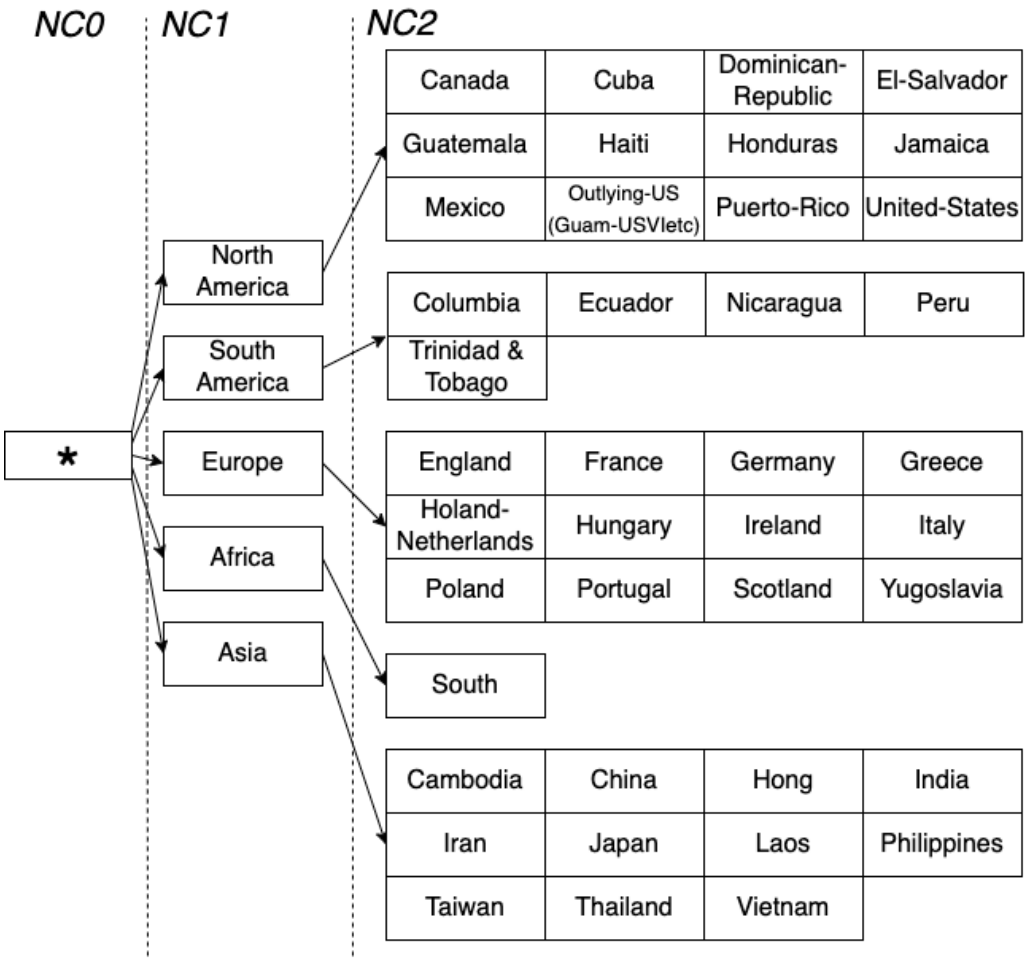


Figure 8.5: Hierarchical Tree of the Attribute "Native Country" in "Adult" Dataset.

Privacy type	Attribute	Type	Range of values	Example value
QI	longitude	numerical	[-124.35,-114.31]	-121.80
	latitude	numerical	[32.54, 41.95]	37.71
	housing median age	numerical	[1, 52]	37
	median income	numerical	[0.49, 15]	4.74
	median house value	numerical	[14999, 500001]	264725
SA	ocean proximity	categorical	3	1
NSA	mean rooms	numerical	[0.84, 59875]	6.08
	mean bedrooms	numerical	[0.12, 3625]	1.10
	population	numerical	[3, 35682]	1725
	households	numerical	[1, 6082]	605

Table 8.2: Attributes of "California Housing" Dataset.

- **Contraceptive method choice dataset**[3] is from the UCI Machine Learning Repository. It is a subset of Indonesian contraceptive study data from the 1987, which includes demographic and economic characteristics of women in relation to the methods used to control pregnancy. It consists of 1473 records and contains the attributes given in Table 8.3. Such attributes as "age", "education" level and "number of children" are given as quasi-identifiers and the "contraceptive method" is used as a sensitive attribute with 3 cases: "no use", "short-term use" and "long-term use".

Privacy type	Attribute	Type	Unique values	Example value
QI	age (Fig.8.6)	numerical	[16, 49]	39
	wife education (Fig.8.7a)	categorical	4	4
	children (Fig.8.7b)	numerical	[0, 16]	4
SA	method	categorical	3	3
NSA	husband education	categorical	4	4
	wife's religion	binary	{0, 1}	1
	wife's working	binary	{0, 1}	1
	husband's occupation	categorical	4	4

Table 8.3: Attributes of "Contraceptive Method Choice" Dataset.

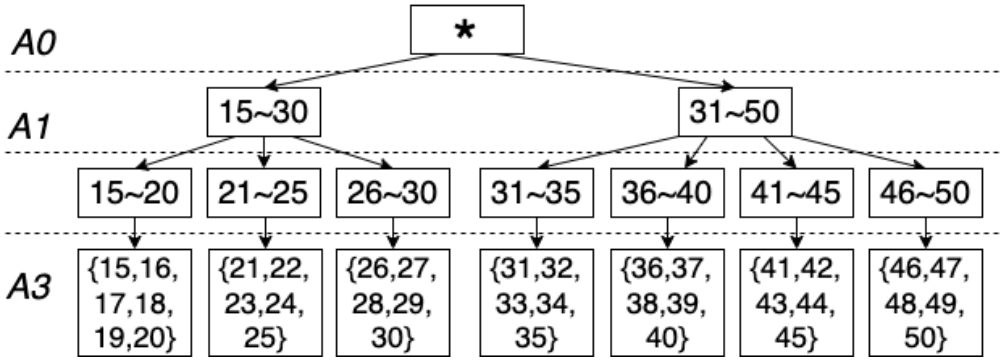
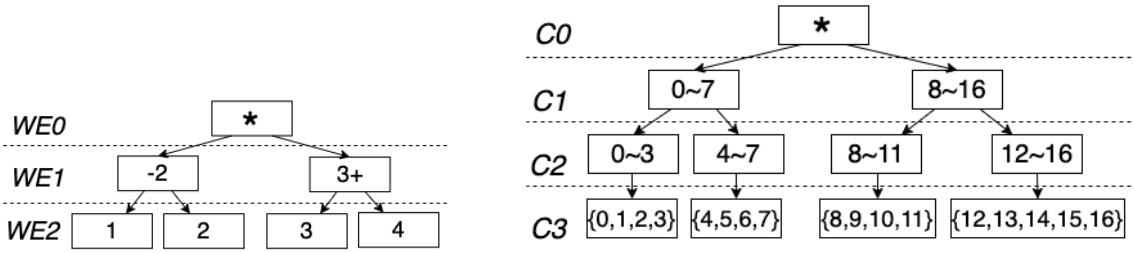


Figure 8.6: Hierarchical Tree of the Attribute "Age" in "Contraceptive Method Choice" Dataset.



(a) "Wife's education" **(b)** "Children"

Figure 8.7: Hierarchical Trees of Attributes "Wife's Education" and "Children" in "Contraceptive Method Choice" Dataset.

[3] see https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice

- **Mammographic mass dataset**[4] is also from the well-known UCI Machine Learning Repository. It gives patient information and test results concerning tissue damage from malignant or benign breast tumours. There are 830 records in total and the attributes are shown in Table 8.4. The sensitive attribute is "severity" with 2 classes: 0- benign and 1- malignant.

Privacy type	Attribute	Type	Range of values	Example value
QI	BI RADS assessment (Fig.8.9a)	categorical	7	5
	age (Fig.8.8)	numerical	[18, 96]	66
	shape (Fig.8.10a)	categorical	4	4
	margin (Fig.8.9b)	categorical	5	4
	density (Fig.8.10b)	categorical	4	3
SA	severity	binary	2	1

Table 8.4: Attributes of "Mammographic Mass" Dataset.

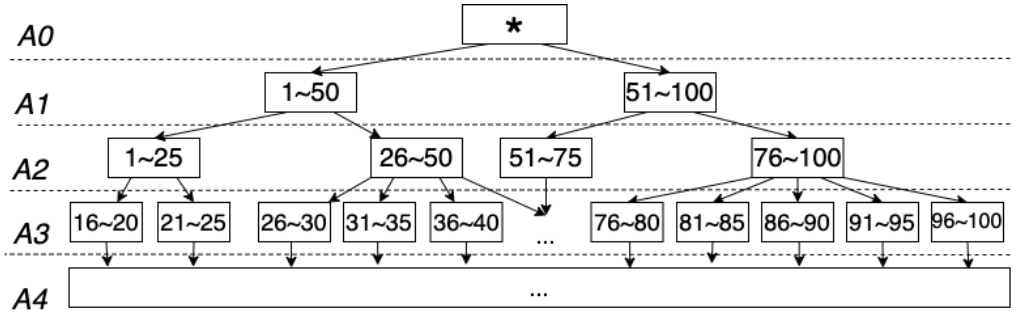
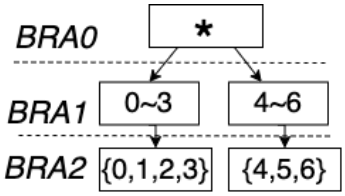
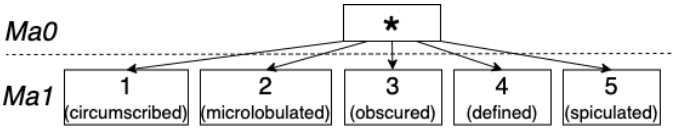


Figure 8.8: Hierarchical Tree of the Attribute "Age" in "Mammographic Mass" Dataset.



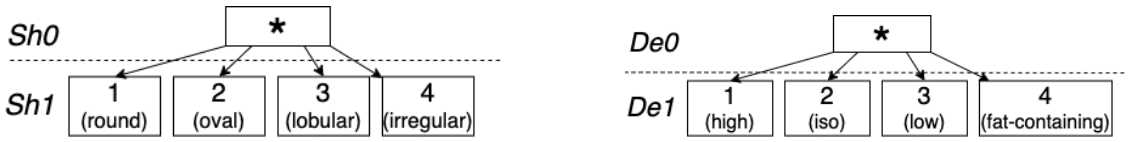
(a) "BI RADS assessment".



(b) "Margin".

Figure 8.9: Hierarchical Trees of Attributes "BI RADS Assessment" and "Margin" in "Mammographic Mass" Dataset.

[4] see http://archive.ics.uci.edu/ml/datasets/mammographic+mass



(a) "Shape". **(b)** "Density".

Figure 8.10: Hierarchical Trees of Attributes "Shape" and "Density" in "Mammographic Mass" Dataset.

8.2 Libraries

IBM Differential Privacy Library *Diffprivlib*[5] is used to implement differential privacy mechanisms in the proposed algorithm. It is a general-purpose library for experimentation, research and algorithm development in the area of differential privacy and is widely used in the research community. Diffprivlib has implementaion of Gaussian, Laplace and exponential mechanisms. In addition, there is a wide set of machine learning models with differential privacy. Currently, Diffprivlib has models for clustering, classification, regression, dimensionality reduction and pre-processing. The library offers a number of general tools for analysing differentially private data, such as differentially private histograms and others.

The **k-anonymity library**[6] provides implementations of several core algorithms: Datafly, Top-down algorithm, Mondrian, which are most commonly used in practice for anonymizing datasets.

8.3 Results

In the following experiments, the proposed algorithm DP-anonym is compared with the algorithms Datafly, Basic Mondiran, Classic Mondiran, Top- down. The evaluation is based on metrics to analyse the quality of data anonymization in the PPDP setting. In this experiment, an average epsilon value equal to 0.5 is used in DP-anonym algorithm in order to avoid introducing a marked difference in the degree of privacy among the other algorithms. It examines how much the anonymized data loses in utility when applying the proposed algorithm compared to standard algorithms, how the utility of the data changes as a function of increasing the level of epsilon differential privacy, and compares the results of responses to common queries in real-world applications.

Data utility analysis

Figure 8.11 shows a comparison of the algorithms on the "Adult" dataset in terms of the utility of anonymized data. The chart illustrates the evolution of the data utility values when increasing a privacy parameter k.

There is a clear indication that Datafly algorithm shows the largest and mostly constant values for all data utility metrics, that once again proves the necessity of developing and implementing enhanced algorithms. An interesting observation is that at $k = 6$ and $k = 11$ a noticeable improvement in all

[5] see https://diffprivlib.readthedocs.io/en/latest/

[6] see https://github.com/kaylode/k-anonymity

three metrics is outlined. This suggests that the “Adult” dataset using Datafly performs better with the $k = 11$ equivalence class size, keeping in mind that $k = 6$ is also included in the class size 11 according to the roll-up principle.

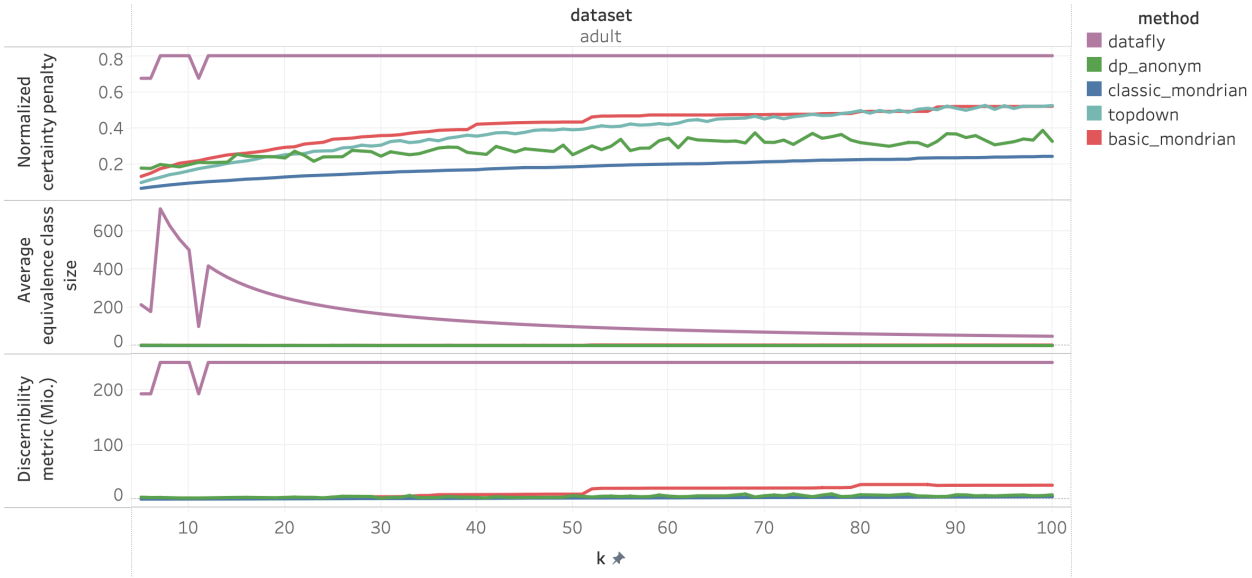


Figure 8.11: Comparison of Data Utility Metrics on “Adult” Dataset among Algorithms: DP-anonym, Datafly, Basic Mondrian, Classic Mondrian, Top-down.

A subsequent decay of the value is noticeable in the Average equivalence class size metric. It is worth mentioning that this, however, is not due to an improvement in the quality of the data. When we recall the metric formula $C_{AVG}(T) = \dfrac{n}{q_{\forall E \in T} \cdot k}$, where n is the number of records, q is the total number of equivalence classes in the anonymized table T and k in the privacy parameter, it becomes clear that this is due to an increase of value k in the denominator, having the same value of the numerator.

For a more detailed consideration of other algorithms, the values of the utility parameters after applying Datafly algorithms are omitted from the graph in Figure 8.12. The basic Mondrian, shown in red, implies the absence of predefined hierarchies for attributes, whereas the classic Mondrian uses them. Similarly, it is interesting to note the prominent gradients in the evaluation of the basic Mondrian method at levels of k equal to 25, 52, and 80.

From the Average equivalence class size metric, as well as Discernibility metric, it can be concluded that at these values of parameter k, equivalence classes have increased dramatically, when two or more equivalence classes are clustered into one. Recalling the metric formula $DM(T) = \sum_{\forall E \in T, |E| \geq k} |E|^2 + \sum_{\forall E \in T, |E| < k} n \cdot |E|$, where n is the number of records, $|E|$ is the size of a class in the anonymized table T, it is clear why there was such a big quadratic spike, as the size of an equivalence class has doubled.

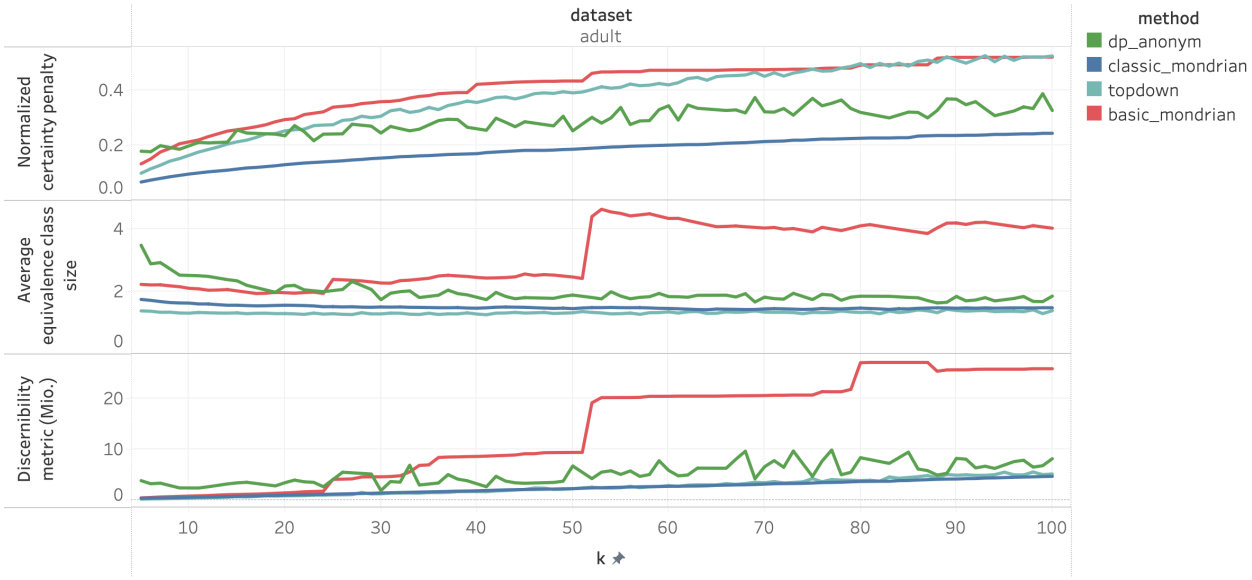


Figure 8.12: Comparison of Data Utility Metrics on "Adult" Dataset among Algorithms: DP-anonym, Basic Mondrian, Classic Mondrian, Top-down.

In addition, a clear increase can be observed in the Normalized certainty penalty metric, which shows how strongly the data are generalized according to their hierarchy level. This also confirms the theory presented earlier. Thus, it can be said that by applying the basic Mondrian method to the dataset "Adults" the privacy parameter more than 52 no longer carries all the necessary guarantees for the data quality.

Let's consider the evaluation of the algorithm DP-anonym proposed in this work. The noticeable fluctuation is due to the randomized nature of the algorithm, which is a core goal in providing differential privacy. Despite this, it can be seen that the algorithm achieves relatively similar values and behavior for all three metrics compared to the most successful algorithms in the experiment, the classical Mondrian and Top- down. Moreover, in the case of generalization quality, the DP-anonym algorithm even exceeds the Top- down algorithm.

When using another dataset "California housing", a contrasting pattern can be observed. The results of this experiment can be seen in Figure 8.13. All quasi-attributes in this dataset are numerical, unlike in "Adult" dataset. In these conditions, the DP-anonym algorithm shows even better values of the data utility metrics, which can be explained by the smoother and more gradual generation of new equivalence classes.

Furthermore, based on the metric values for basic and classical Mondrian, it can be noted that although partitioning in numerical attributes is not a major problem as it is in the case with categorical data due to the semantical difficulties in the attribute values, the evaluation of classical Mondrian with predefined hierarchies shows slightly better results in contrast to basic Mondrian, which partitions data without hierarchical trees.

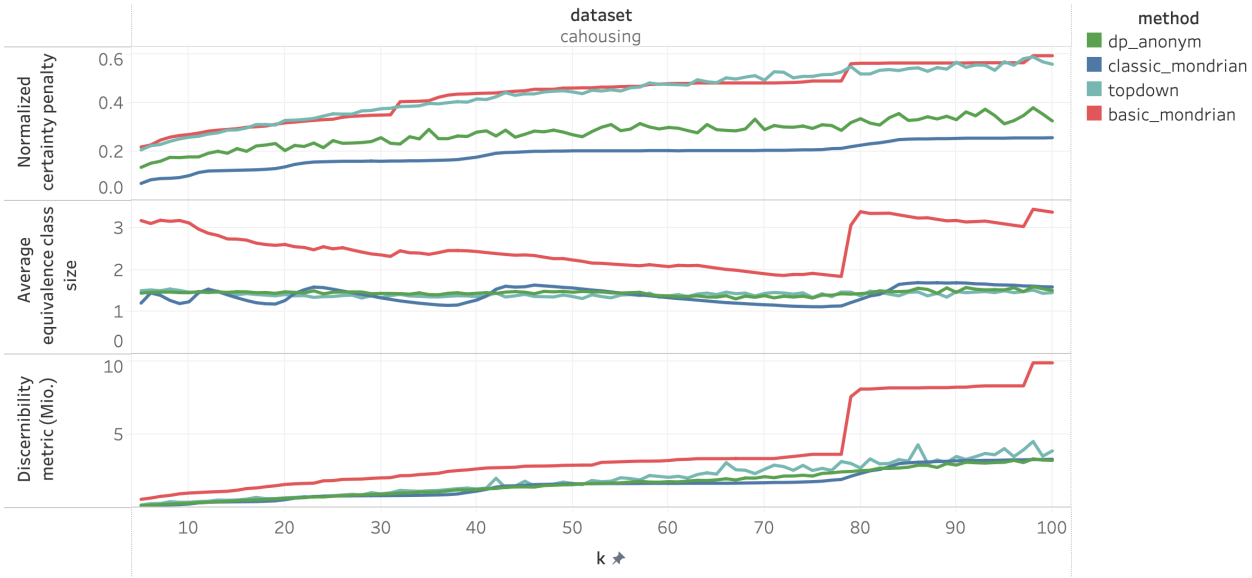


Figure 8.13: Comparison of Data Utility Metrics on "California housing" Dataset among Algorithms: DP-anonym, Basic Mondrian, Classic Mondrian, Top-down.

Figure 8.14: Comparison of Data Utility Metrics on "Contraceptive Method Choice" and "Mammographic Mass" Datasets among Algorithms.

Figure 8.14 shows an analysis of the data anonymization quality of 'Contraceptive method choice" and "Mammographic mass" datasets correspondently. These datasets contain both numeric and categorical attributes with a hierarchy depth of on average 4 levels. Both datasets are not large with sizes of less than 1.5 hundred records. Again in this case, DP-anonym algorithm shows anonymization performances comparable to the leading methods.

Improving the data utility with rising epsilon

The distinctive feature of DP-anonym algorithm is that it provides not only qualitative $k-$anonymity of published data while preserving high data utility, but also a semantic privacy guarantee. As pointed out earlier, DP-anonym algorithm provides epsilon differential privacy, which is a more rigorous mathematical model of preserving privacy.

Figure 8.15 shows two curves for both largest datasets "Adult" and "California housing". Each curve shows an increase in Normalized certainty penalty in relation to the rising privacy parameter k. An important salient feature of this experiment is the use of varying values for differential privacy parameter epsilon. In this example, we consider the algorithm with the lowest acceptable epsilon value of 0.1, the blue line, and the highest epsilon value of 1, the red line. As a reminder, lower epsilon values provide more privacy. An expected improvement in data quality as the epsilon value increases among the presented data utility metrics can be observed. The data quality becomes better because the data undergoes smaller changes by adding noise through the Laplace and exponential mechanisms that are integrated into DP-anonym algorithm.

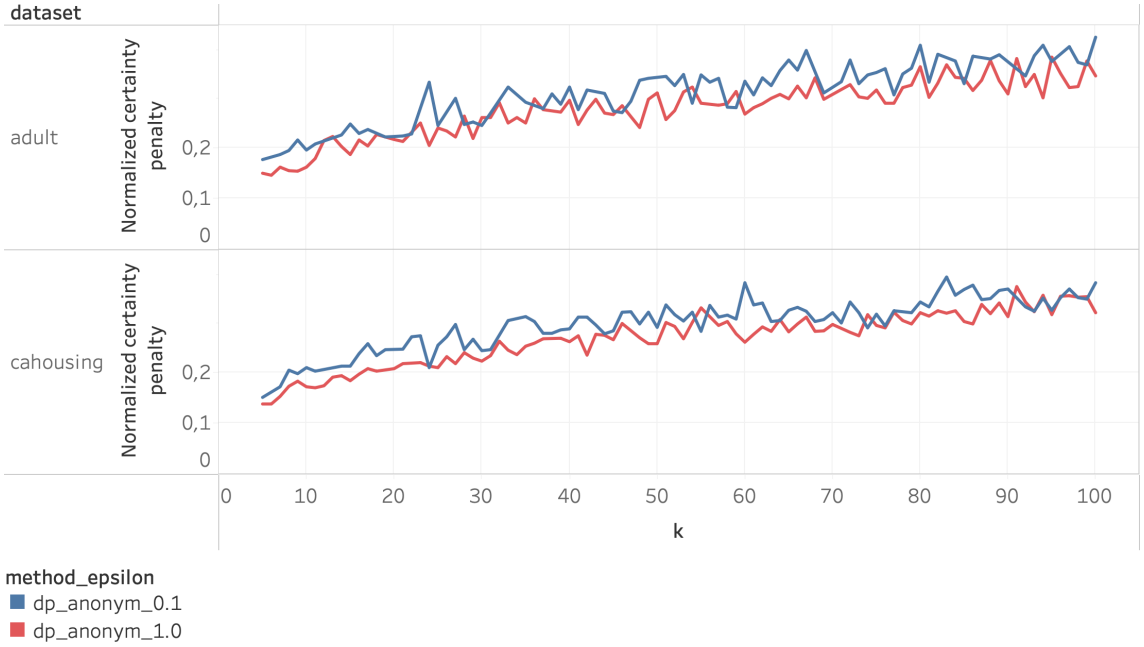


Figure 8.15: Variation of Data Utility Scores using Minimal and Maximal Allowable Epsilon Privacy Parameter in DP-anonym Algorithm.

In the following experiment, the results of which are shown in Figure 8.15, the average estimates of the data utility metrics on the anonymization parameter k values from the interval 2- 100 are given. DP-anonym algorithm with increasing epsilon privacy parameter on the dataset is employed here. In this case, as well, despite the wide variation in the size of equivalence classes (the value of the parameter k), an improvement in data quality is observed with an increase in the epsilon privacy parameter. That is, a minimum epsilon value of 0.1 provides the highest privacy, and consequently, the data changes more and loses more data utility as opposed to applying the highest epsilon value of 1, where less privacy is provided and the data do not lose utility. In addition, in "California housing" dataset, it is worth noting that after the 0.3 mark for epsilon, the values remain almost constant.

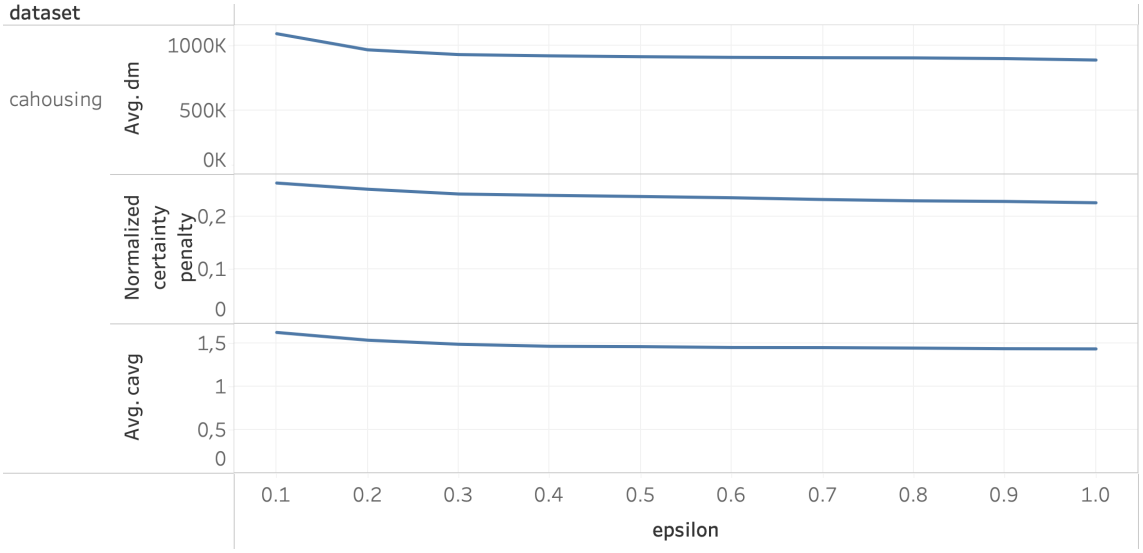


Figure 8.16: Illustration of Data Quality Improvement by Increasing the Epsilon Value in DP-anonym.

From this it can be concluded that in this dataset, it is possible to provide high privacy using epsilon values from 0.1 to 0.3. So increasing epsilon does not bring much quality improvement, because epsilon is an upper possible bound, not an actual bound.

Utility of anonymized data in answering queries

In the next experiment, we look at the quality of data anonymization by means of answering queries. This is a good practice because it is possible to notice how much the data has been generalized and lost the ability to answer queries precisely. In addition, data steward uses these kinds of simple queries to get a better sense of the data.

Let's consider "Adult" dataset and execute the following basic count query: "How many people in the U.S. make less than 50K and more than 50K depending on their race and gender". In SQL it looks like this:

```
SELECT COUNT(*)
FROM Adults
WHERE native-country == ``U.S.''
GROUP BY race, gender
```

This query was executed on the original data, on anonymized table using DP algorithm, and on table anonymized using the Datafly algorithm, with $k = 3$ and $\epsilon = 0.1$. Figure 8.17 illustrates the number of people among races and genders who earn less than or equal to 50K (blue column) and more than 50K (orange column) in the United States. From the graph it can be seen that the number of people from the anonymized table using DP-anonym algorithm does not exactly replicate the answers from the original table, which is again a consequence of the nature of differential privacy. It can be said with certainty, however, that the response after DP-anonym algorithm is comparable to the original. In contrast, responses from Datafly algorithm's anonymized table showed no results at all. This is because even with such a small value of the parameter k, the algorithm has already

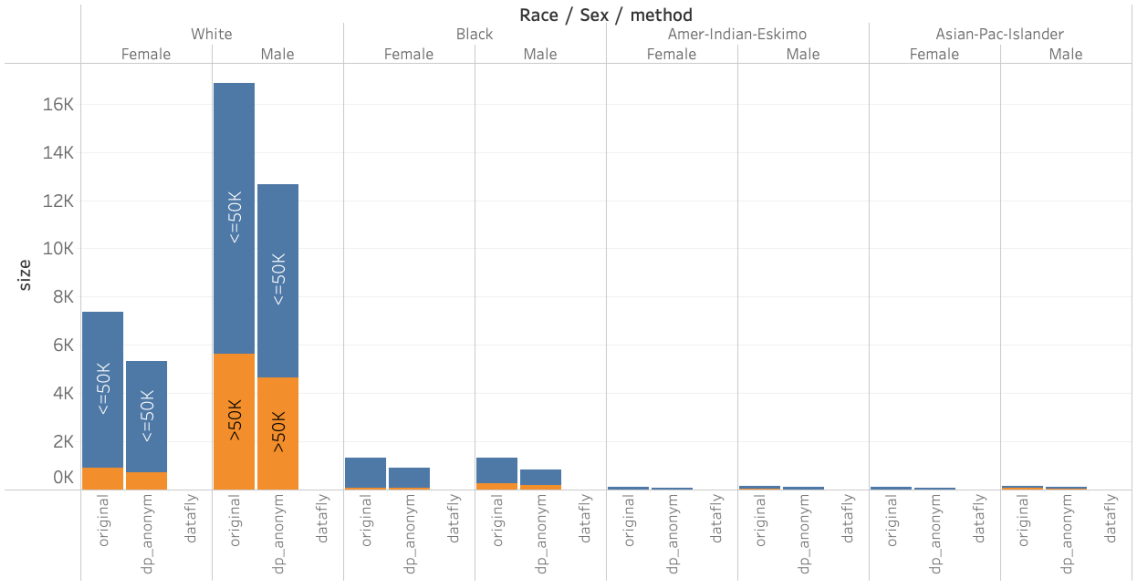


Figure 8.17: Distribution of Salary by Race and Gender in the U.S., "Adult" Dataset

fully generalized the quasi-attribute "native-country" to the highest level. In addition, Datafly uses single-dimensional generalization, which means that all rows of this attribute are generalized at the same level. Thus, the presented DP-anonym algorithm can respond to this query, while the standard anonymization algorithm can not do this.

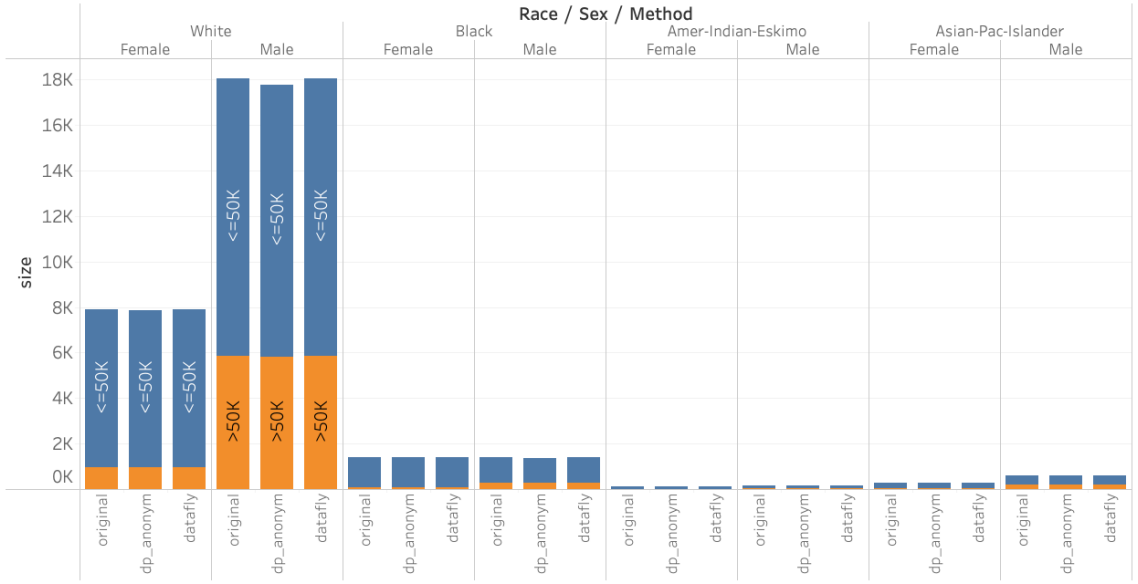


Figure 8.18: Distribution of Salary by Race and Gender in the World, "Adult" Dataset

Figure 8.18 shows the result of the modified query, that is more general. Namely, it ascends two levels in the hierarchy to a "World" or "∗":

```
SELECT COUNT(*)
FROM Adults
GROUP BY race, gender
```

The number of people from the table anonymized by Datafly algorithm can already be distinguished here. In this case, DP-anonym algorithm also demonstrates values close to the original. However, this variant represents extremely aggregated queries, which are unlikely to help the data steward to better understand the data. Thus, the DP-anonym algorithm shows a much better quality of anonymized data than other discussed algorithms, which makes it very useful for common tasks in private data analysis.

8.4 Lessons Learned

The proposed DP-anonym algorithm is discussed in this section in terms of fulfilling the objective of this work. The algorithm performs anonymization of a selected sample from the dataset and satisfies the requirements of $k-$anonymity. From the experimental results, it can be seen that the proposed algorithm demonstrates better or comparable to the leading algorithms data utility, which makes it well suited for pre-processing in a private manner. As the data quality degrades not significantly, the data steward will be able to accomplish the required tasks in processing and then move on to analysis with high data value. DP-anonym anonymizes the data faster than some standard algorithms as it uses median partitioning rather than constructing a large search space. In addition, the algorithm is highly scalable because it uses the frequency set method to find the median. Another distinctive feature is that it can work both with and without the predefined hierarchical trees, in situations where they either cannot be created or are not required by the formulation of the analysis problem. This feature provides a very user-friendly way for the analyst. An important achievement is the use of a multidimensional data generalization model in the proposed algorithm. Therefore, the algorithm generalizes data less and gives much more complete answers to generic database queries, compared to other standard methods.

Another important aspect is that the proposed algorithm provides differential privacy, in contrast to other algorithms. That is, DP-anonym protects data at the syntactic and semantic levels, which is unique in the field of private data publishing.

In this way, the algorithm satisfies and in some aspects exceeds the given requirements for publishing anonymized data to the data steward for pre-processing.

9 Conclusion

In the scope of this study, the possibility of releasing a dataset at the preprocessing stage in a private manner is investigated in order to avoid malicious use of sensitive data of individuals by various parties during data analysis. Consequently, a sampling anonymization method for publishing data DP-anonym ensuring the differential privacy of the algorithm is introduced in this work. The purpose of this method is to provide the data steward with anonymized and private dataset and allow him to be familiar with the dataset and subsequently compose rules to transform the data and transmit it to the data analyst. Since the properties of differential privacy are mostly applied in the field of privacy preserving data mining, query answering, as well as aggregated statistics, this notation is not strictly applicable to an algorithm for releasing anonymized data. However, this work proposes an anonymization algorithm towards achieving $k-$anonymity, in which basic differential privacy mechanisms such as the Laplace and exponential mechanisms are integrated. Due to the post-processing property and the sequential composition of differential privacy, the overall algorithm also satisfies the DP property. The data sampling step preceding the execution of the algorithm is applied to amplify the privateness.

In the experimental part of the work, it is shown that the proposed algorithm outperforms the result of the standard Datafly algorithm by several times, and shows compatible results with the leading Mondrian and Top-down greedy algorithms, in terms of the utility of anonymized data, measured by discernibility metric, average equivalence class size and normalized certainty penalty. Moreover, the proposed algorithm shows much higher payload when answering typical queries, due to the application of multidimensional generalization. In terms of user-friendliness, the algorithm is flexible and allows operating both with and without the pre-defined hierarchy trees for attributes.

The privacy analysis of the algorithm and ensuring $(\epsilon, \delta)-$differential privacy properties are discussed as well. Therefore, the data publishing algorithm proposed in this work has a better and comparable quality of providing $k-$anonymity, in relation to the standard methods. Besides this, it also guarantees differential privacy, which is a stricter mathematical protection of privacy, regardless of the background knowledge of the adversary.

Bibliography

[AMCM+14] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, L. Murphy, et al. “A systematic comparison and evaluation of k-anonymization algorithms for practitioners”. In: *Transactions on data privacy* 7.3 (2014), pp. 337–370 (cit. on pp. 33, 38).

[AP08] C. C. Aggarwal, S. Y. Philip. *Privacy-preserving data mining: models and algorithms*. Springer Science & Business Media, 2008 (cit. on p. 19).

[AS00] R. Agrawal, R. Srikant. “Privacy-preserving data mining”. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 2000, pp. 439–450 (cit. on p. 19).

[BA05] R. J. Bayardo, R. Agrawal. “Data privacy through optimal k-anonymization”. In: *21st International conference on data engineering (ICDE‘05)*. IEEE. 2005, pp. 217–228 (cit. on pp. 21, 53).

[BBG18] B. Balle, G. Barthe, M. Gaboardi. “Privacy amplification by subsampling: Tight analyses via couplings and divergences”. In: *Advances in Neural Information Processing Systems* 31 (2018) (cit. on p. 52).

[BCD+07] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, K. Talwar. “Privacy, accuracy, and consistency too: a holistic solution to contingency table release”. In: *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2007, pp. 273–282 (cit. on p. 41).

[BEM+17] A. Bittau, Ú. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnes, B. Seefeld. “Prochlo: Strong privacy for analytics in the crowd”. In: *Proceedings of the 26th symposium on operating systems principles*. 2017, pp. 441–459 (cit. on p. 50).

[CCF02] M. Charikar, K. Chen, M. Farach-Colton. “Finding frequent items in data streams”. In: *Automata, Languages and Programming: 29th International Colloquium, ICALP 2002 Málaga, Spain, July 8–13, 2002 Proceedings 29*. Springer. 2002, pp. 693–703 (cit. on p. 29).

[CM06] K. Chaudhuri, N. Mishra. “When random sampling preserves privacy”. In: *Annual International Cryptology Conference*. Springer. 2006, pp. 198–213 (cit. on p. 51).

[CSU+19] A. Cheu, A. Smith, J. Ullman, D. Zeber, M. Zhilyaev. “Distributed differential privacy via shuffling”. In: *Advances in Cryptology–EUROCRYPT 2019: 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19–23, 2019, Proceedings, Part I 38*. Springer. 2019, pp. 375–403 (cit. on p. 50).

[CT13] C. Clifton, T. Tassa. “On syntactic anonymity and differential privacy”. In: *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*. IEEE. 2013, pp. 88–93 (cit. on p. 30).

[Dal86] T. Dalenius. “Finding a needle in a haystack or identifying anonymous census records”. In: *Journal of official statistics* 2.3 (1986), p. 329 (cit. on pp. 19, 20).

[DKM+06] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, M. Naor. “Our data, ourselves: Privacy via distributed noise generation”. In: *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*. Springer. 2006, pp. 486–503 (cit. on p. 42).

[DKY17] B. Ding, J. Kulkarni, S. Yekhanin. “Collecting telemetry data privately”. In: *Advances in Neural Information Processing Systems* 30 (2017) (cit. on p. 30).

[DMNS06] C. Dwork, F. McSherry, K. Nissim, A. Smith. “Calibrating noise to sensitivity in private data analysis”. In: *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer. 2006, pp. 265–284 (cit. on pp. 42, 46).

[DR+14] C. Dwork, A. Roth, et al. “The algorithmic foundations of differential privacy”. In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407 (cit. on pp. 41, 45, 48).

[DRSP15] Y.-A. De Montjoye, L. Radaelli, V. K. Singh, A. “ Pentland. “Unique in the shopping mall: On the reidentifiability of credit card metadata”. In: *Science* 347.6221 (2015), pp. 536–539 (cit. on p. 15).

[DS10] C. Dwork, A. Smith. “Differential privacy for statistics: What we know and what we want to learn”. In: *Journal of Privacy and Confidentiality* 1.2 (2010) (cit. on pp. 30, 47).

[DS15] J. Domingo-Ferrer, J. Soria-Comas. “From t-closeness to differential privacy and vice versa in data anonymization”. In: *Knowledge-Based Systems* 74 (2015), pp. 151–158 (cit. on p. 30).

[Dwo08] C. Dwork. “Differential privacy: A survey of results”. In: *International conference on theory and applications of models of computation*. Springer. 2008, pp. 1–19 (cit. on p. 26).

[EPK14] Ú. Erlingsson, V. Pihur, A. Korolova. “Rappor: Randomized aggregatable privacy-preserving ordinal response”. In: *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. 2014, pp. 1054–1067 (cit. on pp. 29, 45).

[FPS96] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. “The KDD process for extracting useful knowledge from volumes of data”. In: *Communications of the ACM* 39.11 (1996), pp. 27–34 (cit. on p. 57).

[FWCY10] B. C. Fung, K. Wang, R. Chen, P. S. Yu. “Privacy-preserving data publishing: A survey of recent developments”. In: *ACM Computing Surveys (Csur)* 42.4 (2010), pp. 1–53 (cit. on pp. 19, 25, 56).

[GHLP12] J. Gehrke, M. Hay, E. Lui, R. Pass. “Crowd-blending privacy”. In: *Advances in Cryptology–CRYPTO 2012: 32nd Annual Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2012. Proceedings*. Springer. 2012, pp. 479–496 (cit. on p. 51).

[HWJ20] J. Hinds, E. J. Williams, A. N. Joinson. “It wouldn‘t happen to me: Privacy concerns and perspectives following the Cambridge Analytica scandal”. In: *International Journal of Human-Computer Studies* 143 (2020), p. 102498 (cit. on p. 15).

[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*. 2002, pp. 279–288 (cit. on p. 54).

[JG17] J. Joy, M. Gerla. “Differential privacy by sampling”. In: *arXiv preprint arXiv:1708.01884* (2017) (cit. on p. 52).

[KLN+11] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, A. Smith. “What can we learn privately?” In: *SIAM Journal on Computing* 40.3 (2011), pp. 793–826 (cit. on p. 51).

[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*. 2005, pp. 49–60 (cit. on pp. 21, 31, 34).

[LDR06a] K. LeFevre, D. J. DeWitt, R. Ramakrishnan. “Mondrian multidimensional k-anonymity”. In: *22nd International conference on data engineering (ICDE‘06)*. IEEE. 2006, pp. 25–25 (cit. on pp. 31, 35, 54).

[LDR06b] K. LeFevre, D. J. DeWitt, R. Ramakrishnan. “Workload-aware anonymization”. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006, pp. 277–286 (cit. on p. 62).

[Lit93] R. J. Little. “Statistical analysis of masked data”. In: *Journal of Official Statistics* 9 (1993), pp. 407–426 (cit. on p. 21).

[LLV06] N. Li, T. Li, S. Venkatasubramanian. “t-closeness: Privacy beyond k-anonymity and l-diversity”. In: *2007 IEEE 23rd international conference on data engineering*. IEEE. 2006, pp. 106–115 (cit. on p. 25).

[LLV09] N. Li, T. Li, S. Venkatasubramanian. “Closeness: A new privacy measure for data publishing”. In: *IEEE Transactions on Knowledge and Data Engineering* 22.7 (2009), pp. 943–956 (cit. on p. 26).

[LQS11] N. Li, W. H. Qardaji, D. Su. “Provably private data anonymization: Or, k-anonymity meets differential privacy”. In: *CoRR, abs/1101.2604* 49 (2011), p. 55 (cit. on p. 52).

[LQS12] N. Li, W. Qardaji, D. Su. “On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy”. In: *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*. 2012, pp. 32–33 (cit. on p. 30).

[McS09] F. D. McSherry. “Privacy integrated queries: an extensible platform for privacy-preserving data analysis”. In: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. 2009, pp. 19–30 (cit. on p. 43).

[MKGV07] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam. “l-diversity: Privacy beyond k-anonymity”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007), 3–es (cit. on p. 24).

[MKM+06] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, J. Y. Halpern. "Worst-case background knowledge for privacy-preserving data publishing". In: *2007 IEEE 23rd International Conference on Data Engineering*. IEEE. 2006, pp. 126–135 (cit. on p. 25).

[MPRV09] I. Mironov, O. Pandey, O. Reingold, S. Vadhan. "Computational differential privacy". In: *Advances in Cryptology-CRYPTO 2009: 29th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 16-20, 2009. Proceedings*. Springer. 2009, pp. 126–142 (cit. on p. 30).

[MT07] F. McSherry, K. Talwar. "Mechanism design via differential privacy". In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE. 2007, pp. 94–103 (cit. on p. 48).

[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*. 2004, pp. 223–228 (cit. on p. 23).

[NRS07] K. Nissim, S. Raskhodnikova, A. Smith. "Smooth sensitivity and sampling in private data analysis". In: *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. 2007, pp. 75–84 (cit. on p. 51).

[NS08] A. Narayanan, V. Shmatikov. "Robust de-anonymization of large sparse datasets". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, pp. 111–125 (cit. on p. 14).

[Pan14] V. Pandurangan. *On taxis and rainbows: Lessons from NYC's improperly anonymized taxi logs*. 2014. URL: https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1 (cit. on p. 14).

[RJR17] K. Rajendran, M. Jayabalan, M. E. Rana. "A study on k-anonymity, l-diversity, and t-closeness techniques". In: *IJCSNS* 17.12 (2017), p. 172 (cit. on p. 31).

[RTG00] Y. Rubner, C. Tomasi, L. J. Guibas. "The earth mover's distance as a metric for image retrieval". In: *International journal of computer vision* 40.2 (2000), p. 99 (cit. on p. 26).

[Sam01] P. Samarati. "Protecting respondents identities in microdata release". In: *IEEE transactions on Knowledge and Data Engineering* 13.6 (2001), pp. 1010–1027 (cit. on p. 21).

[SBB+22] C. Stach, M. Behringer, J. Bräcker, C. Gritti, B. Mitschang. "SMARTEN—A Sample-Based Approach towards Privacy-Friendly Data Refinement". In: *Journal of Cybersecurity and Privacy* 2.3 (2022), pp. 606–628 (cit. on p. 57).

[SD13] J. Soria-Comas, J. Domingo-Ferrer. "Differential privacy via t-closeness in data publishing". In: *2013 Eleventh Annual Conference on Privacy, Security and Trust*. IEEE. 2013, pp. 27–35 (cit. on p. 30).

[SDSM14] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, S. Martınez. "Enhancing data utility in differential privacy via microaggregation-based k-anonymity". In: *The VLDB Journal* 23.5 (2014), pp. 771–794 (cit. on p. 30).

[SGPM20] C. Stach, C. Gritti, D. Przytarski, B. Mitschang. “Trustworthy, secure, and privacy-aware food monitoring enabled by blockchains and the IoT”. In: *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE. 2020, pp. 1–4 (cit. on p. 57).

[SM18] C. Stach, B. Mitschang. “ACCESSORS-A Data-Centric Permission Model for the Internet of Things.” In: *ICISSP* 18 (2018), pp. 30–40 (cit. on p. 13).

[Sta15] C. Stach. “How to Deal with Third Party Apps in a Privacy System–The PMP Gatekeeper–”. In: *2015 16th IEEE International Conference on Mobile Data Management*. Vol. 1. IEEE. 2015, pp. 167–172 (cit. on p. 13).

[Swe00] L. Sweeney. “Uniqueness of simple demographics in the US population”. In: *LIDAP-WP4, 2000* (2000) (cit. on pp. 20, 31).

[Swe01] L. Sweeney. “Computational disclosure control: A primer on data privacy protection”. PhD thesis. Massachusetts Institute of Technology, 2001 (cit. on p. 33).

[Swe02a] L. Sweeney. “Achieving k-anonymity privacy protection using generalization and suppression”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 571–588 (cit. on pp. 21, 33).

[Swe02b] L. Sweeney. “k-anonymity: A model for protecting privacy”. In: *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05 (2002), pp. 557–570 (cit. on pp. 20, 23).

[Tea+17] A. Team et al. “Learning with privacy at scale”. In: *Apple Mach. Learn. J* 1.8 (2017), pp. 1–25 (cit. on p. 29).

[TV06] T. M. Truta, B. Vinay. “Privacy protection: p-sensitive k-anonymity property”. In: *22nd International Conference on Data Engineering Workshops (ICDEW‘06)*. IEEE. 2006, pp. 94–94 (cit. on pp. 23, 24).

[Vad17] S. Vadhan. “The complexity of differential privacy”. In: *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich* (2017), pp. 347–450 (cit. on p. 41).

[VV17] P. Voigt, A. Von dem Bussche. “The EU general data protection regulation (GDPR)”. In: *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10.3152676 (2017), pp. 10–5555 (cit. on p. 13).

[War65] S. L. Warner. “Randomized response: A survey technique for eliminating evasive answer bias”. In: *Journal of the American Statistical Association* 60.309 (1965), pp. 63–69 (cit. on p. 44).

[WBK19] Y.-X. Wang, B. Balle, S. P. Kasiviswanathan. “Subsampled rényi differential privacy and analytical moments accountant”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1226–1235 (cit. on p. 52).

[WBLJ17] T. Wang, J. Blocki, N. Li, S. Jha. “Locally differentially private protocols for frequency estimation”. In: *26th USENIX Security Symposium (USENIX Security 17)*. 2017, pp. 729–745 (cit. on p. 45).

[WFY07] K. Wang, B. Fung, P. S. Yu. “Handicapping attacker‘s confidence: an alternative to k-anonymization”. In: *Knowledge and Information Systems* 11.3 (2007), pp. 345–368 (cit. on p. 23).

[Who14] C. Whong. “FOILing NYC’s taxi trip data”. In: *FOILing NYCs Taxi Trip Data. Np* 18 (2014) (cit. on p. 14).

[XJW+14] L. Xu, C. Jiang, J. Wang, J. Yuan, Y. Ren. “Information security in big data: privacy and data mining”. In: *Ieee Access* 2 (2014), pp. 1149–1176 (cit. on p. 19).

[XT06] X. Xiao, Y. Tao. “Personalized privacy preservation”. In: *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. 2006, pp. 229–240 (cit. on pp. 23, 25).

[XWP+06] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A. W.-C. Fu. “Utility-based anonymization using local recoding”. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006, pp. 785–790 (cit. on pp. 37, 55).

[XZX+13] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, M. Winslett. “Differentially private histogram publication”. In: *The VLDB journal* 22 (2013), pp. 797–822 (cit. on p. 41).

[ZMMS18] Y. Zou, A. H. Mhaidli, A. McCall, F. Schaub. “I‘ve Got Nothing to Lose: Consumers‘ Risk Perceptions and Protective Actions after the Equifax Data Breach.” In: *SOUPS@ USENIX Security Symposium*. 2018, pp. 197–216 (cit. on p. 15).

All links were last followed on 20 of May, 2023.

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