WHITEPAPER

ANONYMISATION

A PRACTICAL GUIDE

MAY 2020

ALEXANDRA INSTITUTE

AUTHOR

Zaruhi Aslanyan, Alexandra Institute

Published by

THE ALEXANDRA INSTITUTE

May 2020

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

With the increasing adoption of technological devices for work, study and leisure, the amount of collected data increases as well. Gartner [1] states that "people are willing to provide companies with information in exchange for convenience and personalised experiences". In fact, collecting data allows companies to improve their services and tune them to their customers' specific needs, but it also spawns great challenges. Data breaches, privacy loss, and the corresponding reputational damage are just some examples of these challenges, only recently tackled by data protection regulations (e.g., GDPR).

Data anonymisation aims at protecting the privacy of an individual in a given dataset, and it is a key enabler for securing collected data and comply with data protection regulations. When properly implemented, anonymisation can minimse personal data security breaches, whilst allowing to reap the benefits of collecting and storing users' information.

In this paper, we introduce the concept of anonymisation. We present a wide range of anonymisation techniques to support data anonymisation and evaluation of data privacy, and identify their weaknesses and strengths. Moreover, we review a number of open-source and commercial tools that implement some of these techniques.

The paper is intended as a practical guide to anonymisation for a reader without prerequisite knowledge of data privacy. However, it may also be of interest to the specialist as an overview of the subject. In particular, it may prove useful to profiles responsible for data privacy and security, such as data controller, security manager or data protection officer (DPO).

2 INTRODUCTION

We move more and more towards the digitalised world, where data is at core of everything. Data transform the way we live, work and socialise. Companies collect and use data to improve their services and customer experience, develop new business models, and measure the productivity of their employees and processes. On the other hand, end users provide and utilise data to access services and communicate with the rest of the world. For example, healthcare diagnostic equipment increasingly relies on patient devices that collect patient data to be used for analysis and diagnosis, and for providing recommendations and instructions tailored to the needs of a specific individual.

The amount of data that we collect and create every day is increasing dramatically. According to the International Data Corporation (IDC, Figure 1) the total amount of data created in the world will grow from 33 zettabytes (ZB) in 2018 to 175 ZB by 2025.

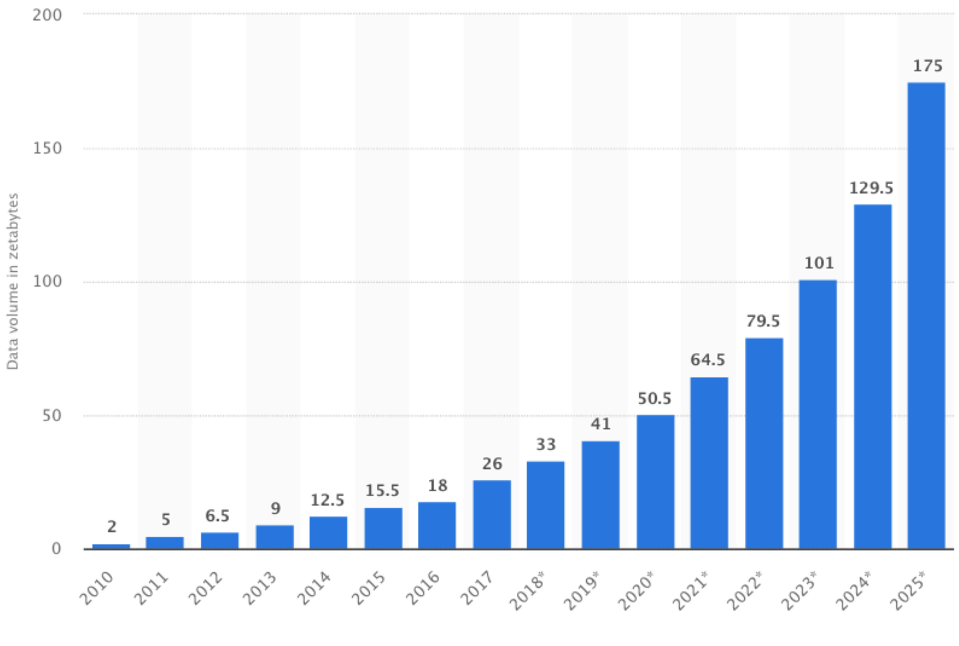


Figure 1: Volume of data/information created worldwide from 2010 to 2025 (in zettabytes). Illustration from https://www.statista.com/statistics/871513/worldwide-data-created

With the advantages of data collection, enormous issues come as well. Cyberattacks and data breaches become increasingly common and costly to handle. Cybercriminals want to steal data that can be used to identify individuals or other valuable information that can be sold. Virtually no company or institution is immune to cyberattacks and data breaches, as the attackers' motives vary from ransom to revenge and vanity.

Figure 2 illustrates the world's biggest data breaches and hacks documented in the last few years. As we can see from the picture, cyberattacks affect companies of all sizes. For example, in September 2019 a database containing Facebook accounts was hacked and 420 million records were exposed. The records contained Facebook users' unique ID and phone number. Moreover, some of the records contained the user's name, gender, and country [2].

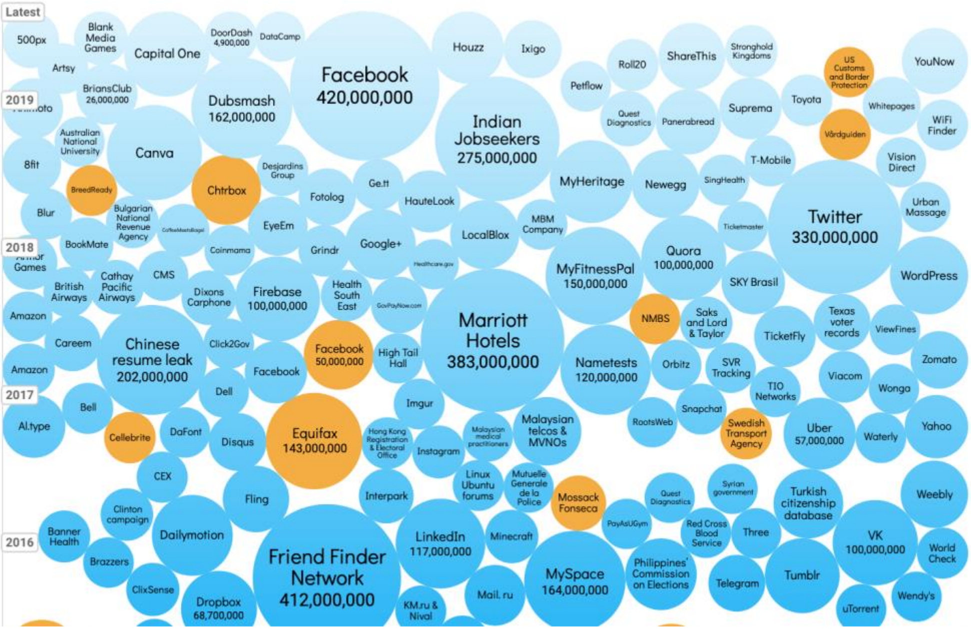


Figure 2: World's Biggest Data Breaches & Hacks. Illustration from https://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks

In this context, a critical issue is the protection of personally identifiable information (PII), that is, data which can be used to identify a given person. Besides protecting PII from potential leakage, companies should comply with various data protection regulations, such as the General Data Protection Regulation (GDPR) in Europe. In case of violation of GDPR, the imposed penalties can be significant. One such recent violation of GDPR resulted in a fine of 9.55 million euro for "insufficient technical and organisational measures to ensure information security" to the German telecom provider 1&1 Telecommunication (December 2019) [3].

Besides threatening the end users to whom the information refers to, cyberattacks bring legal, financial and reputational risks to the organisations collecting data. Hence, the need for protecting and anonymising collected data cannot be stressed enough. Simply removing PII from a dataset, such as names and addresses, would not suffice, for there are other indirect or quasi-identifiers that can be used to pinpoint an individual in a dataset. A known example is the re-identification of William Weld, governor of Massachusetts, from presumably anonymised health data records containing only birth date, sex and ZIP code, in a study carried out by Latanya Sweeney [4].

Various techniques and methods are used to overcome the above-mentioned issues. Data anonymisation is one such method, which can be defined as *the process of protecting personal data by means of irreversibly modifying and/or erasing them*.

According to Recital 26 of the GDPR, “The principles of data protection should […] not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.” In other words, when anonymisation is properly applied, it positions the processing and storage of personal data outside the scope of the GDPR.

Data anonymisation is a complex task. Even understanding which data is sensitive and needs extra protection can be a challenge. A poorly anonymised dataset can still be subject to data leakage, as in the previous example. Thus, it is important to perform anonymisation with due care.

The choice of an anonymisation technique depends on various factors, such as the scenario/use-case, the data, the use of data, the discloser of data, the law, etc. Understanding the context of data will help identify the right techniques. In Appendix B, we present some questions that can be used as a guide to understanding the data and the context in which the data will be used. Note that the legal requirements and limitations as to what one can and cannot do with data are outside of the scope of this paper.

It is worthwhile observing that perfectly anonymised data is not useful. We should always balance between privacy and utility of the database, as presented in Figure 3.

Though challenging, anonymisation needs to be applied aiming at providing data privacy while preserving utility, and this paper will guide the reader in this task.

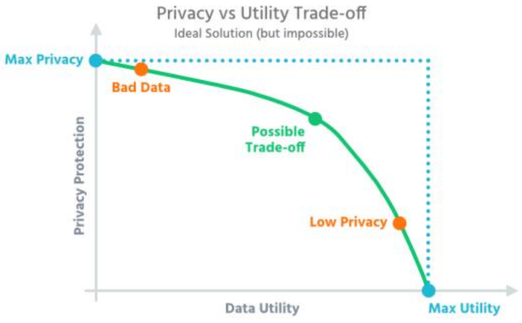


Figure 3: Trade-off between privacy and utility.
Illustration from https://aircloak.com/background/analytics-and-privacy/power-of-anonymization

The paper provides an overview of key concepts in anonymisation to raise awareness in the reader, focusing on personal data. The scope of this paper is two-fold. On one hand, the paper guides organisations and individuals who want to anonymise their data to understand some of the available techniques and suggests anonymisation methods based on the type of data and context. On the other hand, the anonymisation techniques and tools we present, as well as a set of self-assessment questions, assist organisations and individuals who already have anonymised data and want to evaluate their dataset. Throughout the paper, a simple running example is used to explain and clarify the various topics.

In Section 3, we present the terminology used throughout the paper. Anonymisation techniques together with the running example are described in Section 4. State-of-the-art open-source and commercial tools are discussed in Section 5. Finally, we conclude in Section 6.

3 DEFINITIONS

Private data – The information that an individual does not want to make public. This could be, for instance, name, address, phone number, emails, a medical observation, etc.

Personal data – In GDPR article 4, personal data is defined as "any information relating to an identified or identifiable natural person ('data subject') who can be directly or indirectly identified".

Sensitive data – In GDPR sensitive data is defined as any data that reveal:

- Racial or ethnic origin
- Political opinions
- Religious or philosophical beliefs
- Trade union membership
- Genetic data
- Biometric data for the purpose of uniquely identifying a natural person
- Data concerning health or a natural person's sex life and/or sexual orientation

Attributes – An attribute is a column of a table in a database, such as gender, school, or city.

Record – A record is a row of a table in a database, i.e., a tuple of attribute values for an individual.

Sensitive attributes – A sensitive attribute is an attribute whose value must be protected from disclosure to any person who has no direct access to the original dataset. An example of sensitive attribute is occupation, disease or medical status attributes.

Quasi-identifiers – Quasi-identifiers are sets of attributes that might allow an attacker to uniquely identify an individual by means of linking such attributes to additional information/data.

De-anonymisation (re-identification) – Is the process of cross-referencing an anonymised dataset with other additional information/data in order to re-identify the anonymous data.

4 ANONYMISATION TECHNIQUES/METHODS

Various anonymisation methods, such as *k*-anonymity, *l*-diversity and differential privacy have been proposed to make it hard for an adversary to learn anything about individuals from an anonymised dataset. Table 1 compares different anonymisation techniques with respect to key high-level properties, such as the type of data being anonymised or how much the data is altered with respect to the original data.

Technique	**Data type being anonymised**	**How will the anonymisation help?**	**Is the data altered?**	**Complexity**	**Weaknesses**
k-anonymity	Quasi-identifiers	Prevent linkage attacks and re-identification of the quasi-identifiers	Yes, by removing or manipulating given attributes	Easy to implement and intuitive	Focuses only on the quasi-identifiers, non-trivial choice of *k*
l-diversity	Sensitive attributes	Prevent homogeneity and background knowledge attacks and attribute disclosure	No	Easy to implement and intuitive	Focuses only on the sensitive attributes, non-trivial choice of *l*
t-closeness	Sensitive attributes	Prevent skewness and similarity attacks	No	Less intuitive, has complex computation procedure	Focuses only on the sensitive attributes, non-trivial computation of *t*
Differential privacy	Quasi-identifier, sensitive attributes	Prevent attacks with background information, such as compositional attacks	Yes	Non-intuitive	Non-trivial determination of noise; number of analysis depends on the privacy budget

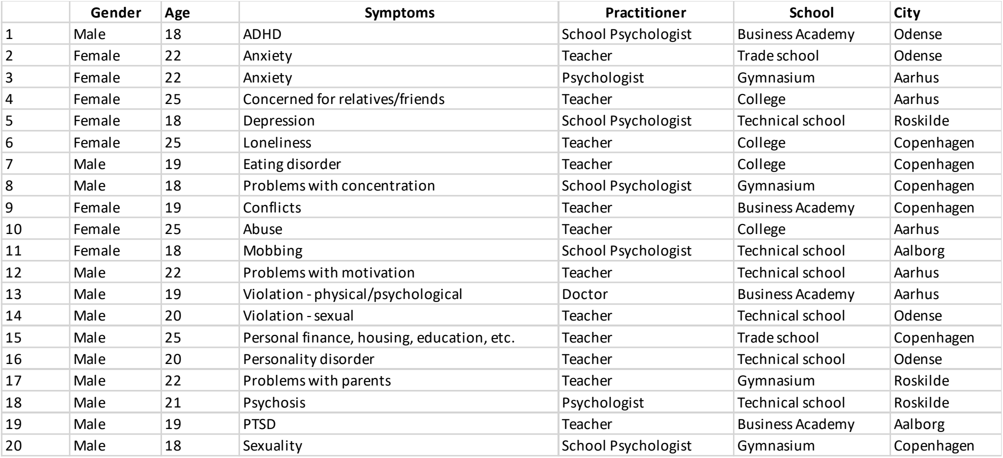
Table 1: Comparison of anonymisation techniques

In the following, we briefly describe the anonymisation techniques from Table 1.

4.1 RUNNING EXAMPLE

In this section, we present a toy example of a healthcare dataset. This example will serve to explain various anonymisation methods throughout the paper.

Consider a medical institution that wants to share part of its dataset for the purpose of statistical analysis. The dataset contains information about students' health records, presented in Table 2. In order to protect the privacy of an individual in the dataset and to comply with GDPR, the institution needs to anonymise the dataset before sharing it.



	Gender	Age	Symptoms	Practitioner	School	City
1	Male	18	ADHD	School Psychologist	Business Academy	Odense
2	Female	22	Anxiety	Teacher	Trade school	Odense
3	Female	22	Anxiety	Psychologist	Gymnasium	Aarhus
4	Female	25	Concerned for relatives/friends	Teacher	College	Aarhus
5	Female	18	Depression	School Psychologist	Technical school	Roskilde
6	Female	25	Loneliness	Teacher	College	Copenhagen
7	Male	19	Eating disorder	Teacher	College	Copenhagen
8	Male	18	Problems with concentration	School Psychologist	Gymnasium	Copenhagen
9	Female	19	Conflicts	Teacher	Business Academy	Copenhagen
10	Female	25	Abuse	Teacher	College	Aarhus
11	Female	18	Mobbing	School Psychologist	Technical school	Aalborg
12	Male	22	Problems with motivation	Teacher	Technical school	Aarhus
13	Male	19	Violation - physical/psychological	Doctor	Business Academy	Aarhus
14	Male	20	Violation - sexual	Teacher	Technical school	Odense
15	Male	25	Personal finance, housing, education, etc.	Teacher	Trade school	Copenhagen
16	Male	20	Personality disorder	Teacher	Technical school	Odense
17	Male	22	Problems with parents	Teacher	Gymnasium	Roskilde
18	Male	21	Psychosis	Psychologist	Technical school	Roskilde
19	Male	19	PTSD	Teacher	Business Academy	Aalborg
20	Male	18	Sexuality	School Psychologist	Gymnasium	Copenhagen

Table 2: Healthcare dataset.

4.2 K-ANONYMITY

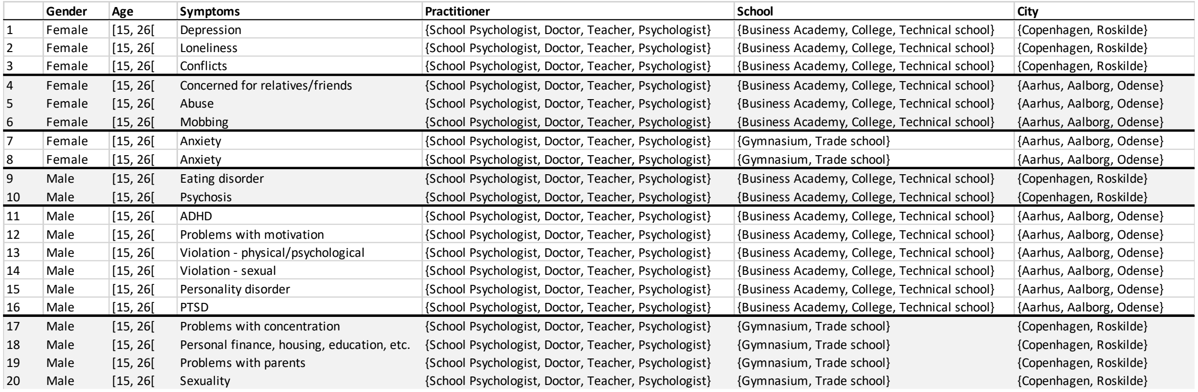
Before releasing a dataset, the first step is to remove all direct identifiers, such as name, address and phone numbers. However, in most of the cases this is not enough. Only removing the direct identifiers will still allow an attacker to re-identify an individual in the dataset by linking the data with other additional information that he/she has.

For example, consider the dataset in Table 2, where all direct identifiers are removed. An attacker knowing some additional information about an individual, e.g., that Trine, 22 years old female living in Aarhus, is in this dataset, can conclude that Trine has anxiety (from row 3), as in the dataset there is only one 22 years old female leaving in Aarhus.

In order to avoid such cases, additional anonymisation steps should be enforced. Several methods have been studied that make it harder for an attacker to learn anything about an individual from an anonymised dataset. One such method is *k-anonymity*.

k-anonymity is an anonymisation method that ensures the information about an individual in the published dataset cannot be distinguished from at least k-1 other individuals in the same dataset. In other words, it divides a dataset into so-called *equivalence classes*, where in each class the records have identical values for all quasi-identifiers. This method focuses on quasi-identifiers, meaning that anonymisation is applied to the quasi-identifier attributes, making these attributes “imprecise”. If an attacker knows only the values of quasi-identifiers of an individual, he/she cannot identify the individual in k-anonymised dataset with high degree of certainty.

Consider the dataset from Table 2, where the attributes “gender”, “age”, “practitioner”, “school” and “city” are quasi-identifiers, while “symptoms” is a sensitive attribute. We apply 2-anonymity to the dataset, hence dividing it into equivalence classes where each record cannot be distinguished from at least 1 other record. In other words, each record in the dataset indistinctly belongs to at least 2 individuals with respect to quasi-identifiers. Table 3 represents the corresponding anonymised dataset with highlighted equivalence classes.



	Gender	Age	Symptoms	Practitioner	School	City
1	Female	[15, 26[	Depression	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Copenhagen, Roskilde}
2	Female	[15, 26[	Loneliness	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Copenhagen, Roskilde}
3	Female	[15, 26[	Conflicts	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Copenhagen, Roskilde}
4	Female	[15, 26[	Concerned for relatives/friends	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Aarhus, Aalborg, Odense}
5	Female	[15, 26[	Abuse	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Aarhus, Aalborg, Odense}
6	Female	[15, 26[	Mobbing	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Aarhus, Aalborg, Odense}
7	Female	[15, 26[	Anxiety	{School Psychologist, Doctor, Teacher, Psychologist}	{Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}
8	Female	[15, 26[	Anxiety	{School Psychologist, Doctor, Teacher, Psychologist}	{Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}
9	Male	[15, 26[	Eating disorder	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Copenhagen, Roskilde}
10	Male	[15, 26[	Psychosis	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Copenhagen, Roskilde}
11	Male	[15, 26[	ADHD	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Aarhus, Aalborg, Odense}
12	Male	[15, 26[	Problems with motivation	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Aarhus, Aalborg, Odense}
13	Male	[15, 26[	Violation - physical/psychological	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Aarhus, Aalborg, Odense}
14	Male	[15, 26[	Violation - sexual	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Aarhus, Aalborg, Odense}
15	Male	[15, 26[	Personality disorder	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Aarhus, Aalborg, Odense}
16	Male	[15, 26[	PTSD	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school}	{Aarhus, Aalborg, Odense}
17	Male	[15, 26[	Problems with concentration	{School Psychologist, Doctor, Teacher, Psychologist}	{Gymnasium, Trade school}	{Copenhagen, Roskilde}
18	Male	[15, 26[	Personal finance, housing, education, etc.	{School Psychologist, Doctor, Teacher, Psychologist}	{Gymnasium, Trade school}	{Copenhagen, Roskilde}
19	Male	[15, 26[	Problems with parents	{School Psychologist, Doctor, Teacher, Psychologist}	{Gymnasium, Trade school}	{Copenhagen, Roskilde}
20	Male	[15, 26[	Sexuality	{School Psychologist, Doctor, Teacher, Psychologist}	{Gymnasium, Trade school}	{Copenhagen, Roskilde}

Table 3: Example of k-anonymity, where k=2.

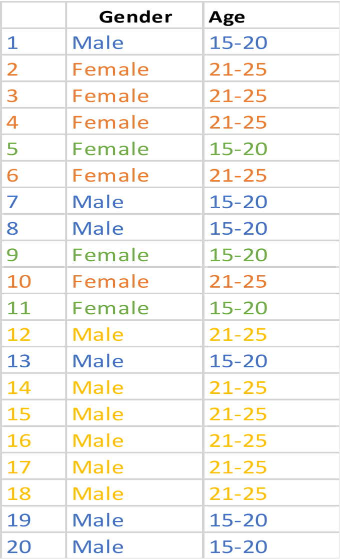
There are different techniques to achieve k-anonymity, the most common being suppression and generalisation of information.

Generalisation

As the name suggests, generalisation makes the values of quasi-identifiers more general so that records with different values are transformed into records with identical values. In this way, we remove specificity from the data and make it less precise.

The transformation of values is done according to the corresponding domain, such that numerical values are transformed into intervals or to more generic values from a numerical domain.

Let us look into the first two columns of Table 2, namely "gender" and "age". Table 4 shows the generalisation of the attribute "age", where the precise values of age have been transformed into intervals, so that the resulting table satisfies 2-anonymity. Different colours represent different equivalence classes.



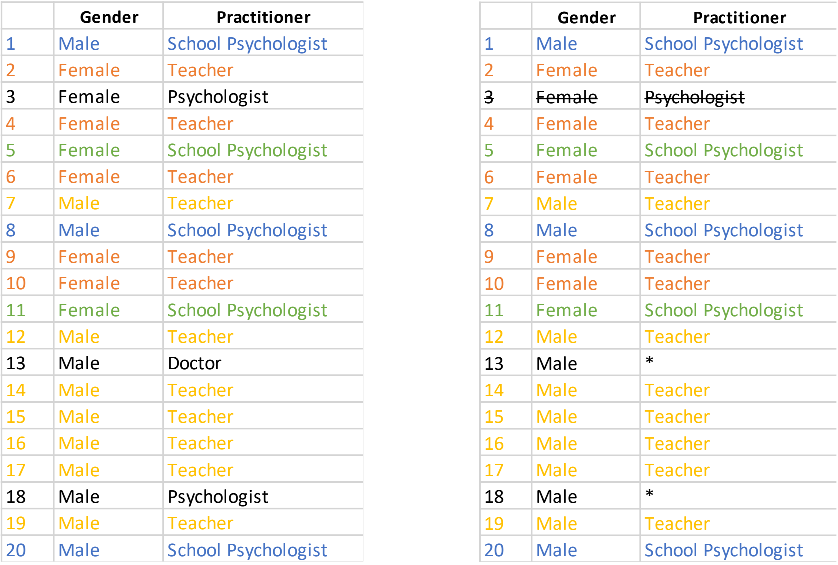
	Gender	Age
1	Male	15-20
2	Female	21-25
3	Female	21-25
4	Female	21-25
5	Female	15-20
6	Female	21-25
7	Male	15-20
8	Male	15-20
9	Female	15-20
10	Female	21-25
11	Female	15-20
12	Male	21-25
13	Male	15-20
14	Male	21-25
15	Male	21-25
16	Male	21-25
17	Male	21-25
18	Male	21-25
19	Male	15-20
20	Male	15-20

Table 4: Example of generalisation of the attribute "Age".

Suppression

In some cases, it is not possible to generalise a record or a particular value in the record because of its uniqueness. Such outlier records or values are usually either removed/omitted or replaced by * (hence the value is lost). This process is called suppression.

Let us consider an example dataset in Table 5(a). In order to obtain 2-anonymity, we group the records such that in each group there are at least 2 other records with the same values. The groups are highlighted with different colours. We can see that rows 3, 13 and 18 cannot be paired with any groups, i.e., there is no other record with the same values. Using the suppression technique and replacing the values for the attribute "practitioner" in rows 13 and 18 with *, we can group these two rows together into an equivalence class. The same approach is not viable with row 3 as by replacing one of the attributes of the record with *, we still would not match any other record in the dataset. Hence, we remove/omit the outlier record (the row 3) from the dataset. The result of the suppression is illustrated in Table 5(b).



	Gender	Practitioner
1	Male	School Psychologist
2	Female	Teacher
3	Female	Psychologist
4	Female	Teacher
5	Female	School Psychologist
6	Female	Teacher
7	Male	Teacher
8	Male	School Psychologist
9	Female	Teacher
10	Female	Teacher
11	Female	School Psychologist
12	Male	Teacher
13	Male	Doctor
14	Male	Teacher
15	Male	Teacher
16	Male	Teacher
17	Male	Teacher
18	Male	Psychologist
19	Male	Teacher
20	Male	School Psychologist

(a)

	Gender	Practitioner
1	Male	School Psychologist
2	Female	Teacher
~~3~~	~~Female~~	~~Psychologist~~
4	Female	Teacher
5	Female	School Psychologist
6	Female	Teacher
7	Male	Teacher
8	Male	School Psychologist
9	Female	Teacher
10	Female	Teacher
11	Female	School Psychologist
12	Male	Teacher
13	Male	*
14	Male	Teacher
15	Male	Teacher
16	Male	Teacher
17	Male	Teacher
18	Male	*
19	Male	Teacher
20	Male	School Psychologist

(b)

Table 5: Example of suppression of the attribute "Practitioner".

Pros and cons of k-anonymity

As with every method, there are advantages and drawbacks of k-anonymity.

Advantages:

- K-anonymity allows to release data while keeping the quasi-identifiers of individuals anonymous, making it harder to re-identify an individual. Hence, it is suitable for a dataset with quasi-identifier attributes.
- The method ensures the anonymity of an individual with respect to quasi-identifier against linkage attacks, i.e., attacks where an anonymised dataset is linked with additional similar data to get the combined overall information about an individual.
- The method is suitable for cases where the quasi-identifier can be easily generalised and even removed.
- K-anonymity is a simple and intuitive method; thus, it is easy to understand, implement and use.

Drawbacks:

- The method is focusing only on the attributes deemed to be quasi-identifiers. In some cases, this will lead to possible re-identification based on sensitive attributes. We will discuss this in more detail in Section 4.3.
- The choice of the parameter *k* is not always trivial and depends on the type of data and the purpose of the anonymisation. For example, Statistics Denmark uses k=3 when showing information about a group and k=5 when computing statistics [5]. In general, high

values of the parameter k lower the chance of re-identification. However, one should not forget about utility, which can be affected with the increase of the value of k.

The k-anonymity process alters the original dataset either by losing some records or by manipulating their content. Hence, it is not suitable for cases in which data cannot be changed.

For more information about k-anonymity we refer to [6] [7].

4.3 L-DIVERSITY

K-anonymity makes it harder for an attacker to re-identify an individual in a dataset. However, in some cases it is still possible to re-identify an individual based on the sensitive attributes of the individual, or, if the attacker knows that an individual is in a given dataset, in some cases it is still possible to learn some of his/her sensitive attributes. Let us discuss two possible attacks.

Homogeneity attacks: Consider the 2-anonymous dataset in Table 3. Assume an attacker knows that Trine, a 22-year-old female living in Aarhus and studying in a gymnasium, is in the dataset. Therefore, the attacker knows that Trine's record is either row 7 or row 8. As both records have the same values for the sensitive attributes, the attacker can conclude that Trine has anxiety.

In this type of attacks, even though the dataset is k-anonymous, it is still possible to predict the values of sensitive attributes due to a lack of their diversity, i.e., the sensitive attributes are the same for a pair of similar entries.

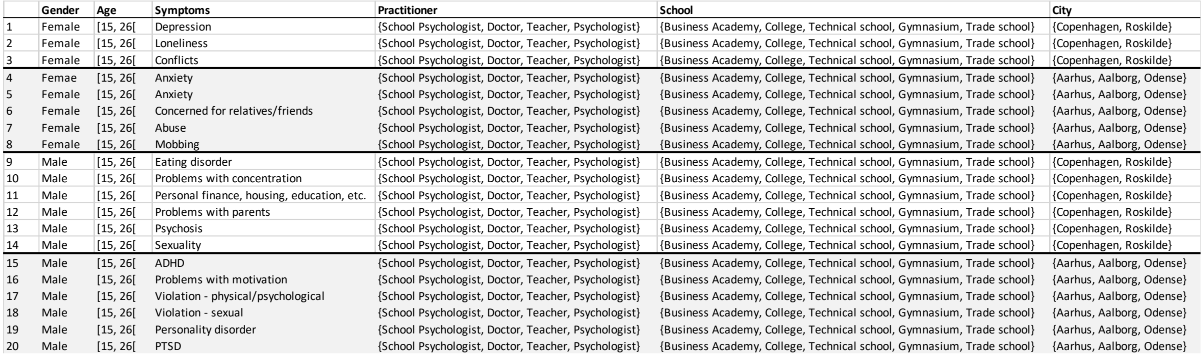
Background Knowledge Attacks: Let us again consider the 2-anonymous dataset in Table 3. An attacker knows that Lars, a male living in Copenhagen and attending a technical school, is in the dataset. Therefore, the attacker knows that Lars's record is either row 9 or row 10. Moreover, an attacker knowing Lars's habits knows that he does not have any eating disorder. This additional background knowledge allows the attacker to conclude that Lars has psychosis.

In this type of attacks, an attacker is able to predict the values of the sensitive attributes due to some background knowledge about an individual in a dataset. Background knowledge can be, for example, knowing that some individual is in a dataset or knowing additional information about an individual in a dataset.

To overcome these issues, a stronger notion of anonymity is needed. From the above-mentioned attacks, we see that they occur as k-anonymity focuses on quasi-identifiers and does not anonymise and/or modify the sensitive attributes. Hence, we need to assure that all records in an equivalence class have different values for the sensitive attributes. This is achieved through *l-diversity*.

A dataset is called l-diverse if every equivalence class has at least l ≥ 2 different values for the sensitive attributes. In this case, we say that the values of sensitive attributes in each equivalence class are at least l-well-defined (l different values). In contrast to k-anonymity, l-diversity focuses on sensitive attributes.

Let us consider the example dataset from Table 2. In order to obtain 3-diversity, the equivalence classes should be merged so that in each class the values for the sensitive attribute “symptoms” are at least 3-well-defined (different). Table 6 presents the anonymised table satisfying 3-diversity with highlighted equivalence classes. Table contains 4 equivalence classes where in each class there are at least 3 different values for the sensitive attributes.



	Gender	Age	Symptoms	Practitioner	School	City
1	Female	[15, 26[	Depression	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Copenhagen, Roskilde}
2	Female	[15, 26[	Loneliness	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Copenhagen, Roskilde}
3	Female	[15, 26[	Conflicts	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Copenhagen, Roskilde}
4	Femae	[15, 26[	Anxiety	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}
5	Female	[15, 26[	Anxiety	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}
6	Female	[15, 26[	Concerned for relatives/friends	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}
7	Female	[15, 26[	Abuse	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}
8	Female	[15, 26[	Mobbing	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}
9	Male	[15, 26[	Eating disorder	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Copenhagen, Roskilde}
10	Male	[15, 26[	Problems with concentration	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Copenhagen, Roskilde}
11	Male	[15, 26[	Personal finance, housing, education, etc.	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Copenhagen, Roskilde}
12	Male	[15, 26[	Problems with parents	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Copenhagen, Roskilde}
13	Male	[15, 26[	Psychosis	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Copenhagen, Roskilde}
14	Male	[15, 26[	Sexuality	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Copenhagen, Roskilde}
15	Male	[15, 26[	ADHD	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}
16	Male	[15, 26[	Problems with motivation	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}
17	Male	[15, 26[	Violation - physical/psychological	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}
18	Male	[15, 26[	Violation - sexual	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}
19	Male	[15, 26[	Personality disorder	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}
20	Male	[15, 26[	PTSD	{School Psychologist, Doctor, Teacher, Psychologist}	{Business Academy, College, Technical school, Gymnasium, Trade school}	{Aarhus, Aalborg, Odense}

Table 6: Example of l-diversity, where l=3.

From Table 6 we see that the above-mentioned attacks are prevented by 3-diversity. For example, assume an attacker knows that Trine’s record is one of the rows from 4 to 8. However, he/she cannot conclude immediately the value for the sensitive attributes for Trine. As the dataset is 3-diverse, an attacker needs to know at least 2 (in general, l-1) additional attributes for identifying the value for Trine. Note that the larger the value for l, the more background knowledge an attacker needs to have.

Pros and cons of l-diversity

Advantages:

- L-diversity ensures the anonymity of an individual against homogeneity and background knowledge attacks.
- L-diversity together with k-anonymity provides a robust privacy model for a great many cases.
- L-diversity does not alter the original values of the sensitive attributes.

Drawbacks:

- L-diversity focuses only on the sensitive attributes; hence it should be applied together with other anonymisation methods such as k-anonymity.
- Similarly to k-anonymity, the choice of the parameter l is not trivial and depends on different aspects, such as a dataset.

For a more detailed explanation of l-diversity and its various flavours we refer to [8].

4.4 T-CLOSENESS

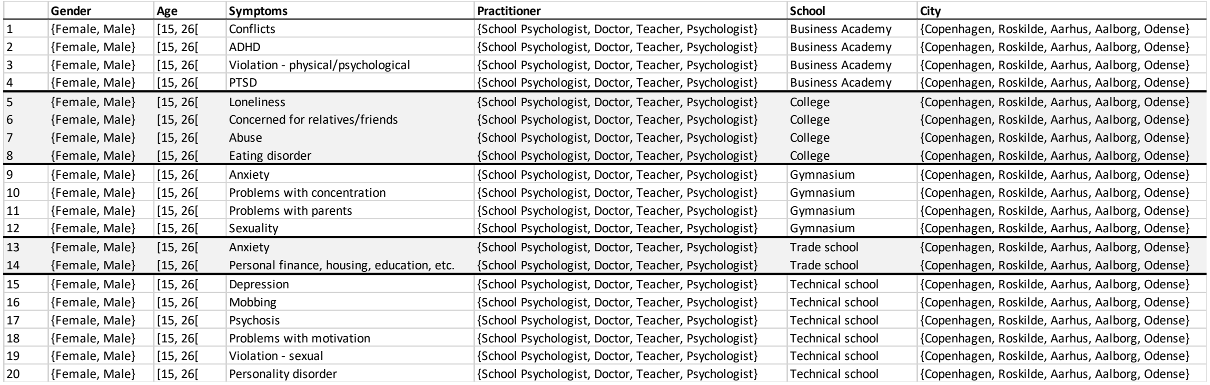
It is possible to predict the values for the sensitive attributes for an individual in an l-diverse dataset when they are either asymmetric (skewness attack), meaning there are only few values for given attributes, or they are semantically close (similarity attacks), meaning that it is possible to map a set of values to a more general category/class.

Let us consider the 3-diverse dataset in Table 6. Assume an attacker knows that Anne's record is one of the rows from 1 to 3. Therefore, the attacker can conclude that Anne has some kind of mood or psychiatric disorder as all three values for the sensitive attribute "symptoms" belong to the general category of mood or psychiatric disorders.

Even though the attacker does not learn the exact value of the sensitive attribute, he/she concludes general sensitive information about an individual, such as a disorder category. This occurs because the values of the sensitive attributes in an equivalence class are so semantically close that in fact, they disclose some information.

In order to overcome this challenge, *t-closeness* has been proposed. The method requires that the distribution of the values of a sensitive attribute in an equivalence class be close to the distribution of that attribute in the overall dataset. The parameter *t* represents the maximum threshold of the distance between these two distributions. T-closeness prevents the above-mentioned attacks; however, it has a complex computational procedure.

Table 7 presents the anonymised version of the dataset from Table 2 using t-closeness, where t = 0.85, meaning that the distance between the distribution of the sensitive attribute "symptoms" in an equivalence class and the distribution of that attribute in the overall dataset is at most 0.85. From the figure we can see that the above-mentioned attacks are prevented by t-closeness. The table contains 5 equivalence classes, and in each class the values for the sensitive attributes are semantically different, i.e., it is not possible to group them into a category that would disclose unambiguous information. For example, assuming that an attacker knows that Anne is in the first equivalence class, he/she would not be able to conclude that Anne has some kind of mood or psychiatric disorder.



	Gender	Age	Symptoms	Practitioner	School	City
1	{Female, Male}	[15, 26[	Conflicts	{School Psychologist, Doctor, Teacher, Psychologist}	Business Academy	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
2	{Female, Male}	[15, 26[	ADHD	{School Psychologist, Doctor, Teacher, Psychologist}	Business Academy	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
3	{Female, Male}	[15, 26[	Violation - physical/psychological	{School Psychologist, Doctor, Teacher, Psychologist}	Business Academy	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
4	{Female, Male}	[15, 26[	PTSD	{School Psychologist, Doctor, Teacher, Psychologist}	Business Academy	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
5	{Female, Male}	[15, 26[	Loneliness	{School Psychologist, Doctor, Teacher, Psychologist}	College	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
6	{Female, Male}	[15, 26[	Concerned for relatives/friends	{School Psychologist, Doctor, Teacher, Psychologist}	College	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
7	{Female, Male}	[15, 26[	Abuse	{School Psychologist, Doctor, Teacher, Psychologist}	College	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
8	{Female, Male}	[15, 26[	Eating disorder	{School Psychologist, Doctor, Teacher, Psychologist}	College	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
9	{Female, Male}	[15, 26[	Anxiety	{School Psychologist, Doctor, Teacher, Psychologist}	Gymnasium	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
10	{Female, Male}	[15, 26[	Problems with concentration	{School Psychologist, Doctor, Teacher, Psychologist}	Gymnasium	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
11	{Female, Male}	[15, 26[	Problems with parents	{School Psychologist, Doctor, Teacher, Psychologist}	Gymnasium	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
12	{Female, Male}	[15, 26[	Sexuality	{School Psychologist, Doctor, Teacher, Psychologist}	Gymnasium	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
13	{Female, Male}	[15, 26[	Anxiety	{School Psychologist, Doctor, Teacher, Psychologist}	Trade school	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
14	{Female, Male}	[15, 26[	Personal finance, housing, education, etc.	{School Psychologist, Doctor, Teacher, Psychologist}	Trade school	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
15	{Female, Male}	[15, 26[	Depression	{School Psychologist, Doctor, Teacher, Psychologist}	Technical school	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
16	{Female, Male}	[15, 26[	Mobbing	{School Psychologist, Doctor, Teacher, Psychologist}	Technical school	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
17	{Female, Male}	[15, 26[	Psychosis	{School Psychologist, Doctor, Teacher, Psychologist}	Technical school	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
18	{Female, Male}	[15, 26[	Problems with motivation	{School Psychologist, Doctor, Teacher, Psychologist}	Technical school	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
19	{Female, Male}	[15, 26[	Violation - sexual	{School Psychologist, Doctor, Teacher, Psychologist}	Technical school	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}
20	{Female, Male}	[15, 26[	Personality disorder	{School Psychologist, Doctor, Teacher, Psychologist}	Technical school	{Copenhagen, Roskilde, Aarhus, Aalborg, Odense}

Table 7: Example of t-closeness, where t=0.85.

Pros and cons of t-closeness

Advantages:

- The method ensures the anonymity of an individual against skewness and similarity attacks.
- T-closeness is useful in cases where we want to keep the distribution of the anonymised data as close as possible to the distribution of the original data.
- T-closeness does not alter the original values of the sensitive attributes.

Drawbacks:

- The method has a computationally complex procedure for computing the parameter t.
- Similarly to l-diversity, t-closeness focuses only on the sensitive attributes, hence it should be applied together with other anonymisation methods such as k-anonymity.
- T-closeness is less intuitive and harder to understand than k-anonymity and l-diversity.

More details on t-closeness can be found in [9, 10].

4.5 DIFFERENTIAL PRIVACY

All the above-mentioned anonymisation methods are to some extent vulnerable to an attack that exploits additional information possessed by the attacker. For example, an anonymised dataset can be targeted by a compositional attack, meaning that an attacker uses other sources such as the Web, public records or domain knowledge to obtain background information for de-anonymisation. Even though the mentioned methods consider that an attacker might have some background information, it is still difficult (or not always feasible) to estimate how much an attacker knows in advance. In most cases, this knowledge is more than we assume.

A stronger privacy model that provides a provable privacy guarantee against the attacks mentioned above has been proposed. Differential privacy is a method that helps reveal useful information about a dataset without revealing any private information about an individual record. The method guarantees that even if an attacker knows all records in a dataset, he/she will not be able to identify the specific record based on the output of a differentially-private method. In other words, the outcome of the analysis is not (significantly) dependent on an individual's record being in the dataset. Hence, the privacy risk is essentially the same with or without the individual's participation in the dataset. Note that differential privacy is not an absolute guarantee of privacy but allows to quantify the risk of privacy loss.

Formally, differential privacy is defined as follows [11]. A randomised function *K* is ε-differentially-private if for all pairs of datasets D_1 and D_2 that differ on at most one element, and for every set *S* of outcomes,

$$\Pr\,[K(D_1) \in S] \leq \exp(\varepsilon) \times \;\Pr\,[K(D_2) \in S]$$

The privacy budget ε measures the privacy loss of an individual from a single query. For multiple *t* queries that use independent randomisation mechanisms with privacy budget ε_i, the total privacy budget will be $\sum_{i=1}^{t} \varepsilon_i$. For this reason, the number of differentially-private analyses on a specific dataset is limited. The value of ε in practice is usually small, such as 0.01, 0.1 or ln2. In general, the smaller ε, the stronger the privacy guarantee.

Differential privacy is achieved by adding random noise to the result, which can be done through various differentially-private mechanisms, such as the Laplace mechanism, the exponential mechanism and the randomised response mechanism. Note that more noise will increase the privacy while reducing the accuracy of data. Hence, one should consider carefully the trade-off between privacy and utility when applying differential privacy techniques.

Consider the dataset from Table 2. Assume an analyst wants to know the ratio between male and female participants in the study. The actual answer is 8 females and 12 males. In order to protect the privacy of the participants, instead of the actual answer, the differentially-private answer of 11 females and 15 males is released by means of adding random noise. The answer retains the approximate ratio of the actual answer.

As another example, consider the healthcare dataset example described in Section 1.2. Assume that an analyst from the medical institution mentions in her/his article that in the medical study 48 female participants have anxiety. The following year another analyst publishes that in the medical study 47 female participants have anxiety. An attacker observing this and having the additional information that Trine who participated in this study left the school, can conclude that Trine has anxiety. Now, assume that instead of the actual answer the analysts use a differentially-private outcome and write that in the study approximately 40 female participants have anxiety. The publication of differentially-private outcomes reduces the risk of identifying Trine.

Below we briefly mention some differentially-private mechanisms. For more details, we refer to [11] and [12].

The Laplace mechanism is one of the commonly used differential-privacy mechanisms for numeric queries. The mechanism adds Laplace-distributed noise of magnitude depending on the privacy budget ε and the sensitivity of the query, that is, the maximum difference in the output that the query may take on a pair of databases that differ for only one record.

Randomised response is a technique that provides plausible deniability for individuals responding to sensitive surveys. It is widely used in statistical analysis for obtaining statistical information about a population without obtaining any information about the individuals in the population.

The approach randomises the data in the following way: for answering a binary question concerning private information a user flips a coin. If the outcome is “head”, the user answers “yes”. Otherwise, if the outcome is “tail”, the user answers truthfully (yes or no). This approach allows to compute an approximate answer for the true response rate without getting a true answer from all the users.

Interactive and non-interactive settings

Differential privacy can be applied in two different settings: interactive and non-interactive. In the interactive setting, a user gets access to the dataset through queries, i.e. the user queries the dataset and gets the differentially-private outcome of each query. On the contrary, in the non-interactive setting, a user gets access directly to the modified, differentially-private dataset.

Local and global models

There are two ways of collecting user data and implementing differential privacy. In the global model, actual user data is collected in a trusted database and then differentially-private analyses are performed on the dataset. In the local model, instead of actual data, a differentially-private version of the user data is collected which may then be used in analyses. Because of the privacy guarantee of differential privacy, any output of an analysis performed on the differentially private data is itself differentially private.

Pros and cons of differential privacy

Advantages:

- The model ensures the privacy of an individual in a range of attacks, e.g., re-identification.
- Differential privacy allows to learn useful information about a population guaranteeing that the information leaked about an individual in the population is limited.
- The privacy guarantee does not depend on the prior knowledge of an attacker.

Drawbacks:

- The amount of noise to add is not trivial to determine and depends on different aspects, such as a dataset.
- The number of differentially-private analyses possible to perform on the same dataset is bounded and depends on the privacy budget.

Differential privacy is exploited by organisations such as Google [[13], [14]] and Apple [15].

More details on differential privacy can be found in [[11], [12]]. Moreover, [16] introduces the concept of differential privacy from a non-technical perspective.

5 ANONYMISATION TOOLS

A dataset can be anonymised manually by following one of the methods described above. However, the manual anonymisation of large datasets is error-prone and time-consuming. The picture becomes even more challenging if the dataset requires a combination of few anonymisation methods for obtaining better results.

Various software tools, both open-source and commercial, are available in the market and can be used for automated anonymisation of datasets. They provide a cost-effective and repeatable in-house solution.

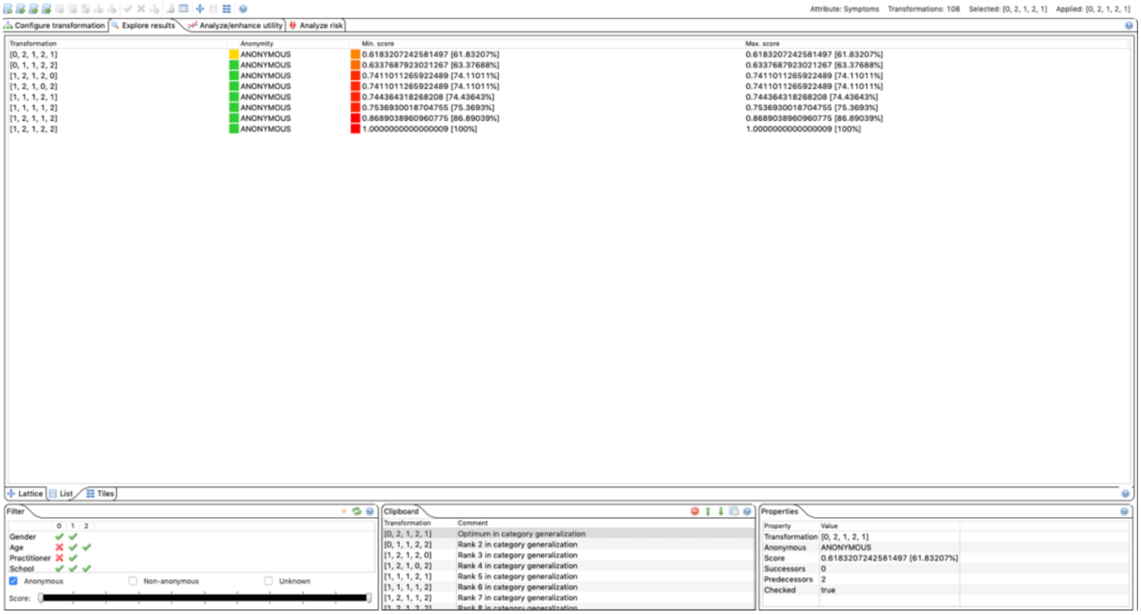
In this section, we focus on and describe in detail the state-of-the-art open-source tool ARX and the commercial tool Aircloak. These tools have user-friendly interfaces, good documentation and are still supported and maintained. In Appendix A we mention briefly some other available open-source and commercial anonymisation tools.

5.1 ARX – DATA ANONYMISATION TOOL

ARX is an open-source tool for structured personal data anonymisation. The tool supports different anonymisation methods, such as k-anonymity, l-diversity, etc. These methods can be applied both separately and in arbitrary combinations. Furthermore, the tool determines privacy risk and performs utility analysis.

After creating a project and uploading a dataset from CSV files, MS Excel spreadsheets and relational database systems, such as MS SQL, DB2, MySQL or PostgreSQL, a user identifies the type and domain of each attribute. For example, in Table 2 the attribute “age” is a quasi-identifier with a numerical domain. Then, the user selects and parametrises the methods that he/she wants to apply for data anonymisation, such as selecting k-anonymity, where k=2. Finally, the tool performs the analysis and suggests an anonymised dataset that can be further inspected and modified.

The ARX tool features a user-friendly graphical user interface, which makes it accessible to non-experts. Moreover, the tool supports different graphical visualisations, e.g., of the solution space (Figure 4) or risk (Figure 5), which allows to explore the results in an intuitive manner. Furthermore, ARX is also available as a Java software library and can therefore be integrated in existing software.


Attribute: Symptoms Transformations: 108 Selected: [0, 2, 1, 2, 1] Applied: [0, 2, 1, 2, 1]
Configure transformation
Explore results
Analyze/enhance utility
Analyze risk
Lattice
List
Tiles

Transformation	Anonymity	Min. score	Max. score
[0, 2, 1, 2, 1]	🟨 ANONYMOUS	🟧 0.6183207242581497 [61.83207%]	0.6183207242581497 [61.83207%]
[0, 1, 1, 2, 2]	🟩 ANONYMOUS	🟧 0.6337687923021267 [63.37688%]	0.6337687923021267 [63.37688%]
[1, 2, 1, 2, 0]	🟩 ANONYMOUS	🟥 0.7411011265922489 [74.11011%]	0.7411011265922489 [74.11011%]
[1, 2, 1, 0, 2]	🟩 ANONYMOUS	🟥 0.7411011265922489 [74.11011%]	0.7411011265922489 [74.11011%]
[1, 1, 1, 2, 1]	🟩 ANONYMOUS	🟥 0.744364318268208 [74.43643%]	0.744364318268208 [74.43643%]
[1, 1, 1, 1, 2]	🟩 ANONYMOUS	🟥 0.7536930018704755 [75.3693%]	0.7536930018704755 [75.3693%]
[1, 2, 1, 1, 2]	🟩 ANONYMOUS	🟥 0.8689038960960775 [86.89039%]	0.8689038960960775 [86.89039%]
[1, 2, 1, 2, 2]	🟩 ANONYMOUS	🟥 1.0000000000000009 [100%]	1.0000000000000009 [100%]

Filter

	0	1	2
Gender	✓	✓	
Age	✗	✓	✓
Practitioner	✗	✓	
School	✓	✓	✓

☑ Anonymous ☐ Non-anonymous ☐ Unknown

Score:

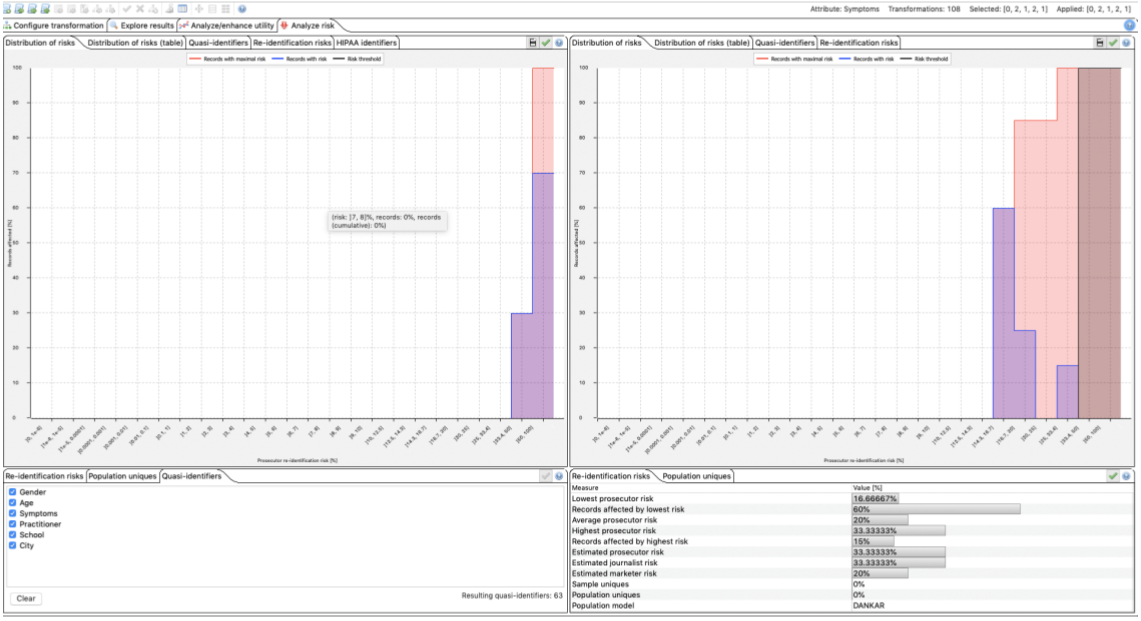
Clipboard

Transformation	Comment
[0, 2, 1, 2, 1]	Optimum in category generalization
[0, 1, 1, 2, 2]	Rank 2 in category generalization
[1, 2, 1, 2, 0]	Rank 3 in category generalization
[1, 2, 1, 0, 2]	Rank 4 in category generalization
[1, 1, 1, 2, 1]	Rank 5 in category generalization
[1, 1, 1, 1, 2]	Rank 6 in category generalization
[1, 2, 1, 1, 2]	Rank 7 in category generalization
[1, 2, 1, 2, 2]	Rank 8 in category generalization

Properties

Property	Value
Transformation	[0, 2, 1, 2, 1]
Anonymous	ANONYMOUS
Score	0.6183207242581497 [61.83207%]
Successors	0
Predecessors	2
Checked	true

Figure 4: Example of the solution space for the healthcare dataset.



☑ Gender
☑ Age
☑ Symptoms
☑ Practitioner
☑ School
☑ City

Clear

Resulting quasi-identifiers: 63

Measure	Value [%]
Lowest prosecutor risk	16.66667%
Records affected by lowest risk	60%
Average prosecutor risk	20%
Highest prosecutor risk	33.33333%
Records affected by highest risk	15%
Estimated prosecutor risk	33.33333%
Estimated journalist risk	33.33333%
Estimated marketer risk	20%
Sample uniques	0%
Population uniques	0%
Population model	DANKAR

Figure 5: Example of the risk for the healthcare dataset.

The tool is under constant development and builds on published research in the field of anonymisation.

For more details about the tool, related publications and download we refer to the official website of ARX: https://arx.deidentifier.org/.

5.2 AIRCLOAK INSIGHTS – ANONYMISATION SOLUTION FOR INSTANT PRIVACY COMPLIANCE

Aircloak Insights is a commercial data anonymisation tool. The tool anonymises the results of the queries to the database instead of anonymising the database itself, like many other tools do.

Aircloak Insights consists of Insights Air and Insights Cloak components, Figure 6. The first component is a web-based control centre that provides an interface for communication between an analyst and the database. The second component is the anonymisation module that analyses and anonymises sensitive data. Insights Cloak works in isolation, ensuring that sensitive data does not leave the secure perimeter.

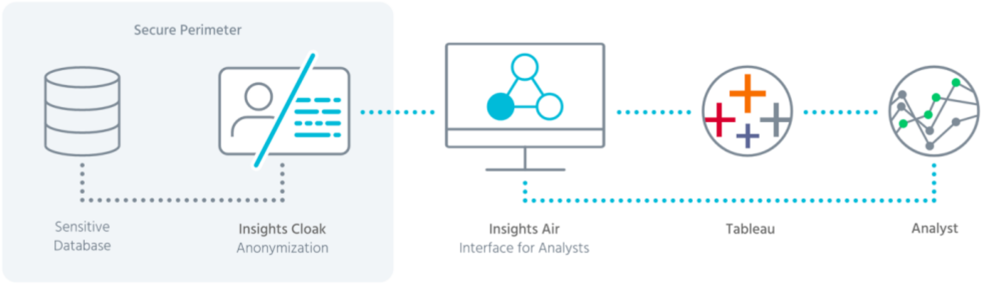

Secure Perimeter
Sensitive Database
Insights Cloak Anonymization
Insights Air Interface for Analysts
Tableau
Analyst

Figure 6: Aircloak Insights deployment. Illustration from aircloak.com

The tool is based on the Diffix framework [17], which adds so-called sticky noise to each condition of a query, i.e., multiple layers of noise is added to each query. For example, the following query [gender='M' AND age=20] would have two noises, one for each condition. In the framework, anonymisation depends on the requested query and the database. Hence, repeated or semantically equivalent queries get the same noise values. This fact allows to ask as many queries as desired, unlike other techniques that have limitations on the number of queries, such as a query budget in differential privacy. Moreover, Diffix reveals a value to a query only if a threshold number of distinct individuals have that value.

The tool works as follows. A user queries the database through a query (e.g., an SQL query). The tool modifies the query to the data backend (e.g., to a structured SQL database), gets the actual result and returns the anonymised requested data to the user by adding multiple layers of noise. Aircloak Insights works both with structured as well as with unstructured data.

Furthermore, it is compatible with many common databases and applicable for any use-case, e.g., healthcare and banking.

Similar to the ARX tool, Aircloak Insights is under constant development and builds on published research in the field of anonymisation.

For more details about the tool, related documentation and download we refer to the official website of Aircloak: https://aircloak.com/.

6 CONCLUSION

Extensive data collection has become essential for companies to offer improved and personalised services to their customers. With the great advantages of data collection come also the responsibility of protecting the information which is collected and stored. Failure to protect personal data can lead to critical incidents, including reputational damage and fines from regulators.

Anonymisation is the process of ensuring that the data of an individual in a dataset is not identifiable and is a key approach to tackling the challenges of protecting personal data. Anonymisation should be performed with due care, as a poorly-anonymised dataset is still subject to data leakage. Before performing any anonymisation, we must consider not only the data itself but also the context of the data, i.e., the use cases, the legal responsibilities, the possible uses of the anonymised dataset after sharing, etc. All these factors can have a role in the choice of the right anonymisation technique.

This paper is intended as an introductory guide to data anonymisation. We presented a wide-range of anonymisation techniques and tools, and for each technique we highlighted its advantages and disadvantages, so as to assist in choosing the suitable techniques for a given use case.

Finally, observe that we have not discussed risk analysis frameworks. However, it is a good practice to assess the privacy loss of an anonymised dataset before disclosing it. ENISA has collected a non-exhaustive list of different risk assessment methods [18].

7 BIBLIOGRAPHY

[1] Gartner, “Smarter With Gartner,” 9 April 2019. [Online]. Available: https://www.gartner.com/smarterwithgartner/how-to-balance-personalization-with-data-privacy/. [Accessed 10 March 2020].

[2] Z. Whittaker, “TechCrunch,” 4 September 2019. [Online]. Available: https://techcrunch.com/2019/09/04/facebook-phone-numbers-exposed/. [Accessed 11 December 2019].

[3] M. J. Schwartz, “Data Breach Today,” 10 December 2019. [Online]. Available: https://www.databreachtoday.com/gdpr-violation-german-privacy-regulator-fines-11-telecom-a-13482. [Accessed 11 December 2019].

[4] L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 10, no. 5, pp. 557-570, 2002.

[5] .. Danmarks Statistik, “Datafortrolighedspolitik,” 18 December 2018. [Online]. [Accessed 4 May 2020].

[6] P. Samarati and L. Sweeney, “Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement though Generalization and Suppression,” Technical report, SRI International, 1998.

[7] L. Sweeney, “k-anonymity: A model for protecting privacy.,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* pp. 557-570, 2002.

[8] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity.,” *Data Engineering,* 2006.

[9] N. Li, T. Li and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity.,” *Data Engineering,* vol. ICDE 2007, pp. 106-115, 2007.

[10] N. Li, T. Li and S. Venkatasubramanian, “Closeness: A new privacy measure for data publishing.,” *IEEE Transactions on Knowledge and Data Engineering ,* vol. 22, no. 7, pp. 943-956, 2010.

[11] C. Dwork, “Differential Privacy,” in *ICALP’06: Proceedings of the 33rd international conference on Automata, Languages and Programming*, Berlin, Heidelberg, 2006.

[12] C. Dwork, “A Firm Foundation for Private Data Analysis,” *Commun. ACM,* vol. 54, pp. 86-95, January 2011.

[13] Ú. Erlingsson, V. Pihur and A. Korolova, “Rappor: Randomized aggregatable privacy-preserving ordinal response.,” in *ACM SIGSAC conference on computer and communications security*, 2014.

[14] R. J. Wilson, C. Y. Zhang, W. Lam, D. Desfontaines, D. Simmons-Marengo and B. Gipson, “Differentially Private SQL with Bounded User Contribution.,” *arXiv preprint arXiv:1909.01917,* 2019.

[15] Differential Privacy Team, Apple, “Learning with Privacy at Scale,” December 2017. [Online]. Available: https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html. [Accessed 23 September 2019].

[16] A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, J. Honaker, K. Nissim, D. R. O’Brien, T. Steinke and S. Vadhan, “Differential Privacy: A Primer for a Non-Technical Audience,” *Vanderbilt Journal of Entertainment & Technology Law,* vol. 21, no. 1, 2018.

[17] P. Francis, S. Probst Eide and M. Reinhard, “Diffix: High-utility database anonymization.,” in *Annual Privacy Forum*, 2017.

[18] European Union Agency for Cybersecurity, “Inventory of Risk Management / Risk Assessment Methods,” [Online]. Available: https://www.enisa.europa.eu/topics/threat-risk-management/risk-management/current-risk/risk-management-inventory/rm-ra-methods. [Accessed 12 March 2020].

[19] “Privacy Analytics Eclipse,” [Online]. Available: https://privacy-analytics.com/files/Privacy-Analytics-Eclipse.pdf. [Accessed 23 07 2018].

[20] Danmarks Statistik, “Datafortrolighedspolitik,” 18 December 2018. [Online]. Available: https://www.dst.dk/ext/formid/datafortrolighed--pdf. [Accessed 4 May 2020].

8 APPENDIX A: ANONYMISATION TOOLS

In this Appendix, we will briefly present a few anonymisation software tools available in the market.

8.1 PRIVACY ANALYTICS ANONYMISATION TOOLS

Privacy Analytics provides commercial software solutions for anonymising personal health data for secondary use.

Privacy Analytics Eclipse anonymises structured data using a risk-based approach. The software evaluates the elements of the dataset, determines the privacy risks, and applies the right level of anonymisation for the context of the data release [19].

Privacy Analytics also provides a solution called Lexicon that works with unstructured data. It enables reduction and anonymisation of unstructured health data for further safe use. Similarly, Lexicon is centred around a risk-based approach.

More details can be found at https://privacy-analytics.com/software/.

8.2 FLEX (OR CHORUS FRAMEWORK)

FLEX is a re-writing engine for SQL queries that enforces differential privacy. It can be easily integrated with existing database environments since it does not require any modifications on the database and the database engine. FLEX works as follows. Whenever an analyst makes an SQL query, FLEX engine transforms it into an intrinsically-private query, where the latter guarantees differential privacy on the output received by the analyst. It supports many of the state-of-the-art differential privacy mechanisms, such as elastic sensitivity, and sample and aggregate framework.

FLEX is currently used by Uber (see https://medium.com/uber-security-privacy/uber-open-source-differential-privacy-57f31e85c57a)

More details can be found at http://www.uvm.edu/~jnear/elastic/ and can be downloaded at https://github.com/uber/sql-differential-privacy.

8.3 PINQ AND WPINQ

Privacy Integrated Queries (PINQ) is a platform for performing differentially-private data analysis done by Microsoft Research. PINQ allows the analysts to get access to the data through Language Integrated Queries (LINQ)-like API. It can be seen as a layer in front of an existing query engine that provides differentially-private implementations of common transformations and aggregations, i.e., it first applies transformations to a dataset and then performs a differentially-private aggregation.

More details can be found at https://www.microsoft.com/en-us/research/project/privacy-integrated-queries-pinq/.

Weighted Privacy Integrated Queries (wPINQ) is a platform for performing differentially-private data analysis of weighted datasets. It is an extension of the PINQ platform and has a similar structure.

More details can be found at http://cs-people.bu.edu/dproserp/wPINQ.html, https://arxiv.org/pdf/1203.3453.pdf and https://tpdp.cse.buffalo.edu/2015/abstracts/TPDP_2015_6.pdf.

8.4 AMNESIA

Amnesia is a data anonymisation tool that anonymises datasets by removing or modifying sensitive information. It is a web-based application that implements data anonymisation algorithms based on k-anonymity and k_m-anonymity. Amnesia reads the original data and then transforms it by using generalisation and suppression. It supports two algorithms of k-anonymity, Incognito, and a parallel version of the Flash algorithm. After the anonymisation has been done, the user is able to tailor the output database in case he/she needs to make it more usable.

Amnesia can be run locally or as a service, and it can be downloaded at https://amnesia.openaire.eu/installation.html.

More information can be found at https://amnesia.openaire.eu/index.html.

9 APPENDIX B: QUESTIONS

In this Appendix, we present some questions that can help understand the data and its context. The presented questions apply to the running example.

What kind/type of data does the organisation record/register? For example, children data, health/medical data, financial data, etc.

The institution has health data of young population (between 18 and 29) containing sensitive information, such as symptoms reported by the patients.

Does the dataset contain *sensitive data*?

Yes, the dataset contains sensitive health data.

FOR WHAT PURPOSES IS THE DATA GOING TO BE USED/ ANONYMISED?

Why is the data being released?

The data is being released for statistical purposes in order to compare the performance of the institution with respect to other medical institutions.

Why is the data being anonymised?

The data is being anonymised in order to protect the sensitive data of any individual in the dataset as the data will be publicly available.

With whom will the data be shared? (Is it going to be shared with a trusted person who signed Data Use Agreement or is it going to be publicly available on a website?)

First of all, the data will be shared with the analysts. Later, the anonymised data will be publicly available for everyone.

What are the people who have access to data allowed to do with it?

The analysts will use the data for statistical analysis, while the publicly available anonymised data will have read-only permission.

How will the anonymisation help?

If an attacker manages to identify sensitive information about a patient, this will affect the reputation of the institution. Moreover, it might damage them financially. Finally, it might lead to unlawful discrimination of the individual. Hence, the anonymisation process can help to protect sensitive data.

DIRECT IDENTIFIERS

What attributes in the data allow to uniquely identify an individual (if any)?

There are no direct identifier attributes in the dataset.

QUASI-IDENTIFIERS

What attributes in the data might allow an attacker to uniquely identify an individual by combining them with additional information (even if these attributes do not allow to uniquely identify any individual)?

The quasi-identifiers are: gender, age, practitioner, school, city.

SENSITIVE ATTRIBUTES

What are the sensitive attributes?

The only sensitive attribute is the attribute "symptoms". One can argue that the attribute "comments" is also sensitive, however, as it appears only for a few records, the best option seems to exclude that attribute from the dataset.

WHAT IS A POTENTIAL ATTACKER'S PROFILE?

Who is a potential attacker?

A potential attacker can be someone who is working for a competitor. It can be a hacker or a criminal.

What additional information might an attacker have? Are you aware of any publicly available information that can help de-anonymise an individual?

An attacker might have access to similar datasets from other medical institutions. Moreover, an attacker might know additional information about the institution or an individual.

Why would an attacker try to de-anonymise the data or an individual?

An attacker might want to damage the reputation of the institution or might want to obtain some sensitive data about an individual in the dataset. For example, health-related information might be sold to insurance companies.

HOW MUCH CAN THE ANONYMISATION PROCESS ALTER THE ORIGINAL INPUT DATA?

Can any attribute be lost?

As the main purpose of the data is to perform statistical analyses, some attributes can be lost, however the remaining data should still provide necessary information for statistical analyses.

What attributes can be manipulated so as to alter partially their values?

All attributes can be made "imprecise".

Which ones must not?

None.

Is there going to be a subsequent release of the data?

There is a possibility of a subsequent release of an updated version of the dataset.

Is there any Data Use Agreement (DUA) Model, i.e., a document that describes what data is being shared and how the data can be used?

The institution has a DUA in place with the data analysts. However, there is no DUA for the publicly available anonymised data.

THE ALEXANDRA INSTITUTE
IT CITY OF KATRINEBJERG
Aabogade 34 ▪ DK-8200 Aarhus N
+45 7027 7012
UNIVATE
Njalsgade 76, 3rd floor ▪ DK-2300 Copenhagen S
+45 7027 7091

ALEXANDRA
INSTITUTE

The Alexandra Institute helps public and private organisations apply the latest IT research and technology to create innovative solutions. Our mission is to contribute to growth and prosperity in Denmark.