university of groningen

Data Anonymisation in the light of the General Data Protection Regulation

Faculty of Science and Engineering

A.F. Bijl (s2581582)

Supervisors:
dr. F.B. Brokken
prof. dr. G.R. Renardel de Lavalette

July 2017

Abstract

The General Data protection Regulation (GDPR) was approved by the European Parliament and Council and will be applied from May 2018 onwards in all countries of the European Union. The GDPR protects the privacy of European citizens by limiting the storage and movement of their personal data. The University of Groningen started a project to analyse the GDPR from a technical perspective. The following thesis is part of this project and takes as subject anonymisation.

Anonymisation is a technique to transform information relating to individuals (personal data) to information from which no individual can be identified (anonymous data). As a result organisations are able to publish data to third parties without the sanctions of the GDPR because the GDPR falls outside the scope of anonymous information. Former studies on re-identification have shown that zero risk of re-identification is impossible for useful data. Therefore a trade-off between data privacy and utility has to be made. With this aspect in mind the following thesis analyses and evaluates several anonymisation techniques.

The models that are evaluated are k-anonymity, l-diversity, t-closeness and differential privacy. It is observed that no model can guarantee zero risk of re-identification because every model is vulnerable to particular re-identification attacks. Every model can be used to provide more privacy but drops the data utility. The GDPR contains no specifications to satisfy the privacy of data subjects. Therefore organisation must show effort in the direction of privacy. Still it has to seen when the GDPR is applied what effort will be enough.

Contents

Chapter 1

Introduction

In this digital age it is more and more common for Internet users to create an account on a Internet web page. Internet users nowadays have multiple accounts to write documents, order clothes, visit a doctor, watch movies, etc. Websites and social platforms enable signing up in a convenient manner. In order to have full access on web pages or social platforms, a personal account is often required. Statistics from a mailbox scanner [1] show that on average, their users have 107 Internet accounts. These statistics show a growing trend that suggests an amount of 207 Internet account per Internet user in 2020. As a result there is a growing public interest in and demand for the re-use of these data. Companies and organisations nowadays are 'data controllers', that collect and store user accounts and their activities in large databases. As a result personal information such as zip code, address, email address etc. of individuals is stored by these controllers in databases containing data of thousands or even millions of users. Internet users can be seen as 'data subjects' from which personal data relates to.

1.1 Project background

As an answer to the increasing storage and movement of personal data the European Parliament and Council formulated a regulation called the General Data Protection Regulation (GDPR) [2]. In 2018 the EU will replace the Data Protection Directive 95/46/EC established in 1995 by the GDPR. Where a directive can be seen as a goal specified by a legislative act, a regulation can be seen as a binding legislative act. Non-compliance with the GDPR can result inchap: fines up to 4% of annual global turnover or €20 Million. The GDPR will not merely affect organisations within Europe but every organisation, possibly located outside the Europen Union, that offers goods or services to, or monitor the behaviour of EU data subjects. The GDPR was approved by the European Parliament and the Council in May 2016. After a transition period of two years the GDPR will be active starting 25 May 2018.

Research data falls under the scope of the GDPR, therefore the University of Groningen started a project to analyse the GDPR. This thesis is part of this project conducted from a computing science point of view. To comply with the GDPR, this project analyses the Standard Data Protection Model [3], a model that built a bridge between the GDPR and technical measures to satisfy its requirements. In this project three technical implementations of the GDPR were found: encryption, anonymisation and pseudonymisation. Encryption is the process through which data is transformed to unintelligible text which looks like random characters. Anonymisation is the process transforming potentially identifiable information into anonymous information from which individuals can not re-identified. Pseudonymisation is the reversible technique of anonymisation; information is replaced but can be retrieved. These three techniques are taken into further research by each individual author. This thesis takes as subject the anonymisation of data.

1.2 The General Data Protection Regulation (GDPR)

The GDPR [2] was written to protect European citizens with regard to the processing and movement of their personal data. The GDPR defined in definition 4 of article 4 what kind of information is seen as personal data.

"Any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."

The GDPR even specifies 'identifiers': fields that can identify individuals. Such identifiers relate to actual persons and allow the identification of actual persons merely using stored information.

1.2.1 GDPR and Anonymisation

The GDPR was built around the protection of personal data of EU citizens. The term personal data is clearly specified within the GDPR, with this definition the question is raised what qualifies as anonymous data. The GDPR provides in recital 26 the definition and consequences of anonymous information.

"The principles of data protection should apply to any information concerning an identified or identifiable natural person... The principles of data protection should therefore 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."

According to the GDPR anonymous information is information from which no individual can be identified. Identifiers which can identify individuals can therefore not be present. Furthermore, according to this recital, the GDPR will not be applied to information considered anonymous. Francis Aldhouse [4] and Matt Wes [5] draw the same conclusion that when anonymisation is performed properly data is taken outside the scope of the GDPR. As this is the case and no individual can be identified with anonymous data, it allows data controllers to store and move this data.

The freedom that anonymous data provides is desirable in certain cases. For example, health service organisations are required to protect the identities of individual patients but may also be required to publish statistics about patient outcomes. Therefore anonymisation, a collection of techniques to transform personal data to anonymous data, is very important for data controllers. For health service organisations, anonymisation is important as it allows them to ensure the privacy of individual patients, and also to publish their statistics. Anonymisation helps to provide privacy to data subjects but also enables transparency within the organisation.

1.3 Anonymisation: Current state

With regard to the GDPR, anonymisation can be seen as the 'holy grail' that takes data outside the scope of the GDPR. Several requirements of anonymisation are required, those are formulated and named in section 1.3.1. However research into anonymisation and anonymous data have shown that identifying individuals from anonymous data is still possible. Certain studies re-identified individuals and set anonymisation in a different light as mentioned in sections 1.3.2 through 1.3.4. Drawing from these publications Ohm [10] observed a big failure of anonymisation, his study is discussed in section 1.3.5. As an answer to these studies the European commission published an article how anonymisation can be seen in line with the GDPR as mentioned in section 1.3.6.

1.3.1 Requirements for anonymisation

In the United States the protection of health information is secured by the Health Insurance Portability and Accountability Act (HIPAA) of 1996 [6]. The HIPAA is the American version of the Directive from 1995 in Europe and is seen as a significant privacy law. The HIPAA privacy rule is protecting most "individually identifiable health information" which is called Protected health information (PHI). The HIPAA does not specify what anonymisation is, instead it uses de-identification as a technique. Their definition of de-identification is:

"De-identification of protected health information. Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information."

In this way de-identification can be seen as a variant of anonymisation, moreover requirements of de-identification are applicable to anonymisation. In order to achieve de-identification the Privacy Rule provides two methods by which health information can both be designated and de-identified.

The first is the 'Expert Determination' method. In this approach an expert determines the change to be very small, or even impossible, to identify someone from the data. The expert must have appropriate knowledge of, and experience with, how to identify individuals from given information. In this approach the judgement must be documented by the expert.

The second is the 'Safe Harbor' method, certain fields from which individuals can be identified have to be removed. The privacy rule names eighteen different so called identifiers that must be removed from information to ensure de-identification. The identifiers that must be removed are listed in appendix 5.1.

In 2010 the Health System Use Technical Advisory Committee and the Data De- Identification Working Group of Canada published guidelines for de-identification [7]. A model was proposed such that the data have to be assessed on reidentification risks related to the data disclosure. Identifying data can be classified in directly or indirectly identifying variables. Directly identifying variables can be used to identify a person with only the variable itself or with other available information. Indirectly identifying variables, or quasi-identifiers, allow the identification of individuals with non-zero probability. These variables result in four levels of decreasing identifiability where the first two are not acceptable:

1. Identifiable data, the data contains directly identifiable variables or enough quasi-identifiers to identify an individual.
2. Potentially de-identified data, identifying variables are de-identified but there can be ways to combine the quasi-identifiers to identify a person.
3. De-identified data, an acceptable level of de-identification is achieved by de-identification of the directly identifiable variables and quasi-identifiers are disguised.
4. Aggregate data, no existence of identifying variables or quasi-variables.

1.3.2 Sweeney

In 2000 Sweeney released a famous paper [8] which identifies people based on two different data sets. Sweeney combined medical data with vote data that resulted in two lists as figure 1.1 shows. In her paper Sweeney experimented with these data sets and published the following remarkable result:

"It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on 5-digit ZIP, gender, date of birth. About half of the U.S. population (132 million of 248 million or 53%) are likely to be uniquely identified by only place, gender, date of birth, where place is basically the city, town, or municipality in which the person resides. And even at the county level, county, gender, date of birth are likely to uniquely identify 18% of the U.S. population."

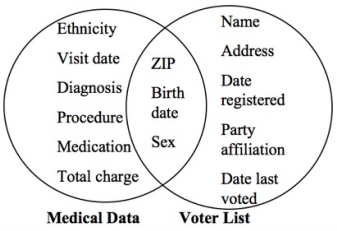


Figure 1.1: The combination of datasets that L. Sweeney enabled to reidentify individuals.

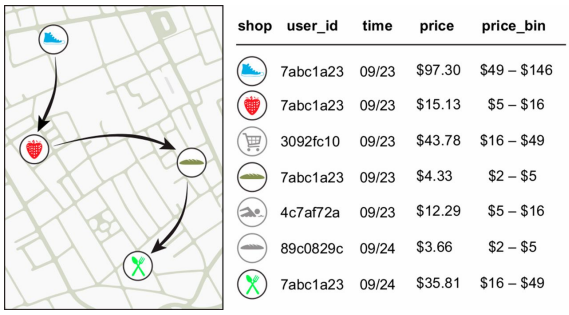
Sweeney found that a combination of even a few characteristics can identify people uniquely and should therefore not have been published or released. Although these characteristics were assumed anonymous, it was shown that these characteristics can be used to identify subjects.

1.3.3 Narayanan and Shmatikov

Narayanan and Shmatikov [9] applied a new de-anonymisation attack on real life data. The new de-anonymisation methodology was applied on a large Netflix prize dataset of 500.000 subscribers. This Netflix prize dataset was published by Netflix to improve recommendations they give to their customers. In order to ensure privacy, Netflix removed obvious identifiers such as names, addresses etc. Netflix therefore considered their database to be anonymous. Combining the 'anonymous' Netflix data with the public data of the Internet Movie Database (IMDB) enabled the authors to identify persons based on their ratings and dates. As a conclusion the authors raised certain questions whether there would be anonymisation techniques that can beat their de-anonymisation techniques.

1.3.4 Reidentifiability of credit card metadata

In his paper published in 2015 de Montjoye studied three months of credit card records of 1.1 million people. This dataset was said to be simply anonymised because obvious identifiers as names, addresses and phone numbers were removed from the data as you can see in figure 1.2.



shop	user_id	time	price	price_bin
	7abc1a23	09/23	$97.30	$49 – $146
	7abc1a23	09/23	$15.13	$5 – $16
	3092fc10	09/23	$43.78	$16 – $49
	7abc1a23	09/23	$4.33	$2 – $5
	4c7af72a	09/23	$12.29	$5 – $16
	89c0829c	09/24	$3.66	$2 – $5
	7abc1a23	09/24	$35.81	$16 – $49

Figure 1.2: An example of the dataset de Montjoye used to re-identify user 7abc1a23.

In his study de Montjoye showed that only four 'spatiotemporal' points are enough to uniquely re-identify 90% of individuals. De Montjoye added the following example to illustrate the reidentification of a person called Scott represented by user 7abc1a23 in figure 1.2:

"We know two points about Scott: he went to the bakery on 23 September and to the restaurant on 24 September. Searching through the data set reveals that there is one and only one person in the entire dataset who went to these two places on these two days. |S(Ip)| is thus equal to 1, Scott is reidentified, and we now know all of his other transactions, such as the fact that he went shopping for shoes and groceries on 23 September, and how much he spent"

1.3.5 Ohm

Ohm (2009) wrote an article [10] about the overestimation's of anonymisation. In addition to the studies of Sweeneys and Narayanan he formulated that data can be either useful or perfectly

anonymous, but never both. Ohm considers the masking of certain identifiers as a huge failure and concludes that less privacy can be assured than assumed by privacy law, regulation, and debate. Therefore anonymisation is overestimated by authorities, laws and organisations in order to avoid critical questions about privacy. As a conclusion Ohm suggests a different concept of anonymous information.

"Better yet, we need a new word for privacy-motivated data manipulation that connotes only effort, not success. I propose "scrub." Unlike "anonymize" or "deidentify," it conjures only effort. One can scrub a little, a lot, not enough, or too much, and when we hear the word, we are not predisposed toward any one choice from the list."

Scrubbing implies here that perfectly anonymisation of data can not be reached but one can only take efforts into that direction. To take steps in the direction of anonymisation Ohm names Five Factors for Assessing the Risk of Privacy Harm: Data-Handling Techniques, Private Versus Public Release, Quantity, Motive and Trust. Surprisingly Ohm still names data-handling techniques that affect the risk of easy reidentification. For this purpose it is still useful to have anonymisation techniques that could help improving the privacy of individuals.

1.3.6 Position of the EU Working party on anonymisation

The Article 29 working party (the working party), founded in 1996, is an independent group of people that provides the European Parliament and EU-states with expert advice on data protection. The working party provides guidelines on the regulations of the EU Parliament and Council. As an answer to the critical studies above, the European Parliament by means of the Article 29 working party published its opinion on anonymisation techniques [11]. In their opinion the working party validates the fact that proper anonymised data will fall outside the scope of the GDPR. As proper anonymisation will have value for individuals and societies as well as mitigating risks. The opinion acknowledges the problem that comes along with anonymisation and analysed certain anonymisation techniques. The working party conjectures that by understanding and proper building of these techniques anonymisation can be obtained.

"Knowing the main strengths and weaknesses of each technique helps to choose how to design an adequate anonymisation process in a given context... The Opinion concludes that anonymisation techniques can provide privacy guarantees and may be used to generate efficient anonymisation processes, but only if their application is engineered appropriately "

The opinion does not mean by privacy guarantees that identification is excluded. The Working Party tries to bring anonymisation to an acceptable level for the risk of re-identification. This approach overlaps with the definition of scrubbing suggested by Ohm in section 1.3.5. Scrubbing only takes effort and gives no guarantees, some level of scrubbing can be called sufficient. This is also what the working party suggests in their opinion on an acceptable level of anonymisation:

"The Working Party has therefore already clarified that the "means ... reasonably to be used" test is suggested by the Directive as a criterion to be applied in order to assess whether the anonymisation process is sufficiently robust, i.e. whether identification has become "reasonably" impossible."

To establish this acceptable level the opinion is clear about the approaches such as the 'Safe Harbor method' where identifiers were removed from the data:

"Generally speaking, therefore, removing directly identifying elements in itself is not enough to ensure that identification of the data subject is no longer possible. It will often be necessary to take additional measures to prevent identification, once again depending on the context and purposes of the processing for which the anonymised data are intended."

1.4 Solving the contradiction

According to the GDPR itself and the additional opinion of the Article 29 working party it can be concluded that proper anonymisation will take data outside the scope of the GDPR. That means

that the regulations do not affect anonymised data.

Earlier studies [8][9][10] to data anonymisation have shown that data considered anonymous could still identify users. A contradiction emerges because how can anonymous data, which by definition could not identify individuals, still be used to identify those individuals? Ohm showed that real anonymous data does not exist and that an acceptable level of reidentification must be established. The Working Party also acknowledged this fact, the robustness of anonymisation process must be evaluated.

1.5 Research question

The goal of anonymisation techniques is to bring the data to an acceptable level of reidentification. As stated by the working party, this acceptable level must be considered from case to case. Certain anonymisation techniques are known and used to de-identify data. In their opinion the working party has considered multiple anonymisation techniques and this thesis takes them as a foundation to work on. This thesis' research question, therefore, is:

which anonymisation techniques can ensure guarantees of privacy to data in line with the GDPR?

In order to answer this question, the different anonymisation models introduced in chapter 2 are evaluated. Chapter 2 introduces different privacy attacks from which these anonymisation models can be evaluated. The evaluation of these models is performed in chapter 3 and further discussions and conclusions are written down in chapter 4.

Chapter 2

Methods

The current chapter covers the methods that were used to evaluate the various anonymisation techniques as well as their properties. First the way how data are organised and the corresponding scope and definitions are discussed. Then the requirements of anonymisation are introduced and measures how to find this. The chapter concludes with different anonymisation models that are measured.

2.1 Scope and definitions

Typically, data are stored in tables where each row corresponds to one individual. Each row holds the data of individual subjects, the columns hold the data of the collected attributes (like date of birth, gender, etc.). According to Fung et al. (2010) [12], each row has a number of attributes that can be separated in the following groups:

- Explicit-identifier, set of attributes that directly identifies a record owner. Examples are a full name and social security number
- Quasi-identifier, attributes that could potentially identify record-owners. Example is the combination of zip code, date of birth, and sex to uniquely identify someone.
- Sensitive information, attributes that contain personal specific sensitive information. Examples could be disease and salary.
- Non-sensitive information, the attributes that are not covered by the previous three groups.

As we have seen in chapter 1, a table containing these four groups of attributes cannot be published. Currently, data tables can only be published when the data they contain is anonymized and data subjects cannot be identified. An anonymous table must satisfy a chosen privacy model and the data have to remain as useful as possible. Non-anonymized and anonymized tables differ as follows:

- Explicit-identifiers are removed from anonymised tables.
- Quasi-identifiers are anonymized.

As we have seen in the introduction, the removal of explicit-identifiers is not sufficient (1.3.2 and 1.3.3). Therefore, the important part is the anonymisation of quasi-identifiers. Different operations on these quasi-identifiers are used to apply anonymisation. The article 29 working party [11] specified two families of anonymisation: randomisation and generalisation/suppression. Randomisation changes the data to an extent that reidentification from these data is sufficiently uncertain. Generalisation is a combination of techniques that generalises data: the scale of the values is increased and set to different categories that reveal less about a data subject. As an example, the value 'The Netherlands' of attribute 'Country' can be generalised to 'Europe'. Suppression is a operations that removes part of quasi-identifiers. For example a postal area code can be generalised by suppression to remove the last digits or characters. Unlike randomisation, generalisation and suppression does not affect the truthfulness of data but alters the granularity of data. Less specific information is published after generalisation and suppression.

2.2 Measures

Ohm (2009) stated that data can be either useful or perfectly anonymous but never both. Maximum privacy is easy to establish by not publishing a table at all, obviously that gives no utility. In the same manner maximum utility does not satisfy the privacy of individuals. A tension exists between data privacy and utility and a trade-off between these two dimensions must be made. The tension between these two dimensions is shown in figure 2.1.

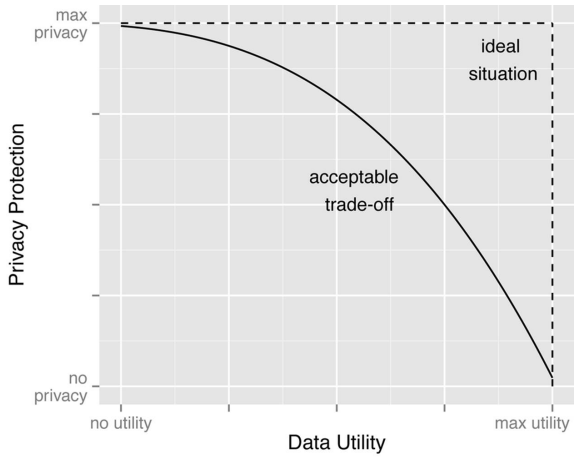


Figure 2.1: The trade-off between data utility and data privacy.

An acceptable level of both privacy and utility must be established to publish data. To put anonymisation models in the light of the GDPR this research analyses anonymisation models on the basis of privacy. A high data utility is desirable but no requirement of the GDPR.

Chapter 1 showed that zero privacy cannot be accomplished, however attacks can be excluded by certain privacy models. In this research privacy is measured by determining/measuring the vulnerability of data to specific types of attacks. An attack is an opportunity for an attacker to gain knowledge about a sensitive value of an individual. The following attacks are defined:

- Record linkage, the attacker knows in advance a value of a quasi-identifier. The amount of data subjects that have this value can be extremely small or even reduced to a single person. In that case the attacker is absolutely certain to identify an individual from the table.
- Attribute linkage or homogeneity attack, the attacker knows some quasi-identifiers in advance but those cannot identify the precise record of the data subject. Still the data set contains a group with these quasi-identifiers that have the same sensitive value. As a result the attacker is still able to know a sensitive value of a data subject.
- Table linkage, an attack where different tables that are related to each other reveal information about data subjects. The name and quasi-identifiers can exist in one table and combining these quasi-identifiers could reveal the record of a person in the other table.

2.3 Anonymisation models

Different privacy models to implement anonymisation by generalisation/suppression exist. In this research three of these models will be measured on privacy and utility.

2.3.1 k-Anonymity

K-anonymity is a protection model formulated by Sweeney (2002) [13]. This model is based on the k-anonymity requirement that a single or group of quasi-identifiers satisfy the k-anonymity if and only if each sequence of values in the data table appears with at least k occurrences in that data table. As a result a person in the table can not be distinguished from k-1 individuals who also appear in the table.

2.3.2 l-Diversity

L-diversity is an addition to k-anonymity by Machanavajjhala et al. (2007) [14]. To satisfy l-diversity a group of quasi-identifiers contains at least l "well represented" values for certain sensitive fields. The authors define "well represented" in three different ways. The easiest model to understand is; within each equivalence class there must at least be l different sensitive fields.

2.3.3 t-Closeness

T-closeness is a further improvement to k-anonymity and l-diversity proposed by Li et al. (2007) [15]. t-Closeness takes care of the distribution of each group of quasi-identifiers (equivalence class). A group of quasi-identifiers meets t-closeness when each published distribution mirrors the initial distribution of each attribute till a threshold t.

2.3.4 Differential Privacy

Differential privacy is a different model to publish data. It has a query/response technique where each response must satisfy ϵ-differential privacy. Moreover each response table must differ in precisely one record.

Chapter 3

Evaluation

In this chapter the anonymisation models are evaluated with respect to data privacy.

3.1 k-Anonymity

The model of k-Anonymity is defined by the following rule: *Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals.* Therefore rows with a unique combination of quasi-identifiers must be generalised or suppressed to equivalent classes with size k.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 3.1: The 4-anonymity of inpatient micro data

The right hand table in figure 3.1 shows generalisation on age, and suppression on zip code and nationality of the left table with micro data. The resulting table satisfies the k-anonymity requirement with k=4: every equivalence class of quasi-identifiers has a minimum size of 4.

3.1.1 Privacy

k-Anonymity prevents record linkage because an attacker can not be 100% certain which row corresponds to an individual. Given a combination of quasi-identifier corresponding to an individual there is a probability of 1/k that a row corresponds to that individual.
k-Anonymity is still vulnerable to two attacks known as the homogeneity and background knowledge attack. The homogeneity attack can be shown by figure 3.1, here equivalence class, containing number 9 till 12, for every row contains the similar condition. Given the quasi-identifiers of an individual, it can be concluded that this person corresponds to this group. Therefore an attacker is able to conclude that this person has cancer. This conclusion is sensitive and cannot be concluded from a released data table. The background knowledge attack is based on additional knowledge of the table. Given quasi-identifiers from a person, an attacker can exclude certain sensitive information from an equivalence group. If an attacker reduces this to one sensitive value, the attacker has knowledge about that particular value. For example: consider that an attacker knows that some one is in the first equivalence group (1-4) and has no heart disease. So the attacker can conclude that this individual has a viral infection.

There are practical problems with k-anonymity as well. First the number of k can be chosen too small. Let k=2, given a combination of quasi-identifiers an attacker get to know a sensitive value with 50% confidence. Second quasi-identifiers can be missed in the data set. When quasi-identifiers are missed, modification such as generalisation or suppression will not be applied. As a result those quasi-identifiers are exposed when data is published. Third optimal k-anonymisation is a NP-hard problem as shown by [17]. Although there are some greedy-algorithms that approaches optimal k-anonymisation.

3.2 l-Diversity

Another problem of k-Anonymity is the diversity within equivalence classes. The property of l-diversity is an addition of k-anonymisation that tackles this problem. The definition is: *An equivalence class is l-diverse if it contains at least l "well-represented" values for each confidential attribute. A data set is l-diverse if every equivalence class in it is l-diverse.*

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Figure 3.2: 3-diverse of inpatient micro data

The table in figure 3.2 shows 3-diversity of the micro data of figure 3.1. The generalisation of age is increased but the zip code suppression is reduced to only the last digit.

3.2.1 Privacy

l-Diversity preserves variability within equivalence classes, therefore it is impossible to interfere with the sensitive value for equivalence classes. The probability that when applying l-diversity and given a combination of quasi-identifiers, a row corresponds to an individual is 1/l. Therefore l-diversity prevents attribute linkage.
Still l-diversity has some major drawbacks known as the skewness and similarity attack [15]. The skewness attack is best illustrated by a table containing individuals that are tested on HIV. Assume that 1% of this table is positive and 99% negative. Suppose that each equivalence class has an equal number of positives and negatives, therefore it meets the requirements of 2-diversity. Accordingly this means that each individual in an equivalence has a change of 50% being HIV positive. This is a serious privacy risk compared to the initial percentage of 1%. The similarity attack means that l-diversity leaves room to gain information from the table itself. An example here is to consider an equivalence class with 3 different stomach diseases. An attacker can still conclude that an individual who meets the quasi-identifiers of that class has some stomach related disease.
Sometimes l-diversity is difficult to achieve consider again the example of the HIV test. If this table contains 10000 rows there can only be 100 distinct equivalence classes. This will result in an enormous loss of information.

3.3 t-Closeness

The property of t-closeness deals with the distribution of the whole attribute and the distribution of this attribute within each equivalence class. The difference between these distributions can not be higher higher than t. The differences between the distribution is calculated by the Earth

mover's distance (EMD). The EMD can be calculated for both numerical and categorical values as shown by Li et al. [15].

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Figure 3.3: Table that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

The table in figure 3.3 shows 0.167-closeness for the salary attribute. This is caused by the second equivalence class containing the set $\{6, 8, 11\}$. It is calculated to compute the distance between the whole distribution and this equivalence class.

3.3.1 Privacy

The t-closeness property preserves distribution equality. A table satisfying t-closeness for a low number of t is not vulnerable to a skewness attack. An attacker can not see significant difference between the original table and the table satisfying t-closeness for a low value of t.
The problem here is that t-closeness is still vulnerable to attribute linkage or the homogeneity attack illustrated in section 3.1.1. Another undesirable effect of t-closeness is the decrease of the relation between quasi-identifiers and sensitive values. If one equivalence class contains a sensitive value that doesn't appear that often in the overall distribution it will be generalised or suppressed in order to satisfy t-closeness. Data utility is strongly decreased by this fact. Solving this problem can be done by increasing t, but this will result in higher vulnerability to the similarity attack.

3.4 Differential privacy

Differential privacy is a model that preserves data utility and privacy by randomisation of the data [18]. In contrast to the previous techniques, differential privacy provides the dataset to an authorised third parties as an answer to queries on the dataset. Instead of a privacy model for the table itself, differential privacy is a mechanism to release data. An algorithm that satisfies differential privacy has its input in the initial dataset of the data controller. As its output it produces tables that differ in one single record. The property of differential privacy is that the probabilities of these different tables will be almost similar. The difference between these tables is notated as ϵ, and the corresponding technique is called ϵ-differential privacy. In order to satisfy this principle random noise is added to the response to each query.

3.4.1 Privacy

Differential privacy will prevent attackers to identify persons within the dataset because the dataset itself is not published. Due to noise addition an attacker cannot find the differences of the data sets because noise will fluctuate more than the difference of one record. When an attacker cannot see the difference between two data-sets that differ in one person, the attacker can't learn the data of that single person.
There are some drawbacks of differential privacy. First a query based model offers much privacy for data subjects, but for data analysts this may not be desirable. If a table is released analysts can interact with it in their own manner but this will not be the case when dealing with differential privacy. An attacker may perform a counting attack by query the same table multiple times. At a certain point the attacker can draw conclusions with some certainty about the initial dataset which is highly undesirable. Therefore a differential privacy model must keep track and set a limit on the amount of particular queries of a third party. The last problem with differential privacy is establishing the value of ϵ and the corresponding noise. If ϵ is too low more utility is preserved but this threshold will preserve less privacy.

Chapter 4

Discussion

This thesis was conducted in the light of the General Data Protection Regulation (GDPR) established by the European council and parliament in 2016. The GDPR preserves the privacy of EU citizens by limiting the movement and storage of their personal data.
In this thesis it is shown that the General Data Protection Regulation (GDPR) places anonymous information; information from which no individual can be identified, outside the scope of the GDPR. As former studies have shown (chapter 1.3.1) anonymisation techniques are overestimated. Data with zero risk to reidentification can not be established. The EU working party acknowledged this fact but still stressed the potentials of anonymisation techniques, therefore the following research question emerged:

which anonymisation techniques can ensure guarantees of privacy to data in line with the GDPR?

Multiple anonymisation techniques are evaluated on individual privacy and data utility.
k-Anonymity, l-diversity, t-closeness and differential privacy are described and analysed to measure their privacy. Each anonymisation model can satisfy privacy requirements of the data subjects and can be taken as an endpoint of anonymisation. Still, as shown in chapter 3, each model has several potential drawbacks and vulnerabilities to different privacy attacks.
To set these anonymisation models in the scope of the GDPR is an inconvenient task. In order to be 'future proof', the GDPR does not name particular methods to establish anonymisation. The analysed methods are named by the working party in their opinion as shown in 1.3.6 and leave freedom to make choices for every individual data set. Therefore the opinion lacks specific requirements for anonymisation techniques. Absolute guarantees for values such as k, l, t and ϵ therefore can not be given at this moment. As Ohm [10] suggested data controllers have to show effort in the direction towards data anonymisation. When an organisation publishes data it must be shown that anonymisation techniques are applied to a particular level. When this is considered a proper anonymisation, and therefore falling outside the scope of the GDPR, is a matter of discussion.
Related work will consist of finding better and new techniques to perform data anonymisation. The potentials and theories of differential privacy were found in a late stadium by the present author. Considering the scope of the thesis, full coverage of the relevant theory and progression over recent years has to be postponed to future research projects. Nowadays differential privacy is seen as the potential Holy Grail within data anonymisation, so this can be an interesting subject for future research. Another direction for future research is to find good implementations of data anonymisation according to the GDPR. When data breaches appear from anonymized data, it has to be seen what judges according to the GDPR will decide.

References

[1] [INFOGRAPHIC] Tom Le Bras, Online Overload – It's Worse Than You Thought, 21 July 2015, https://blog.dashlane.com/infographic-online-overload-its-worse-than-you-thought/

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 27 April 2016, http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

[3] "The Standard Data Protection Model: A concept for inspection and consultation on the basis of unified protection goals". Die Bundesbeauftragte für den Datenschutz und die Informationsfreiheit, March 1 2017. https://www.datenschutzzentrum.de/uploads/SDM-Methodology_V1_EN1.pdf

[4] Francis Aldhouse, Anonymisation of personal data - A missed opportunity for the European Commission, 16 July 2014, http://www.sciencedirect.com/science/article/pii/S0267364914000946

[5] Matt Wes, Looking to comply with the GDPR? Here's a primer on anonymization and pseudonmyization, 25 april 2017, https://iapp.org/news/a/looking-to-comply-with-gdpr-heres-a-primer-on-anonymization-and-pseudonymization.

[6] National Institutes of Health, HIPAA Privacy Rules for Researchers, http://privacyruleandresearch.nih.gov/faq.asp.

[7] Health System Use Technical Advisory Committee and the Data De-Identification Working Group, 'Best Practice' Guidelines for Managing the Disclosure of De-Identified Health Information http://www.ehealthinformation.ca/wp-content/uploads/2014/08/2011-Best-Practice-Guidelines-for-Managing-the-Disclosure-of-De-Identificatied-Health-Info.pdf

[8] Latanya Sweeney, Simple Demographics Often Identify People Uniquely, 2000, Carnegie Mellon University.

[9] Arvind Narayanan and Vitaly Shmatikov, Robust De-anonymization of Large Sparse Datasets, May 18 - 21, 2008, SP '08 Proceedings of the 2008 IEEE Symposium on Security and Privacy Pages 111-125, http://dl.acm.org/citation.cfm?id=1398064.

[10] Paul Ohm, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, August 13 2009, UCLA Law Review, Vol. 57, p. 1701, 2010; U of Colorado Law Legal Studies Research Paper No. 9-12., https://ssrn.com/abstract=1450006

[11] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, 10 April 2014, https://cnpd.public.lu/fr/publications/groupe-art29/wp216_en.pdf

[12] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. ACM Comput. Surv. 42, 4, Article 14 (June 2010).

[13] Latanya Sweeney. 2002. k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10, 5 (October 2002), 557-570.

[14] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1, 1, Article 3 (March 2007).

[15] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, 2007, pp. 106-11

[16] Vijay S. Iyengar. 2002. Transforming data to satisfy privacy constraints. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02). ACM, New York, NY, USA, 279-288.

[17] Adam Meyerson and Ryan Williams. 2004. On the complexity of optimal K-anonymity. In Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '04). ACM, New York, NY, USA, 223-228.

[18] Cynthia Dwork. 2006. Differential privacy. In Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II (ICALP'06), Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.), Vol. Part II. Springer-Verlag, Berlin, Heidelberg, 1-12.

Chapter 5

Appendix

5.1 Safe Harbor method

The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

1. Names
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000
3. All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
4. Telephone numbers
5. Vehicle identifiers and serial numbers, including license plate numbers
6. Fax numbers
7. Device identifiers and serial numbers
8. Email addresses
9. Web Universal Resource Locators (URLs)
10. Social security numbers
11. Internet Protocol (IP) addresses
12. Medical record numbers
13. Biometric identifiers, including finger and voice prints
14. Full-face photographs and any comparable images
15. Account numbers
16. Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section "Re-identification"]; and
17. Certificate/license numbers