Literature Survey on Privacy Preserving Mechanisms for Data Publishing

November 1, 2013

Andrei Manta

Literature Survey on Privacy Preserving Mechanisms for Data Publishing

Literature Survey

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Andrei Manta
born in Bucharest, Romania

Multimedia and Signal Processing Group
Department of Intelligence Systems
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

IBM Netherlands
Johan Huizingalaan 765
Amsterdam, the Netherlands
www.ibm.com/nl/nl

Literature Survey on Privacy Preserving Mechanisms for Data Publishing

Author: Andrei Manta
Student id: 1520172
Email: manta.s.andrei@nl.ibm.com

Thesis Committee:

Chair:	Prof. Dr. Ir. R. L. Lagendijk, Faculty EEMCS, TU Delft
University supervisor:	Dr. Zekeriya Erkin, Faculty EEMCS, TU Delft
Company supervisor:	Ing. Bram Havers, IBM Netherlands, CEWM
Company supervisor:	Drs. Robert-Jan Sips, IBM Netherlands, CAS

Contents

Chapter 1

Introduction

In recent years, the idea of Open Data has become increasingly popular [38]. The belief is that most government data should be freely available for reuse and redistribution, to be leveraged as fuel for innovation [24]. Studies have shown that Open Data, in general, has big economic value[1]. To support the growth of the Open Data movement, the Open Government Partnership (OGP)[38] has been created. Each of the member countries of the OGP has presented their country plans for Open Data [38]. The goal for the Netherlands is to open up public institution data that respects certain policies by 2015 [1]. To stimulate the usage of Open Data, the Ministry of Economic Affairs and the Ministry of Interior and Kingdom Relations have even created a temporary Knowledge Center for Open Data[2].

The Open Data movement, as we can see from the list of participating countries in the OGP, is of global proportions. Even so, there are still challenges to be dealt with. Data cannot simply be made open since issues, for example about individual privacy and national security, may arise from the usage and combination of some data sets. Therefore, guidelines have been developed that describe the process of opening up and publishing the data in a safe way. For the Netherlands, one such guideline has been developed by the dutch government[3]. Looking into the future, we can see that more data will be freely available, easier to access, and with inter-links and semantics already included (Linked Data movement [7]).

1.1 The Data

When thinking about data in terms of security/protection, one can distinguish two types: sensitive and non-sensitive data. The one that requires protection is the former. Sensitive data may concern national security (e.g. strategic dam positions), company trade secrets, individual privacy etc. To the best of our knowledge, from all these types, the literature has only considered individual privacy. From the survey we have noticed that they look at the problem in the following way: a privacy threat exists, in a given context, and a solution,

[1] TNO, Open overheid, 2011

[2] Digitale Agenda.nl: `http://www.rijksoverheid.nl/documenten-en-publicaties/notas/2011/05/17/digitale-agenda-nl-ict-voor-innovatie-en-economische-groei.html`

[3] https://data.overheid.nl/handreiking

theoretical and/or practical, is devised. There is little material, again, to the best of our knowledge, that considers other types of threat. In other words, there is no paper which tries to explain *data sensitivity*, a concept that encapsulates data privacy.

1.2 The Focus

The data, whether sensitive or non-sensitive can be split into different categories: transactional data, relational data, graph data, location/movement data. Each of these categories requires a different solution to the problem. Due to the broad nature of the topic, the main focus of this survey will only be about protecting individual privacy.

1.3 The Problem

From several interviews with representatives of different public institutions of The Netherlands (Appendix A for the interviews and Chapter 6 for the discussion) we will show that transforming a data set into Open Data is not easy. Rules and regulations are still in an incipient stage which makes the decision whether or not some data sets may be opened difficult. The responsibility to open up the data falls mostly on the shoulders of the department that owns the data, yet, as our reviews shall demonstrate, these departments do not always have the right knowledge to deal with the issues surrounding safe data publication. The guidelines which they follow cover only the obvious cases and are too general to be used as the only decision support tool. The problems appear when they are dealing with not so obvious cases, when the data looks safe but contains hidden risks. As mentioned above, there is an increase in data availability. There are more sources of data which can be combined to reveal those hidden risks and create, for example, a privacy breach. There is a big gap in knowledge on how to handle the data sanitization properly within each department. There are a few institutions which have expertise in the domain (such as CBS, O&S), but will they be able to handle all the data publishing needs of the Netherlands alone?

Since there is a clear gap between privacy preserving data publishing in practice and what happens in the academia (Section 6.1), we try to close that gap by asking ourselves the following question:

> What are the necessary steps in order to achieve privacy preserving data publishing, in the context of Open Data?

This is a difficult question and it is highly probable that a complete answer is not yet possible. However, interesting research directions, based on this concern, are discussed in Chapter 6.

1.4 Outline

The rest of the report is structured as follows. Chapter 2 explains the terminology used throughout the report. Privacy models for countering linkage and probabilistic attacks will

be discussed in Chapter 3 and Chapter 4, respectively. In Chapter 5 we present metrics for measuring data set privacy and anonymity levels. Chapter 6 concludes with the discussion and future work. The interview summaries can be found in Appendix A.

Chapter 2

Terminology and definitions

In this chapter, we will introduce important concepts that will be used through this thesis.

2.1 Privacy Preserving Data Publishing (PPDP)

There are many research directions that deal with data anonymization. This survey is focused on the Privacy Preserving Data Publishing (PPDP) direction since it relates the most to the Open Data idea (publishing without knowing the eventual usage of the data). It aims at providing methods and tools for publishing useful information while preserving data privacy. It extends the official statistics community through works that consider background attacks, inference of sensitive attributes, generalization and various notions of data utility measures [18].

2.2 Privacy Protection

Before we can start talking about different anonymization techniques, we first need to understand what it means to preserve published data privacy - a definition for privacy protection. A first such definition came from Dalenius in 1977. He states the following.

Definition 1 (Privacy protection) *Access to the published data should not enable the attacker to learn anything extra about any target victim compared to no access to the database, even with the presence of any attacker's background knowledge obtained from other sources [11].*

As good as this definition may be, it is too strict. Dwork proved that such an absolute privacy protection is impossible, due to the existence of background knowledge [14]. To circumvent this limitation, the PPDP literature has adopted a more relaxed and more practical definition.

Definition 2 (Privacy protection revised) *Access to the published data should not enable the attacker to learn anything extra about any target victim compared to no access to the database, given that the attacker has only a* limited *amount of background knowledge.*

2.3 Tables and attributes

Most of the literature on PPDP start off by making the assumption that the data publisher releases one single table which contains several types of attributes. This assumption is usually dropped later in the respective work. The literature distinguishes between four different type of attributes (table columns).

Identifiers These are attributes, or a set there-of, that fully and non-ambiguously identify a person (also referred to as "victim") to some pieces of **sensitive** information in a certain table. When a data-set is published, such attributes are always removed. Examples include SSN, passport number and name.

Quasi-identifiers(QID) Represent a set of attributes used for linking with external information in order to uniquely identify individuals in a given anonymized table. These include attributes which at first glance may seem harmless - postcode, gender, age. But according to Sweeney [40], 87% of the US population can likely be uniquely identified based only on the QID (zip code, gender, age). It falls to the data publisher to establish which attributes are to be treated as quasi-identifiers. As it will be shown later, adding too many attributes to this category will greatly impact the utility of the data. Adding too little will pose a risk to the privacy of the individuals. The PPDP literature considers that an attacker has full knowledge of an individuals QID values (full QID disclosure). Note that, usually, only privacy models that deal with linkage attacks use this notion. Throughout this paper we will be using *QID* to refer to the quasi-identifier and *qid* to refer to the value of a quasi-identifier.

Sensitive attributes (S) These attributes contain values that are considered to be sensitive to the victim. Examples of such attributes are salary and disease.

Non-sensitive attributes (NS) These attributes are composed of column in the table that do not fall under any of the previously mentioned categories.

The sequence is which the attributes are categorized is the same as above: first identifiers, then QIDs, then sensitive attributes. The rest are then considered non-sensitive attributes. One might ask that, since these attributes are sensitive, why publish them at all. The problem lies in the fact that these values are most of the time the reason why such a data set is published. Thus, the solution must rely on hindering an attacker's ability to link an individual to sensitive information. Chapter 2.4 discusses different types of privacy models, based on the attack they are trying to prevent.

2.4 Privacy models

When reasoning about privacy one can observe a duality. On the one hand, you have an individual's privacy, on the other, that of an entity. There are many privacy models discussed in the literature, yet all of them only consider the former. Throughout this report, we will be talking about an individual's privacy. One way of grouping the models is based on the

type of attack they are trying to prevent [18]. Fung et al. identify two categories: privacy models that counter *linkage attacks* (Chapter 3) and *probabilistic attacks* (Chapter 4).

Chapter 3

Countering Linkage Attacks

Linkage attacks try, as the name suggests, to link one individual to a record or to a value in a given table or to establish the presence of absence in the table itself. These type of attacks are called *record linkage*, *attribute linkage* and *table linkage*, respectively, and are discussed in more detail below.

3.1 Record linkage

As mentioned above, record linkage attacks try to link an individual to a record in a published data set. Take for example the patient table shown in Table 3.1. Here *Disease* is considered to be the sensitive attribute. Not very much can be inferred about the identity of the individuals. If an attacker, however, knows that Alice went to that hospital and also knows that Alice is a 36 year old female dancer (from a publicly available external table, e.g. a voter list), by joining the two tables on $<age, sex, job>$, the attacker infers which record belongs to Alice. The linkage is done based on the QID, which takes the values of $qid = \{age = 36, sex = female, job = dancer\}$ for Alice.

Job	Sex	Age	Disease
Engineer	male	35	Hepatitis
Engineer	male	38	Hepatitis
Lawyer	male	38	HIV
Writer	female	35	Flu
Writer	female	35	HIV
Dancer	female	35	HIV
Dancer	female	36	HIV

Alice

Table 3.1: Patient table

Name	Job	Sex	Age
Alice	Dancer	female	36
...	...	...	...

Table 3.2: External table

3.1.1 K-ANONYMITY

To prevent from such record linkage attacks, Sweeney [39] proposed k-anonymity. It requires that for each *qid* in the table, there should be at least $k-1$ other records with the same *qid*. The sensitive values that are grouped under one such *qid* are said to reside in a q^**-block* . The result of such an anonymization technique can be seen in Table 3.3. The effect of k-anonymity is that an attacker can link an individual to a record with a maximum probability of 1/k.

The proposed method to obtaining the $k-1$ other identical *qids* is through generalisation. There are two types of values for the attributes, w.r.t. the current field: numerical and categorical. Other types exist, e.g. audio, video, images etc., but are beyond the scope of this survey. For the numerical values, generalisation takes place by converting numbers to intervals that contain that number. For example, an age of 36 can be generalised to [30,40) or 3*. For categorical values, a taxonomy tree may be used used. Such a tree can be seen in Fig.3.1. Here we can see that, for example, *Dancer* can be generalised to *Artist* which can be generalised to *Any*. In our example, only Job and Age have been generalised in order to achieve 3-anonymity. This means that an attacker can link and individual to a record with at most 1/3 probability.

Job	Sex	Age	Disease
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	HIV
Artist	female	[35-40)	Flu
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV

Table 3.3: 3-anonymous table

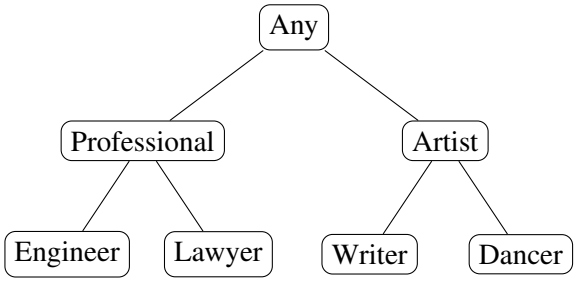

Figure 3.1: Job taxonomy tree

3.1.2 (X-Y)-ANONYMITY

In our previous example, the table contained at most one record per individual. Now consider a table in which an individual can have more than one record. If every individual would have three records in the table, the in every q*-block there would be k/3 distinct individuals. In the case of 3-anonymity, this means that a q*-block could contain only one distinct individual. This leads to a record linkage probability of 1.

To overcome this problem, (X-Y)-anonymity has been proposed by Wang and Fung [43]. Let X and Y be two disjoint sets of attributes over all attributes in the table. The attributes contained in each set are determined by the data publisher. The model states that each value on X has to be linked (co-occur) to at least k distinct values on Y. For example, take Table 3.4, which extends the previous example with a person ID *pid* for each individual. We set $X = \{job, sex, age\}$ (same as the QID) and $Y = \{pid\}$. If (X,Y)-anonymity is satisfied, then X would be linked to k distinct *pids*. This means that each q*-block would then contain k different individuals, overcoming the above mentioned problem.

Pid	Job	Sex	Age	Disease
1	Professional	male	[35-40)	Hepatitis
2	Professional	male	[35-40)	Hepatitis
1	Professional	male	[35-40)	HIV
...	...	...	...	...

Table 3.4: 3-anonymous table

Pid	Job	Sex	Age	Disease
1	Professional	male	[35-40)	Hepatitis
2	Professional	male	[35-40)	Hepatitis
1	Professional	male	[35-40)	HIV
...	...	...	...	...

Table 3.5: (<job,sex,age>-<pid>)-anonymous

3.1.3 MULTIR-ANONYMITY

Since most of the literature on k-anonymity only considered anonymizing one table, the notion of MultiRelational-anonymity has been introduced by Nergiz et al. [34]. In real life, a published data set usually contains relations between its tables. As such, anonymizing each table separately, might leak private information when these are combined. To model this, the authors consider that the set has a person table PT which contains a person identifier *pid* and some sensitive values. Furthermore, the data sets contain n tables $T_1 \dots T_n$ which contain a foreign key, some attributes in QID and some sensitive attributes. T is then defined as the join of all the tables:

$$T = PT \bowtie T_1 \bowtie \dots \bowtie T_n$$

In order to achieve MultiR-anonymity for each record owner o in T, there must be at least $k-1$ other record owners who share the same QID. See for example Table 3.6 for the initial tables and Table 3.7 for the situation after the join. After the join, our initial example is obtained, which then when anonymized, satisfies MultiR-anonymity.

Pid	Job	Disease
1	Engineer	Hepatitis
2	Engineer	Hepatitis
3	Lawyer	HIV
4	Writer	Flu
5	Writer	HIV
6	Dancer	HIV
7	Dancer	HIV

Pid	Sex	Age
1	male	35
2	male	38
3	male	38
4	female	35
5	female	35
6	female	35
7	female	36

Table 3.6: Multirelational data set

Pid	Job	Sex	Age	Disease
1	Professional	male	[35-40)	Hepatitis
2	Professional	male	[35-40)	Hepatitis
3	Professional	male	[35-40)	HIV
4	Artist	female	[35-40)	Flu
5	Artist	female	[35-40)	HIV
6	Artist	female	[35-40)	HIV
7	Artist	female	[35-40)	HIV

Table 3.7: Join of all the tables on pid - anonymized

3.2 Attribute linkage

In the case of an attribute linkage attack, the goal is to determine which sensitive value belongs to the victim. k-anonymity and its variations only protect against record linkage attacks. Two attacks fall under the attribute linkage category: the homogeneity attack and the background knowledge attack [32].

1. Homogeneity attack In the case of the homogeneity attack, the problem is the fact that all sensitive values in a q*-block have the same value. Take for example Table 3.8, a 3-anonymous patient table. Assume the attacker knows Alice (35, female, writer) and that she has been to the hospital that published this table. Due to skewness in the data, the attacker can see that all the females of age 35 who are writes suffer from the same disease: HIV. The attacker then concludes that Alice also has HIV. This attack is formally known as *positive disclosure*.

Job	Sex	Age	Disease
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV

Table 3.8: 3-anonymous patient table

2. Background knowledge attack In this attack, an attacker uses background information to eliminate possible values for the sensitive attribute of a victim. Consider the example in Table 3.9, a 3-anonymous table. This time, there is no skewness in the data, so a homogeneity attack cannot take place. Assume the attacker would like to look up Bob (38, male, lawyer). By looking at the table, he could only infer that Bob either has the flu or hepatitis. Now assume the attacker knows Bob and also knows that he does not have the flu, for example, due to lack of visible symptoms. Using this information, the attacker infers that the only possible disease Bob can have is hepatitis. The privacy principle that encapsulates this attack is called *negative disclosure*.

Job	Sex	Age	Disease
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	Flu
Artist	female	[35-40)	Flu
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV

Table 3.9: 3-anonymous patient table

3.2.1 Bayes Optimal Privacy

Machanavajjhala et al. [32] present an ideal notion of privacy called "Bayes Optimal Privacy". This notion assumes that both the data publisher and the attacker have full knowledge of the joint distribution of the sensitive and non-sensitive attributes. It requires this assumption in order to model background knowledge as a probability distribution over the attributes and uses Bayesian inference to reason about privacy. They also quantify an attackers prior (before the data has been published) and posterior beliefs (after the data has been published) based on Bayesian formulae.

Based on the model of prior and posterior belief, Machanavajjhala et al. formally define three privacy principles.

Definition 3 (Positive Disclosure) *An adversary can correctly identify the value of the sen-*

sitive attribute with high probability, that is, the attacker's posterior belief that the individual has sensitive value s is greater than $1-\alpha, \alpha > 0$.

Definition 4 (Negative Disclosure) *An adversary can correctly eliminate some possible values of the sensitive attribute with high probability, that is, the attacker's posterior belief that the individual has sensitive value s is smaller than* $\varepsilon, \varepsilon > 0$.

Definition 5 (Uninformative principle) *Published data should provide an adversary with little extra information beyond the background knowledge. This means that the difference between prior and posterior belief should be small.*

The *Bayes Optimal Privacy* model cannot be applied in practice due to several shortcomings.

Insufficient Knowledge: The publisher is unlikely to know the complete joint distribution of sensitive and nonsensitive attributes in the general population.

Adversary knowledge in unknown: It is also very unlikely that an adversary would posses such knowledge of the joint distribution. Still, the data publisher cannot know how much background knowledge an attacker may posses.

Instance-Level Knowledge This privacy model cannot protect against background knowledge that cannot be probabilistically modeled. For example, Bob's son tells Alice Bob does not have the flu.

Multiple adversaries It may be possible that different attackers exist, each with a different level of background knowledge. This implies that the data publisher must take all this into account, since different levels of background knowledge can lead to different inferences.

3.2.2 ℓ-DIVERSITY

To overcome the limitations of the optimal model, the Machanavajjhala et al. [32] propose the notion of "ℓ-diversity".

Definition 6 (ℓ-diversity) *A table is said to be ℓ-diverse if every q^*-block in the table contains at least ℓ "well-represented" values for the sensitive attribute S.*

This notion of privacy can be instantiated based on how "well-represented" is defined. There are three definitions in [32]: *distinct ℓ-diversity*, *entropy ℓ-diversity* and *recursive ℓ-diversity*. These are presented below, together with some interesting extensions [32].

DISTINCT ℓ-DIVERSITY

This definition of ℓ-diversity is the most simple. It implies that every q^*-block should contain at least ℓ distinct values for the sensitive attribute S. This also leads to k-anonymity in the case of $k = \ell$. For example, Table 3.10 is a distinct 2-diverse table.

Job	Sex	Age	Disease
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	Flu

Table 3.10: Distinct 2-diverse patient table

Job	Sex	Age	Disease
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	Flu

Table 3.11: Entropy 1.8-diverse patient table

ENTROPY ℓ-DIVERSITY

But distinct ℓ-diversity is not a very strong notion of privacy since having ℓ distinct values does not say anything about the distribution of the values. To overcome this, Machanavajjhala et al. give a definition of "well-defined" based on entropy. A table is said to be ℓ-diverse if for every q*-block in the table the following holds:

$$-\sum_{s\in S} P(qid,s)log(P(qid,s)) \geq log(\ell)$$

Here $P(qid,s)$ is the fraction of tuples in the q*-block *qid* that have the *sensitive value s*. It captures the notion of "well-represented" by the fact that the entropy increases as the frequencies of the values become more uniform. An example is Table 3.11, which is 1.8-diverse. The drawback of this definition for ℓ-diversity is that it is not very intuitive. The fact that the table is 1.8-diverse does not reflect, for example, the fact that the probability of HIV in the second group is 0.75.

RECURSIVE (C,ℓ)-DIVERSITY

Job	Sex	Age	Disease
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	HIV
Professional	male	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	Hepatitis
Artist	female	[35-40)	Hepatitis
Artist	female	[35-40)	Flu

Table 3.12: Recursive (1.1,2)-diverse patient table

The third definition for "well-represented" is more intuitive. The idea is that frequent values should not be too frequent while less frequent values should not be too rare.

Let the values of the sensitive attribute S be sorted based on frequency. f_i is the i^{th} most frequent sensitive value in the q*-block . A table is called recursive (c,ℓ)-diverse if for every q*-block the following holds:

$$f_1 < c\sum_{i=l}^{n} f_i$$

c is a user defined value. Table 3.12 is an example of a recursive (1.1,2)-diverse patient table. This means that 2-diversity is ensured if the value of c is greater or equal to 1.1. Taking a closer look we can see that by using this definition, each q*-block will contain at least ℓ distinct sensitive values for S, with an emphasis on at least. Another observation is that higher values for c lead to a higher privacy risk. For $c < n, n = |q* - block|$, one piece of background information could, in the worst case, eliminate $\frac{c}{n}$ of the tuples inside the q*-block .

For example, let $n = 10$ for a given q*-block (5 Cancer, 4 HIV, 1 Flu), and we want the q*-block to be (6,3)-diverse. $5 < 6 \cdot 1$ so it is. An attacker has now linked Bob to this q*-block . But the attacker also knows that Bob does not have cancer. He has, thus, successfully eliminated 5/10 of the tuples in the block.

Extensions

An interesting extension to ℓ-diversity discussed in [32] are the *don't care sets*. It may be the case that some of the sensitive attribute values may be disclosed without causing a privacy breach. For example, it might not be a problem if it would be disclosed that an individual was negative on a certain STD test. In [32] entropy ℓ-diversity and recursive (c,ℓ)-diversity have been extended to handle such a situation. Due to the complex nature of the extension, we refer the reader to [32] for details.

MULTI-ATTRIBUTE ℓ-DIVERSITY

When considering multiple sensitive attributes, new issues arise. Mainly, the due to correlation between the attributes. For example if an attacker can use background knowledge to eliminate some sensitive attribute s, then he can also eliminate any other sensitive value associated with s. To overcome this, Multi-Attribute ℓ-diversity has been proposed [32]. It assumes the QID attributes $Q_1 \dots Q_n$ and the sensitive attributes $S_1 \dots S_m$.

Definition 7 (Multi-Attributeℓ-diversity) *A table is said to be multi-attribute ℓ-diverse if for all i=1..m, the table is ℓ-diverse when S_i is treated as the sole sensitive attribute, and $Q_1 \dots Q_n S_1 \dots S_{i-1} S_{i+1} \dots S_m$ as the QID attributes.*

3.2.3 T-CLOSENESS

We saw in the previous section how ℓ-diversity protects against attribute linkage attacks. Yet there are two types of attacks for which ℓ-diversity cannot prevent attribute disclosure.

1. Skewness attack Take for example a table which has one sensitive attribute: the result for a STD test. 99% of the people are negative while the rest positive. Due to the skewness of the data you could have partitions of the data in which half the values are positive and have are negative. In such a case, someone linked to this q*-block would have 50% chance of being positive, while in the general population only, this would be only 1%. Another example would be when a q*-block has 98% positive values and 2% negative values. Someone linked to this block would almost certainly be categorised as having the STD. In both examples distinct 2-diversity is satisfied.

2. Similarity attack Another problem arises when the values in a q*-block are distinct, but semantically similar. For example, all the patients in a q*-block have some form of lung disease. Having ℓ distinct values does not protect against attribute disclosure.

A solution to this problem is the notion of t-closeness [28]. The authors use the notion of "equivalence class"(EC) instead of q*-block .

Definition 8 (t-closeness) *A table is said to achieve t-closeness if for every equivalence class (q*-block) in the table, the distribution of a sensitive values in the group is within t of the distribution of values in the the whole population.*

But how to properly measure the distance between distributions. In general, the variational distance or the Kullback-Leibler divergence would achieve this. However, in our case, we have one extra constraint. It also needs to measure the semantic distance. The authors decided to use the Earth Mover Distance (EMD) to this end. Intuitively, it is measured as the minimum work required in order to transform one distribution into the other, by moving distribution mass between the two.

There are two formulae presented in [28]. One for numerical and one for categorical values. To understand how EMD works, the formula for the numerical case is presented bellow, since it is easier to understand. For the categorical case we refer the reader to [28].

Let P and Q be two distributions with p_i and q_i probabilities of element i in each distribution, respectively. Further, let $r_i = p_i - q_i$. Then, the distance between P and Q is:

$$D[P,Q] = \frac{1}{m-1} \sum_{i=1}^{i=m} \left| \sum_{j=1}^{j=i} r_j \right| \tag{3.1}$$

A numerical example might help. Let $P = \{3k, 4k, 5k\}$ and $Q = \{3k,\ 4k, 5k, 6k, 7k,$ 8k, 9k, $10k, 11k\}$ be two salary distributions. Thus $p_i = \frac{1}{3}$ and $q_i = \frac{1}{9}$. This leads to the following values for $r_{1...9} = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, -\frac{1}{9}, -\frac{1}{9}, -\frac{1}{9}, -\frac{1}{9}, -\frac{1}{9}, -\frac{1}{9}\}$ Intuitively, we can solve this by moving r_1 amount from the first element to the second, $r_1 + r_2$ from the second to the third element etc. If we apply formula 3.1, we get the following distance between P and Q

$$D[P,Q] = \frac{1}{9-1}(\frac{2}{9} + \frac{4}{9} + \frac{6}{9} + \frac{5}{9} + \frac{4}{9} + \frac{3}{9} + \frac{2}{9} + \frac{1}{9}) = 0.375$$

Intuitively, the distance between P and Q is large. If an individual gets linked to this EC, then the attacker could conclude that the individual has a small salary (compared to the full

Postcode	Age	Salary
476**	2*	3k
476**	2*	4k
476**	2*	5k
4790*	≥40	6k
4790*	≥40	11k
4790*	≥40	8k
476**	2*	7k
476**	2*	9k
476**	2*	10k

Table 3.13: 3-diverse / 0.375-close table

Postcode	Age	Salary
4767*	2*	3k
4767*	2*	5k
4767*	2*	9k
4790*	≥40	6k
4790*	≥40	11k
4790*	≥40	8k
4760*	2*	4k
4760*	2*	7k
4760*	2*	10k

Table 3.14: 0.167-close table

distribution), since all the values in the EC are small. $D[P,Q] = 0.375$ proves the intuition right. Good values for t are usually small, around 0.1. For example, one version of the table before t-closeness could look like Table 3.13. After t-closeness has been applied, it would look like Table 3.14.

t-closeness solves the problems of k-anonymity and l-diversity, but has some limitations of its own. First, it lacks the flexibility to specify different protection levels for different sensitive values. Second, the EMD distance may not be suited for numerical values [27]. Third, the distortion of the data is very large because each EC must have a similar distribution.

3.2.4 (N,T)-CLOSENESS

(n,t)-closeness tries to compensate for the limitations of t-closeness. It distorts the data less and also has an algorithm to achieve it. It is an extension of t-closeness. Instead of forcing the distribution of all equivalence classes (EC) be within t distance to the full table distribution they now require that the EC be within t of a large enough population of the data set.

Definition 9 ((n,t)-closeness) *An equivalence class E_1 is said to have (n,t)-closeness if there exists an EC E_2 such that E_2 is a natural superset of E_1, has more than n elements and the distance between the sensitive attribute distributions of E_1 and E_2 is at most t. A table has (n,t)-closeness if all equivalence classes in the table have it [29].*

A natural superset of an equivalence class is an equivalence class which contains it by broadening the set value boundaries. Take for example E_1 to be the first EC in table 3.16. E_1 is thus defined as (zipcode='476**', age=[20,29]). A natural superset would then be (zipcode='476**', age=[20,39]) since it "naturally" contains it.

Take the tables 3.15 and 3.16 for example. Table 3.16 does not respect 0.1-closeness, but respects (1000,0.1)-closeness. There are three equivalence classes. The second one achieves (1000,0.1)-closeness simply because it is a natural superset of itself (it contains 2000 elements). The first and third do not have 1000 elements, but their union (which is also their superset), does and the distribution does not change.

Zipcode	Age	Disease	Count
47673	29	Cancer	100
47674	21	Flu	100
47605	25	Cancer	200
47602	23	Flu	200
47905	43	Cancer	100
47904	48	Flu	900
47906	47	Cancer	100
47907	41	Flu	900
47603	34	Cancer	100
47605	30	Flu	100
47604	36	Cancer	100
47607	32	Flu	100

Table 3.15: Patient table

Zipcode	Age	Disease	Count
476**	2*	Cancer	300
476**	2*	Flu	300
479**	4*	Cancer	200
479**	4*	Flu	1800
476**	3*	Cancer	200
476**	3*	Flu	200

Table 3.16: (1000,0.1)-closeness

Once again, the authors mention the problem of choosing an appropriate metric for measuring the distance between two distributions. To make it more concrete, they came up with a five point desiderata.

Identity of indiscernibles $D[P,P] = 0$; there is no information gain if the belief does not change.

Non-negativity The information gain can only be positive; $D[P,Q] \geq 0$

Probability scaling Increasing one's belief probability by a constant γ should be more significant for smaller initial belief than for larger; $(\alpha \to \alpha+\gamma) > (\beta \to \beta+\gamma); \alpha < \beta$

Zero-probability definability The distance metric should be able to handle zero probability values within the distributions (KL divergence, for example, does not).

Semantic awareness The metric should reflect the semantic distance between the values, if any.

The previously proposed EMD metric does not have the probability scaling. As such, they proposed a new metric which respects all the points mentioned above. It consists of two steps. First, kernel smoothing. They apply the Nadaraya-Watson kernel weighted average over both distributions:

$$\hat{p} = \frac{\sum_{j=1}^{m} p_j K(d_{ij})}{\sum_{j=1}^{m} K(d_{ij})}$$

where $K(.)$ is the kernel function and d_{ij} is the distance between the sensitive values s_i and s_j. The main advantage of this is that it requires a distance matrix between the values, distance which considers semantic meaning. There are two ways of computing the distance.

Age	Sex	Zipcode	Disease	GN
5	m	12000	gastric ulcer	stomach disease
9	m	14000	dyspepsia	dyspepsia
6	m	18000	pneumonia	respiratory infection
8	m	19000	bronchitis	bronchitis
12	m	22000	pneumonia	pneumonia
19	m	24000	pneumonia	pneumonia
21	f	58000	flu	ø
26	f	36000	gastritis	gastritis
28	f	37000	pneumonia	respiratory infection
56	f	33000	flu	flu

Table 3.17: Patient table with guarding node

Given two sensitive values s_1 and s_2, the left formula is for the numerical case, the right one for the categorical case:

$$d_{ij} = \frac{|s_i - s_j|}{R} \qquad d_{ij} = \frac{h(s_i - s_j)}{H}$$

where R is the attribute domain range, H the height of the taxonomy tree and $h(\_)$ is the height of the lowest common ancestor in the distance tree.

From the kernel smoothing we obtain $\hat{P}$ and $\hat{Q}$. Now instead of computing the distance $D[P,Q]$ they compute the estimate of the distance, $D[\hat{P},\hat{Q}]$ using the Jensen-Shanon divergence:

$$JS[P,Q] = \frac{1}{2}(KL[P,avg(P,Q)] + KL[Q,avg(P,Q)])$$

where $avg(P,Q)$ is the average distribution $(P+Q)/2$.

They also propose an algorithm to actually achieve (n,t)-closeness. This is based on the Mondrian algorithm [26]. It has tree parts: choosing a dimension to partition, choosing a value to split and checking whether the partitioning violates the privacy conditions. Their modification consist of a check algorithm for the third part.

3.2.5 PERSONALIZED PRIVACY

This privacy models tries to compensate over-generalisation, and thus unnecessary utility loss, by taking into account the privacy levels desired by each individual. So instead of applying a uniform privacy level, which may be overprotective for some, and under-protective for others, it customizes it to the needs of the individual. This is modeled by adding one extra column to the data set. Given a taxonomy tree, this column specifies which subtree the individual does not want to be associated with. For example, take Table 3.17 and taxonomy tree in Figure 3.2. The first person does not want to be linked to any stomach disease (gastric ulcer, dyspepsia, gastritis). Thus he does not mind being linked to a respiratory infection, for example. On the other hand, the seventh person has flu and does not mind fully opening her data (her guarding node is ø).

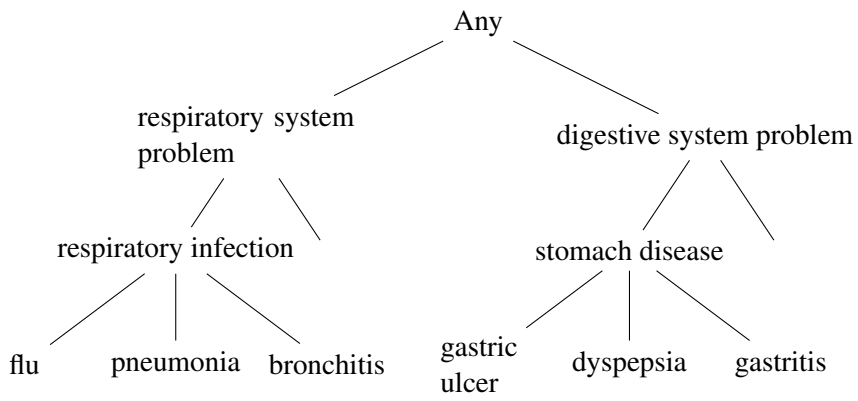

Figure 3.2: Disease taxonomy tree

The algorithm guarantees that an adversary can infer from the published sanitized table T* a value that breaches an individuals privacy with at most probability $\mathcal{P}_{breach}(t)$. In order to compute this probability one needs to know the external table P to which the adversary will link. This, however, is very hard to do in practice since one cannot know the background knowledge of an adversary [18]. The idea is that one can compute $\mathcal{P}_{breach}(t)$ by looking at the number of possible combinations of assigning people and values from P to T, which breach the individuals desired level of privacy.

The algorithm for personalized privacy consists of three steps. First, it generalises the QI attributes using the same function for each tuple, similar to what k-anonymity, ℓ-diversity and t-closeness do. Then, equivalence classes are formed based on the QID. The last step involves a generalisation of the sensitive attributes, but this time using a custom function for each tuple. Overall, the utility is greater because the QIDs are generalized less. The result of the algorithm applied to Table 3.17 can be seen in Table 3.18.

The idea is good in theory but has several drawbacks in practice:

- $\mathcal{P}_{breach}(t)$ depends on external table
- individuals do not have access to the distribution of their sensitive values, as such they tend to be overprotective in selecting their guarding node, which leads to higher utility loss
- getting people to fill in their guarding nodes is not easy, especially post data gathering.

Due to these drawbacks, this privacy model could only be applied in practice to a very limited number of data sets

3.3 Table Linkage

In a table linkage type of attack, the adversary tries to infer with a high enough probability that a certain individual is or is not present in a certain published sanitized table. Take for example someone who has diabetes and took part in a study. This study is then published in

Age	Sex	Zipcode	Disease
[1,10]	m	[10001,20000]	gastric ulcer
[1,10]	m	[10001,20000]	dyspepsia
[1,10]	m	[10001,20000]	respiratory infection
[1,10]	m	[10001,20000]	respiratory infection
[11,20]	m	[20001,25000]	respiratory infection
[11,20]	m	[20001,25000]	respiratory infection
21	f	58000	flu
[26,30]	f	[35001,40000]	gastritis
[26,30]	f	[35001,40000]	pneumonia
56	f	33000	respiratory infection

Table 3.18: Anonymized patient table

a sanitized way. Diabetes is quite an expensive disease as treatments can reach well above 10000$ per year. Employers may abuse such study information to infer if a job candidate has diabetes or not, since this may incur extra costs for the company. On the other hand, if an employer has two candidates to choose from, he could use such information to see which of the individuals has a higher chance of not having diabetes. This section presents an algorithm that tries to prevent such attacks.

3.3.1 δ-PRESENCE

	Name	Zip	Age	Nationality	Sen
a	Alice	47906	35	USA	0
b	Bob	47903	59	Canada	1
c	Christine	47906	42	USA	1
d	Dirk	47630	18	Brazil	0
e	Eunice	47630	22	Brazil	0
f	Frank	47633	63	Peru	1
g	Gail	48973	33	Spain	0
h	Harry	48972	47	Bulgaria	1
i	Iris	48970	52	France	1

Table 3.19: Public table

	Zip	Age	Nationality
b	47903	59	Canada
c	47906	42	USA
f	47633	63	Peru
h	48972	47	Bulgaria
i	48970	52	France

Table 3.20: Research subset

The goal of δ-presence is to bound the probability of inferring that an individual is or is not in the table by $\delta = (\delta_{min}, \delta_{max})$. More formally

$$\delta_{min} \leq \mathcal{P}(t \in T|T^*) \leq \delta_{max}$$

where T* is the generalised table. This reads: the probability of tuple t belonging to the original table, given that the adversary sees the sanitized table, is bounded by δ_{min} and δ_{max}.

The authors provide a way to compute this probability, namely

$$\mathcal{P}(t \in T|T^*) = \frac{C(p_1^*, T^*)}{\sum_{t|p_1^* \in \Psi(t)} C(t,P)}$$

but in requires knowledge of the external table P. $C(t,P)$ is the cardinality of tuple t in table P. By cardinality, we mean the number of tuples with the same QID. p_1^* is the tuple in T^* that is in the generalisation set $\Psi(t)$ (e.g. $\Psi(LosAngeles)$ = {California, USA, North-America, America}). To give a concrete example, please consider the following example [35] tables P (3.19), T (3.20), P*(3.21) and T*(3.22).

$\mathcal{P}(tuple\,a \in T|T*) = \frac{|\{b,c,f\}|}{\{a,b,c,d,e,f\}} = \frac{1}{2}$. In the same way we obtain 1/2 for b,c,d,e,f. For g,h,i we have $\mathcal{P}(tuple\,a \in T|T*) = \frac{|\{h,i\}|}{\{g,h,i\}} = \frac{2}{3}$. Thus table T* satisfies ($\frac{1}{2}$, $\frac{2}{3}$)-presence.

	Name	Zip	Age	Nationality	Sen
a	Alice	47*	35	America	0
b	Bob	47*	59	America	1
c	Christine	47*	42	America	1
d	Dirk	47*	18	America	0
e	Eunice	47*	22	America	0
f	Frank	47*	63	America	1
g	Gail	48*	33	Europe	0
h	Harry	48*	47	Europe	1
i	Iris	48*	52	Europe	1

Table 3.21: Sanitized public table

	Zip	Age	Nationality
b	47*	*	America
c	47*	*	America
f	47*	*	America
h	48*	*	Europe
i	48*	*	Europe

Table 3.22: Sanitized research subset

It is a nice privacy model but the main downside is that it assumes that the adversary and the data publisher have access to the same external table, which is not very likely to happen in practice.

Chapter 4

Countering Probabilistic Attacks

Probabilistic attacks try to increase the confidence of an attacker that a certain individual has some sensitive value x. In other words, an attacker tries to gain as much information as possible about an individual, from a published table, beyond his own background knowledge. For example, an attacker, based on his background knowledge, could say prior to the publishing that the victim has disease HIV with probability 0.1. If he can somehow conclude that, after the table has been published, the victim has HIV with probability 0.8, then the attacker has successfully carried out a probabilistic attack. It is successful, because the difference between the prior (0.1) and posterior (0.8) probabilities is big enough. This section presents several privacy models that try to prevent such attacks. This is done in most cases by trying to limit the difference between the prior and posterior beliefs, also known as the uninformative principle [32].

4.1 (C,T)-ISOLATION

This privacy model is similar to the ones trying to prevent record linkage, but still handle probabilistic attacks. The idea is that given a statistical database, an adversary should not be able to isolate any points in the database, corresponding to an individual. This is another way of hiding in the crowd. Formally, given a real database (RDB) point y and an adversary inferred point p, obtained after seeing the sanitized database (SDB) and using auxiliary information, let $\delta_y = \|q - y\|$ be the distance between the two points.

Definition 10 ([10]) *Point p (c,t)-isolates y if $B(p, c\delta_y)$ contains less than t points in the RDB, that is, $|B(p, c\delta_y) \cap RDB| < t$, where $B(p, c\delta_y)$ is the ball of radius $c\delta_y$ around the point p.*

To achieve this, they start off by researching two intuitions. First, that histograms preserve privacy. One cannot infer anything from a histogram about an individual. The downside is, histograms offer only limited utility. Second, they say that density based perturbation (noise with magnitude a function of the local density of points) offers more utility. This is true since points in a dense region require less perturbation. Another benefit would

be that high density regions ensure the "hiding in the crowd" of points. Based on this they propose an anonymization algorithm which combines the two methods mentioned above.

The first step consists of randomly dividing the dataset in two sets: A and B. From B they construct a histogram (each cell containing at most 2t points). Then set A is sanitized (through data perturbation) using the position of the points in the histogram of B. Take a point $v \in$ A. Let ρ_v be the side length of the cell containing v in B. The next step is to pick at random a point N_v from a spherical Gaussian with variance ρ_v^2 in each coordinate. Finally, release $v' = v + N_v$.

This privacy model has some properties with respect to utility and privacy.

Utility. Chawla et al. state that after the anonymization process has been applied, learning a mixture of Gaussians is still possible. Also, it is possible to find k clusters each of cardinality at least t. Given an algorithm that can approximate the optimal clustering to within a constant factor on the sanitized data, it can also approximate the optimal clustering of the original data to within a constant factor.

Privacy. To ensure proper privacy, high-dimensionality is required. It is also not possible to make the sanitization arbitrarily safe. The authors state that a polynomial time adversary can succeed without any background knowledge with a probability $\Omega(e^{-d}/log(n/t))$.

As final remarks, since the model involves distances between high-dimensional points, it seems more useful when applied to a numerical database. One big limitation of this model is that it reasons about an adversary which does not have access to any auxiliary information.

4.2 ε-DIFFERENTIAL PRIVACY

Most privacy models that deal with probabilistic attacks usually try to limit the difference between an attackers prior and posterior beliefs about the sensitive value of a victim. ε-differential privacy [14], on the other hand, tries to limit the increase of the risk to one's privacy when an individual contributes his data to a statistical database. This means that if one's privacy is breached, that would have happened anyway, with or without the person's records in the database.

Dwork has two major contributions to the field through [14]. The privacy model and the formal proof that privacy cannot be achieved when an adversary has access to unlimited auxiliary information.

The privacy model proposed ensures that the outcome of the randomized sanitization algorithm does not significantly vary in the presence or absence of a single record. Formally

Definition 11 (ε-differential privacy) *A randomized function $\mathcal{K}$ gives ε-differential privacy if for all datasets D_1 and D_2, differing on at most one element, and all $S \subseteq Range(\mathcal{K})$,*

$$Pr[\mathcal{K}(D_1) \in S] \leq exp(\varepsilon)[Pr[\mathcal{K}(D_2) \in S]$$

This privacy model, however, is applied in an interactive setting, when the user sends queries to the database. Between the user and the database lies the privacy mechanism. It is the mechanism which returns a perturbed answer to the user. The noise added is usually

based on the Laplace distribution. To make it efficient, the noise depends on the magnitude of the largest change a single participant could have on the output to the query function. This change is also called the sensitivity of the output function (Δf). Based on this sensitivity, one aims to achieve ($\Delta f/\sigma$)-differential privacy, which can be done be respecting $\sigma \geq \varepsilon/\Delta f$.

In a follow up paper ([15]) Dwork showed that such a sanitized database can only have a limited life-span. She shows that when the **total** number of queries is sublinear in n, then the magnitude of the noise required to achieve ε-differential privacy is $o(\sqrt{n})$.

In the survey on differential privacy [16] it can be seen that some non-interactive database model can also be anonymized using their model, basically, by simulating the queries over the database. As far as the usefulness of the model goes, several algorithms and applications are proven to work:
Algorithms

- statistical data inference (this is the non-interactive model example)
- contingency table release
- learning (nearby) halfspaces

Concepts/Applications

- concepts learnable in the statistical query model are also privately learnable (e.g. SuLQ database [8])
- private PAC learning is possible
- differentially private queries of classes with polynomial VC dimension [1]

There are claims that differential privacy works without any assumptions about the data. This is, however, not the case [23]. Kifer and Machanavajjhala show that the claim is not valid for example in social networks or when other deterministic statistics are already publicly available. In the former case, the presence of an individual may cause certain edges to form between nodes. Thus simply removing such a node will not delete the evidence of participation. In order to be able to properly apply differential privacy, they show that assumptions about how the data was generated are required. In the latter case, previously released statistics might add correlation between the released noisy answers. The noise added by the mechanism can then be filtered out allowing for the initial table to be reconstructed with a very high probability.

In both [23] and [33] it is noted that differential privacy might return answers to queries that may not be useful in practice. In the later paper it is showed what the effects of differential privacy are on the *Age* of Lithuanian women and *Salary* of American citizens. Using the sensitivity of the dataset (see above) to determine ε, they show how skewness of the data has a big impact on utility. Furthermore, the more specific the target group is(e.g. Lithuanian women, older than 50 who have a job), the less utility can be gain from the data. To better grasp the magnitude of the distortion, a short example follows. In the salary experiment,

[1] Vapnik-Chervonenkis dimension is a measure of the capacity of a statistical classification algorithm.

if the full set is considered, then the noise added to the salary is about ±250$. This leads to a privacy breach for wealthy people (for an income of 5 million, ±250 is meaningless noise). On the other hand, if only a state would be selected, then the noise added would be ±250000, leading possibly to negative incomes.

4.3 (d,γ)-PRIVACY

The privacy model tries to limit the difference between the prior and posterior probability that an individual's record is present in the table. One important contribution of this paper is the formal guarantee on privacy and utility. But in order to achieve this, the authors make the assumption of *d-independence* (see below), which may not always be possible in practice.

The algorithm presented in the paper to anonymize the dataset uses data synthesis based on the existing table and the attribute domain. There are two steps. First, each record in the original table T gets added to the sanitized table T* with probability $\alpha+\beta$. Second, tuples from the domain of all possible records which are not in T, are added to T* with probability β. Hence the name of the algorithm, $\alpha\beta$ algorithm.

Definition 12 (d-bounding) *An adversary is called d-bounded if for all tuples $t \in D$ either the prior probability $Pr[t] \leq d$ or $Pr[t] = 1$.*

Definition 13 (d-independence) *An adversary is called d-independent if he is d-bounded and tuple-independent (table tuples are independent of each other).*

A *d-independent* adversary thus either has a small enough prior or he already knows the individual's record in the table.

The attacker's prior probability that on individual's record is in the table, k, is defined as $Pr_1[t] \leq k\frac{n}{m}$. n is the number of records in the table, m is the size of the domain D. k represents the power of the adversary. $Pr_{12}[t]$ is called the attacker's posterior belief and depends on the published anonymized table (view) and the attacker's prior.

Definition 14 ((d,γ)-privacy [35]) *An algorithm is called (d,γ)-private if the following holds for all d-independent adversaries with Pr_1 and for all possible table sanitizations T* and tuples t s.t. $Pr_1[t] \leq d$:*

$$\frac{d}{\gamma} \leq \frac{Pr|12[t|T*]}{Pr_1[t]} \qquad \textit{and} \qquad Pr_{12} \leq \gamma$$

If the equation on the right fails, then a positive leakage ((d,γ) breach) has takes place. This simply means that the difference between the prior and posterior is too large. The left hand side can be seen as a negative leakage. That is, the posterior probability may not drop by a factor greater than $\frac{d}{\gamma}$.

Utility. Their definition of utility is based on the error bounds of estimates of counting queries.

Definition 15 ((ρ,ε)-useful [35]) *A randomized algorithm is called (ρ,ε)-useful if there exists a query estimator $\hat{Q}$ s.t. $Pr[|Q(T)-\hat{Q}(T^*)| \geq \rho\sqrt{n}] \leq \varepsilon$.*

In other words, with high probability $1-\varepsilon$, the error on the counting query estimation is at most $\rho\sqrt{n}$.

Privacy. With respect to privacy, the authors prove that if k, the adversary's prior, is $k = \Omega\sqrt{m}$, which in turn implies that $d = kn/m = \Omega(n/\sqrt{m})$, then no algorithm can achieve both privacy and utility in this setting.

Improvements. They also provide two extensions to their algorithm. First, they show how one can publish updated datasets while still preserving privacy. This is achieved by taking into account all previously published tables in the tuple selection step. Second, they show how to model a variable adversary's prior for each tuple. It may be the case that some tuples have a higher prior than others.

4.4 DISTRIBUTIONAL PRIVACY

Blum et al. [9] present a privacy model which is similar to differential privacy but strictly stronger. Distributional privacy states that any privacy mechanism that is applied to a database that is drawn from a distribution, should only reveal information about the underlying distribution and nothing else.

In the paper they show, that for any discretized domain and for any concept class with polynomial VC-dimension, it is possible to publish a differentially private database which can answer all queries in the concept class. Furthermore, there is no limit on the number of queries that can be answered (as opposed to differential-privacy which requires sublinear number in n for a noise addition of $o(\sqrt{n})$).

One downside of this model is that it has high computational costs. The authors only show an efficient algorithm for the class of range queries over a finite interval and with bounded precision. They also show that for non-discretized domains, this is not possible under the current usefulness definition (similar to the usefulness definition in (d,γ)-privacy). The definition is based on the interactive database model:

Definition 16 ((ε,δ)-usefulness [9]) *A database mechanism A is (ε,δ)-useful for queries in class C if with probability $1-\delta$, for every $Q \in C$ and every database D, for $\widehat{D} = A(D)$, if $|Q(\widehat{D}) - Q(D)| \leq \varepsilon$.*

To overcome this problem, they loosen the definition of usefulness to the following:

Definition 17 ($(\varepsilon,\delta,\gamma)$-usefulness [9]) *A database mechanism A is $(\varepsilon,\delta,\gamma)$-useful for queries in class C according to some metric d if with probability $1-\delta$, for every $Q \in C$ and every database D, for $\widehat{D} = A(D)$, if $|Q(\widehat{D}) - Q'(D)| \leq \varepsilon$ and $d(Q,Q') \leq \gamma$.*

This makes answering queries about (nearby) halfspaces possible. Instead of answering the exact query, the algorithm replies with an approximately correct answer to a nearby query.

There are two formal definitions for the distributional privacy model. One for the interactive, and one for the non-interactive database model.

Definition 18 (**(α,β)-distributional privacy - interactive model [9]**) *An interactive database mechanism A satisfies (α,β)-distributional privacy if for any distribution over database elements $\mathcal{D}$, with probability $1-\beta$, two databases D_1 and D_2 consisting of n elements drawn without replacement from $\mathcal{D}$, for any query Q and output x satisfies $Pr[A(D_1,Q)=x] \leq e^{\alpha} Pr[A(D_2,Q)=x]$.*

Definition 19 (**(α,β)-distributional privacy - non-interactive model [9]**) *A non-interactive database mechanism A satisfies (α,β)-distributional privacy if for any distribution over database elements $\mathcal{D}$, with probability $1-\beta$, two databases D_1 and D_2 consisting of n elements drawn without replacement from $\mathcal{D}$ and for all sanitized outputs $\widehat{D}$, $Pr[A(D_1)=\widehat{D}] \leq e^{\alpha} Pr[A(D_2)=\widehat{D}]$.*

Chapter 5

Measures of privacy and utility

Different algorithms provide different levels of privacy and utility. Some algorithms use such measurements internally, to guide them during the process of anonymization. Others come with provable guarantees that certain levels of privacy or utility are met. There are also ways to measure the privacy and utility levels post-anonymization. The only problem is - there is no general solution. What something is useful for someone, may not be useful for somebody else. What something is safe enough on a record level, it might not be safe on an aggregate level (e.g. what patterns can be extracted).

Below we start by discussing metrics for privacy and utility. We continue by presenting the R-U maps, which rely on such metrics to visualize the trade-off between risk and utility. We continue with frameworks to help choose the best anonymization technique based on the desired level of privacy and utility - based on information theory and based on data mining.

Lambert [25] formally defines the risk and the hard of possible disclosures. She states that an adversary does not need to make a correct inference, for it to be harmful to an individual. She shows how to reason and compute the risk and harm based on what the goals of the attacker might be.

Two survey papers are of importance when speaking about metrics in a general sense. First, Venkatasubramanian [42] presents different ways to measure anonymity, depending on the technique used to anonymize the data set. Second, Fung et al. [18] present among others different metrics used to measure the utility retained by the data set. An overview will be presented here of the most important methods. For more details please refer to the mentioned papers.

5.1 Measures of Anonymity

Venkatasubramanian [42] distinguishes between three types of measures for anonymity: statistical methods (based on the variance of key attributes), probabilistic methods (based on Bayesian reasoning to quantify information gain and information loss) and computational methods (assume the adversary is resource bounded). We briefly present a few methods from the first two types. For the complete list of measures please refer to [42].

5.1.1 Statistical measures

Adding noise to the data is one of the methods used to preserve privacy. As such, one way to measure privacy is to determine the amount of variance in the perturbed data, the more, the better [13]. Another way is to measure the length of the confidence interval of the estimator [4]. The longer the interval, the better the anonymization.

5.1.2 Probabilistic measures

Mutual Information [3] One way to measure a privacy leak is based on mutual information. $I(A;B) = H(A) - H(A|B)$ is the difference between the entropy of the random variable A and the conditional entropy of A given B. The first is seen as the uncertainty, the second, as the privacy after seeing B. The amount leaked is then the following: $P(A|B) = 1 - 2^{H(A|B)}/2^{H(A)} = 1 - 2^{-I(A;B)}$. It is represented as a power of 2 since entropy is usually expressed in bits of information.

Transfer of information [17] This relates to the Uninformative principle [32]. The idea is to measure whether the information gain is bounded. A privacy breach occurs when $P(X) \leq \rho_1, P(X|Y=y) \geq \rho_2$, where $\rho_1 \ll \rho_2$.

Generalization based methods include already discussed metrics for k-anonymity, ℓ-diversity and t-closeness (e.g discernibility metrics, earth-mover distance, KL-divergence).

5.2 Measures of Utility

It has been pointed out that utility is hard to quantify. Still, there are different types of metrics to achieve this. Fung et al. [18] distinguish between general purpose metrics (which should be used when it is not known how the data will be used), specific purpose metrics (measure utility based on the usage goal) and trade-off metrics (which consider the trade-off between privacy and utility).

5.2.1 General purpose metrics

The first general purpose metric introduces is the minimal distortion metric (MD) [36, 39, 40, 43]. In this case, a penalty is induced for each value that has been changed w.r.t. the original; it is a single attribute measure.

ILoss [46, 26, 32] is a data metric used to measure information loss by generalisation. For each attribute in a record, $\frac{|v|-1}{|D_A|}$ is determined, where $|v|$ is the number of leaves in the taxonomy tree under v and D_A are the number of values in the domain. It is possible to assign a weight to each attribute in the record. To obtain the information loss for a table, the sum of *ILoss* of all records is taken.

The discernibility metric incurs a penalty for every record (QID of such record) that is indistinguishable from others. It goes against k-anonymity in the sense that the more the records differ, the more utility one has.

5.2.2 Specific purpose metrics

The only mentioned metric is the classification metric (CM) [22] which measures the classification error on the data. "A penalty is incurred for each record that is generalised or suppressed to a group in which the records class is not the majority class". The method was developed to cope with the fact that future data is usually not available, making it impossible to see what the classification error is.

5.2.3 Trade-off metrics

There is always a trade-off between utility and privacy. Focusing only on one will make the other be impossible to achieve. Usually, such trade-off metrics help in the process of anonymization by guiding the search. One proposed metric is the information gain / privacy loss metric (IGPL) [19, 20]. $IGPL(s) = \frac{IG(s)}{PL(s)-1}$. IG(s) is a specialization operation which gains information and at the same time loses privacy PL(s). Defining IG and PL depends on the goal of the anonymization.

5.3 Risk-Utility Maps

The risk-utility map (R-U map) has been initially proposed by Duncan et al. [12]. The idea was initially proposed for noise addition to determine which amount of noise was best to minimize disclosure risk (privacy leaks) while still retaining enough privacy. It requires a way to measure risk and one to measure utility. Then the anonymization algorithms are run using different values for their parameters. Each run generates a 2D point based on the value of R (risk metric) and U (utility metric). This plot can then be used to determine appropriate values for the parameters. Duncan et al. [12] use three scenarios to help them determine R: when the adversary cannot determine that the target is present in the dataset, when the adversary can isolate the target to the data set and third, when the adversary has enough external information to perform record linkage.

The flexibility and simplicity of the model makes it a very powerful tool. The difficulty lies in finding good metrics for utility and disclosure risks. Duncan et al. show how it can be applied on to relational data. Loukides et al. [31] present in their paper how R-U maps can be used for transactional data. They compare three anonymization techniques: Apriory Anonymization (AA) [41], Constrained-based Anonymization of Transactions (COAT) [30] and Privacy-constrained Clustering-based Transaction Anonymization (PCTA) [21]. In this case, the utility, for example, is measured as 1/ARE, where ARE is the average relative error in query answering (in this case, how many incorrect transactions have been retrieved). On deciding what a good trade-off is, they chose the *knee point* - the most significant change in the curve of the plot - to represent the best trade-off possible.

Xia et al. [45] tackle the problem a the trade-off a bit differently. Instead of a simple curve, one has a scatter plot based on the generalisation space. They introduce the notion of de-identification policy and formalize the problem of de-identification policy frontier discovery. A de-identification policy represents a set of domain partitions for each QI attribute. Their goal is then to find the set of policies with the best risk-utility trade-off (how to best

partition the data set through generalisation). One of their findings is that the frontier proposed by the Safe Harbor standard of the HIPAA Privacy Rule is suboptimal to the one their algorithm has discovered.

5.4 Information theoretic measures

Bezzi [6] presents a framework, based on one-symbol information (the contribution to mutual information by a single record), to help determine which anonymization technique provides the best privacy guarantee for a given case. To do so, they show how to redefine existing metrics into information theoretic ones. To achieve this they introduce two notions: the Surprise (j-measure) and the Specific Information (i-measure).

$$Surprise(x) = \sum p(y|x) log_2 \frac{p(y|x)}{p(y)}$$

The surprise represents the KL-distance between the marginal distribution $p(y)$ and the conditional probability distribution $p(y|x)$. It is called "surprise" since it increases with less likely events.

$$i - measure = H(Y) - H(Y|x)$$

The specific information captures how diverse entries are for the random variable Y for a given value x. Using these notions, Bezzi translates the specifics of k-anonymity, ℓ-diversity and t-closeness into the information theory domain.

Askari et al. [5] propose a framework to compare the privacy and utility of different sanitization techniques. They do this by treating a privacy mechanisms as a noisy channel. For a certain input distribution $P(a)$, the channel is assigned a channel matrix of probabilities $P(o|a)$. It encodes the probability of output o given input a. They also model the adversary's and user's background knowledge as a priori probability distributions. The reason for the latter is that utility also depends on how much knowledge a user has about the data.

Using this model, privacy is measured as the error probability of a guess by an adversary. This is computed by using the maximum a posteriori probability method. The higher the error, the less success an adversary has in guessing the correct input. When it comes to utility, the authors believe that this depends on the specific application and on the dataset. Still, they provide a general metric which can be later customized towards a specific goal. They define the utility of a single attribute as the difference between the maximum entropy of that attribute (log k) and the actual entropy. If there is a correlation between an attribute and several others, the utility gain can also be expressed as mutual information.

Though the framework is interesting, one of its challenges is to compute the channel matrix, which explodes combinatorially for each attribute added.

5.5 Data mining measures

Sramka et al. [37] propose a data mining approach to comparing different anonymization techniques based on the privacy and utility levels they provide. They define a utility metric

which can be used for legitimate purposes (good utility) or for privacy purposes (bad utility). So the difference between and adversary and a user is simply in the goal and idea of the mining task. In essence, the idea behind the framework is to run many different data mining algorithms on the sanitized data and computing the good and bad utility for each of them. In their experiments they show that ε-differential privacy has a smaller good utility loss than k-anonymity, but this is circumstantial (only two attributes were numerical and could be perturbed).

To measure utility the following formula is used:

$$U_{good}^{(san)}(DB, S, M^i) = \sum_{x \in DB} w(x) E_i(M^i(S(DB), S(x)), \bar{x}_i)$$

where E is the error implication function. In this case, the error between the mined estimation and the true value of x. For bad utility the formula is similar and can be seen as computed with a different miner, field and error and interest functions. The authors test this by using the WEKA package [44] with the default settings. This also shows how the techniques perform against "script kiddies" - people who use already existing tools for their goals.

Chapter 6

Discussion and future work

In this chapter we will be looking at the interviews in more detail, analysing some challenges and open directions for future work

6.1 The Interviews

From different interviews we have noticed that decision making is not always easy when dealing with Open Data. Even if there are rules in place, such as the "open tenzij"[1] policy, which states that a dataset can be published unless it has sensitive information (e.g. zipcode and address of an individual, trade secrets, information that may create threats to national security, information that might create unfair competitive advantage), it is not always easy to decide whether or not a data set should be published. Rijkswaterstaat, for example, is not yet sure how to publish their NAP bolt data, Kadaster might need to decide how to open up parcel information without bringing all the individuals (owners, buyers) at risk. The Amsterdam Economic Board has mostly information that is not related to individuals (street light conditions, trash collection points, traffic etc), but which may still be used in unforeseen ways to harm others. For example, an app[2] has been developed which shows which houses can be more easily broken into). A legal debate could now follow whether such data breaks the privacy law[3]. The last example is beyond privacy issues but shows that there can always be hidden threats in the data, especially through combinations of the data sets. This has lead to the idea that we need to analyze how **sensitive** the data is.

The departments usually do not have the necessary expertise to sanitize and evaluate datasets. For example, Rijkswaterstaat and Kadaster currently have no process in place to actually sanitize and remove and privacy risks from the data set. Privacy is kept safe(to some extent) based on secrecy (not publishing). Even if the *publish unless* rule seems to be black-and-white, in reality it is not easy to choose. If too strict, one may prevent useful information to reach the public; if too tolerant, one may release sensitive data. The problem is that we do not know how the law may change. From the Kadaster interview it was

[1] tr. from dutch: open unless

[2] Makkie Klauwe: `http://www.appsforamsterdam.nl/en/apps/makkie-klauwe`

[3] `http://wetten.overheid.nl/BWBR0005252/geldigheidsdatum_14-05-2013`

suggested that even if one does not publish now, he might be forced to later. Furthermore, such departments, who do not have the expertise to publish sensitive data, rely on other institutions (CBS, O&S) to handle data which may present risks.

When reasoning about the data, whether it's safe or not and how it should be published, experience plays an important role and so does intuition. Some departments have more experience with publishing (CBS), others have less. For example, CBS has statisticians which can make assumptions about which attributes in a table are more important for research - this can impact utility loss when anonymizing certain attributes (see Section 6.2).

6.2 Challenge in choosing the QID

One of the many challenges that a data publisher has to face is the choice of the Quasi IDentifier(QID). It may not always be straightforward which attributes should be part of the QID and which not. If one selects too few attributes, then one risks privacy breaches. If on the other hand one selects too many attributes, then "the curse of dimensionality" [2] will manifest itself. This states that as the QID gets bigger (more attributes), so will the sparsity of the data. The sparser the data is, the more one has to generalize attribute values in order to achieve k-anonymity, for example, which leads to less data utility. Sparsity simply makes data grouping harder. Take for example three zip codes: 135678, 112345, 170000. In order to group them under the same q^*-block , one might have to generalize the zip code to 1***** (the interval [100000,199999]), which has a big impact on utility.

In practice (see CBS interview), the choice is mostly done based on experience and a gut feeling. This puts non-experts at a disadvantage when having to make decisions about the data.

6.3 Future work

There are many challenges ahead privacy preserving data publishing. We started off by using Dalenius' definition of Privacy Preserving[11]. Then Dwork proved that such a privacy definition is impossible[14] since it assumed an adversary with unlimited resources. To make the definition viable, it has been revised with the assumption that an adversary only has access to limited resources. With the Open Data movement we need to ask ourselves whether or not we are reverting back to the impossibility proof. Open Data is potentially making more and more resources available to such adversaries. What we need is a decision support tool to help us quantify the risks and understand how hard it is for an adversary to create a privacy breach.

Privacy Preserving Data Publishing is the core focus of this thesis. It looks at how to publish data without knowing beforehand what the data will be used for. There are many data types (transactional, relational, location, social, graphs) each requiring different techniques to achieve anonymity. This survey is mostly focused on relational data. We have first presented in this report the most popular techniques in the domain of relational data (k-anonymity, ℓ-diversity, t-closeness, ε-differential privacy) and their variations. These can be seen as the building blocks from which many other techniques have been inspired. They

have been structured based on the type of attack they were trying to prevent:linkage and probabilistic attacks. We continued with ways to measure privacy and utility levels. This lead the research into a new direction: that of existing frameworks to compare anonymization techniques.

Besides helping understand the field, one important contribution of this survey are the interviews with representatives of different public institutions in The Netherlands. These interviews support the thought that there is a need for support to close or shrink the knowledge gap for publishing data sets under various privacy, sensitivity or security constraints (legal and technical).

6.3.1 Research questions

It is clear that a framework or a methodology is required to help the non-expert data publisher release proper information to the public. In order to get started on finding an answer for the main question "What are the necessary steps in order to achieve privacy preserving data publishing, in the context of Open Data?", we first have to split it up into smaller pieces.

RQ 1 *Where does privacy preserving data publishing fit in the context of Open Data?*

We have already mentioned in the introduction that data privacy is no longer the only big threat. The focus of the literature has been on privacy, but we feel that a broader concept should also be considered, that of data sensitivity. This includes threats to states, companies and any other non-individual entities. We will try to create a picture of the current context surrounding Open Data and data sensitivity.

RQ 2 *How are decisions taken when publishing privacy preserving Open Data?*

From the interviews we can conclude that there is a knowledge gap between data publishing in theory and practice. This can also be seen when looking at guidelines on publishing data. Theory concerns itself with how to clean the data. In practice, people look more at what data to publish, the rules and regulations they have to follow, and how to make the data more accessible. There are few institutions who know how to do both: to manage the data and to sanitize the data. We will be giving an overview of the current dutch Open Data publishing guidelines, guidelines which unfortunately do not include anything about data sanitization. This gap needs to be filled in by a decision support tool which helps the data publisher understand the *hidden* risks associated with the data, risks which the guidelines cannot uncover.

RQ 3 *How to anonymize the data?*

This is the main question of the master thesis. The coming months we will be looking into the following steps. First, we will start off by designing a framework which can analyze a data set for privacy risks, run different algorithms on the data to anonymize it, run metrics on the anonymized data to show the levels for privacy and utility. These results will then be presented to the user who then selects what he considers to be the proper anonymization.

After the framework has been designed, we will continue with implementing (n,t)-closeness[29]. This hopefully overcomes the limitations of t-closeness. To test this, we will be comparing (n,t)-closeness with t-closeness[28] and other algorithms (yet to be determined). We will then continue with implementing the rest of the framework.

The data we will be using will consist of synthetic data. The reason is twofold. First, quality synthetic data, that contains personal information or other privacy sensitive information, is already freely available. Second, from the interview with CBS, obtaining real data is a cumbersome process, requiring many approvals which eventually might not even yield raw data (usually provided in an anonymized format - $k \geq 100$). The data needs to consist of different data sets which can be linked to each other. This way we can test how the privacy levels degrade when background information is used. For this we will be using different existing data set attack tools (yet to be determined).

Different privacy/utility metrics provide different values. These need to be interpreted so that the data publisher understands what his privacy and utility guarantees are. Below several sub-questions are formulated which encapsulate the ideas presented above.

RQ 3A *Which algorithms should be considered as candidates for anonymization for which type of data, with respect to applicability in practice?*

There are a lot of techniques for each data type (relational, transactional, location etc). As seen throughout the report, there are some techniques that are too theoretical and are hard to apply in practice. We will use the knowledge in this survey to point out those methods which could be used in practice. Unfortunately, this thesis will not be able to handle all the data categories, due to time constraints, and will focus on relational data.

RQ 3B *Which frameworks exist, that can be used in practice, that can compare different anonymization techniques, with respect to data privacy and utility?*

There are some frameworks based on information theory and data mining techniques which help the user decide on which algorithm to use for the anonymization process. There are also the R-U maps, which are simpler, but require a good choice for the metrics used to measure privacy and utility. We will show which frameworks can be used in practice by people with limited expertise.

RQ 3C *How to interpret the measured values for privacy and utility and what guarantees do these values provide?*

Different frameworks give different values for utility and privacy, sometimes expressed in a different way. We will be looking into different types of measurements and analyse what the values actually say about the sanitized data set. To this end, we will be conducting several experiments on synthetic data, where we will compare different anonymization mechanisms.

RQ 3D *How does privacy / utility change when the data set is combined with external sources?*

This question is of real interest with the upcoming Open Data movement. We will try to breach the privacy of an anonymized data set using external information. Then we will try to anonymize the data set taking these external sources into consideration and see whether or not we get better privacy / utility results. It might be possible that due to time constraints, this question will not be investigated.

Acronyms

EC Equivalence Class

EMD Earth Mover Distance

PPDP Privacy Preserving Data Publishing

QID Quasi-IDentifier

Bibliography

[1] The dutch open government draft action plan. http://www.opengovpartnership.org/sites/www.opengovpartnership.org/files/country_action_plans/Draft%20Action%20Plan%20The%20Netherlands_0.pdf. Online; visited May 2013.

[2] *On k-Anonymity and the Curse of Dimensionality*, 2005.

[3] Dakshi Agrawal and Charu C Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 247–255. ACM, 2001.

[4] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. *ACM Sigmod Record*, 29(2):439–450, 2000.

[5] Mina Askari, Reihaneh Safavi-Naini, and Ken Barker. An information theoretic privacy and utility measure for data sanitization mechanisms. In Elisa Bertino and Ravi S. Sandhu, editors, *CODASPY*, pages 283–294. ACM, 2012.

[6] Michele Bezzi. An information theoretic approach for privacy metrics. *Transactions on Data Privacy*, 3(3):199–215, 2010.

[7] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[8] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In Chen Li, editor, *PODS*, pages 128–138. ACM, 2005.

[9] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In Cynthia Dwork, editor, *STOC*, pages 609–618. ACM, 2008.

[10] Shuchi Chawla, Cynthia Dwork, Frank McSherry, Adam Smith, and Hoeteck Wee. Toward privacy in public databases. In Joe Kilian, editor, *TCC*, volume 3378 of *Lecture Notes in Computer Science*, pages 363–385. Springer, 2005.

[11] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15(429-444):2–1, 1977.

[12] George T. Duncan, Sallie A. Keller-mcnulty, and S. Lynne Stokes. Disclosure risk vs. data utility: The r-u confidentiality map. Technical report, Chance, 2001.

[13] George T Duncan and Sumitra Mukherjee. Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise. *Journal of the American Statistical Association*, 95(451):720–729, 2000.

[14] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin Heidelberg, 2006.

[15] Cynthia Dwork. Ask a better question, get a better answer a new approach to private data analysis. In *In Proc. ICDT Int. Conf. Database Theory*, pages 18–27, 2007.

[16] Cynthia Dwork. Differential privacy: A survey of results. *Theory and Applications of Models of Computation*, pages 1–19, 2008.

[17] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222. ACM, 2003.

[18] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, June 2010.

[19] Benjamin C. M. Fung, Ke Wang, and Philip S. Yu. Top-down specialization for information and privacy preservation. In Karl Aberer, Michael J. Franklin, and Shojiro Nishio, editors, *ICDE*, pages 205–216. IEEE Computer Society, 2005.

[20] Benjamin C. M. Fung, Ke Wang, and Philip S. Yu. Anonymizing classification data for privacy preservation. *IEEE Trans. Knowl. Data Eng.*, 19(5):711–725, 2007.

[21] Aris Gkoulalas-Divanis and Grigorios Loukides. Pcta: privacy-constrained clustering-based transaction data anonymization. In *Proceedings of the 4th International Workshop on Privacy and Anonymity in the Information Society*, PAIS '11, pages 5:1–5:10, New York, NY, USA, 2011. ACM.

[22] Vijay S. Iyengar. Transforming data to satisfy privacy constraints. In *KDD*, pages 279–288. ACM, 2002.

[23] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In Timos K. Sellis, Rene J. Miller, Anastasios Kementsietsidis, and Yannis Velegrakis, editors, *SIGMOD Conference*, pages 193–204. ACM, 2011.

[24] Nellie Kroes. The big data revolution. `http://europa.eu/rapid/press-release_SPEECH-13-261_en.htm#PR_metaPressRelease_bottom`, March 2013. Online; visited May 2013; EIT Foundation Annual Innovation Forum /Brussels.

[25] Diane Lambert. Measures of disclosure risk and harm. *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-*, 9:313–313, 1993.

[26] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In Ling Liu, Andreas Reuter, Kyu-Young Whang, and Jianjun Zhang, editors, *ICDE*, page 25. IEEE Computer Society, 2006.

[27] Jiexing Li, Yufei Tao, and Xiaokui Xiao. Preservation of proximity privacy in publishing numerical sensitive data. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 473–486, New York, NY, USA, 2008. ACM.

[28] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In Rada Chirkova, Asuman Dogac, M. Tamer zsu, and Timos K. Sellis, editors, *ICDE*, pages 106–115. IEEE, 2007.

[29] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. Closeness: A new privacy measure for data publishing. *IEEE Trans. Knowl. Data Eng.*, 22(7):943–956, 2010.

[30] Grigorios Loukides, Aris Gkoulalas-Divanis, and Bradley Malin. Coat: Constraint-based anonymization of transactions. *Knowledge and Information Systems*, 28(2):251–282, 2011.

[31] Grigorios Loukides, Aris Gkoulalas-Divanis, and Jianhua Shao. Assessing disclosure risk and data utility trade-off in transaction data anonymization. *Int. J. Software and Informatics*, 6(3):399–417, 2012.

[32] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):146, 2007.

[33] Krishnamurty Muralidhar and Rathindra Sarathy. Does differential privacy protect terry gross' privacy? In Josep Domingo-Ferrer and Emmanouil Magkos, editors, *Privacy in Statistical Databases*, volume 6344 of *Lecture Notes in Computer Science*, pages 200–209. Springer, 2010.

[34] M.E. Nergiz, C. Clifton, and A.E. Nergiz. Multirelational k-anonymity. *Knowledge and Data Engineering, IEEE Transactions on*, 21(8):1104 –1117, aug. 2009.

[35] Mehmet Ercan Nergiz, Maurizio Atzori, and Chris Clifton. Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, SIGMOD '07, pages 665–676, New York, NY, USA, 2007. ACM.

[36] Pierangela Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.*, 13(6):1010–1027, 2001.

[37] Michal Sramka, Reihaneh Safavi-Naini, Jörg Denzinger, and Mina Askari. A practice-oriented framework for measuring privacy and utility in data sanitization systems. In *Proceedings of the 2010 EDBT/ICDT Workshops*, EDBT '10, pages 27:1–27:10, New York, NY, USA, 2010. ACM.

[38] Daniel Stauffacher, Sanjana Hattotuwa, and Barbara Weekes. `http://ict4peace.org/wp-content/uploads/2012/03/The-potential-and-challenges-of-open-data-for-crisis-information-management-and-aid-efficiency.pdf`, 2012.

[39] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

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

[41] Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. Privacy-preserving anonymization of set-valued data. *Proc. VLDB Endow.*, 1(1):115–125, August 2008.

[42] Suresh Venkatasubramanian. Measures of anonymity. In Charu C. Aggarwal and Philip S. Yu, editors, *Privacy-Preserving Data Mining*, volume 34 of *Advances in Database Systems*, pages 81–103. Springer, 2008.

[43] Ke Wang and Benjamin C. M. Fung. Anonymizing sequential releases. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 414–423, New York, NY, USA, 2006. ACM.

[44] I.H. Witten, E. Frank, and M.A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.

[45] Weiyi Xia, Raymond Heatherly, Xiaofeng Ding, Jiuyong Li, and Bradley Malin. Efficient discovery of de-identification policy options through a risk-utility frontier. In Elisa Bertino, Ravi S. Sandhu, Lujo Bauer, and Jaehong Park, editors, *CODASPY*, pages 59–70. ACM, 2013.

[46] Xiaokui Xiao and Yufei Tao. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, SIGMOD '06, pages 229–240, New York, NY, USA, 2006. ACM.

Appendix A

Interview transcripts

This chapter summarizes the discussions carried out with different people at public institutions in the Netherlands.

A.1 Rijkswaterstaat (RWS)

This interview has been carried out with Aart van Sloten at RWS. Rijkswaterstaat is the institution that handles the practical execution of public works and water management. They have a lot of data in their databases which is either sensitive or non-sensitive. Very few datasets are known to be somewhere in the middle.

The approach they use to publish data is very simple. If there is the slightest proof that a data set contains some sensitive data, then do not publish (e.g. the data contains addresses of individuals). Examples of types of data that do not get publish include company and competition data, country security, information about prisons, information on Defense, full building layouts, technical details of tunnels, environment information (to prevent rare specie hunting), information on water depth (can be misused by fishermen). What they usually publish and are not worried about is general geographical data (e.g. roads / waterways). One of the gray area data sets is about the NAP bolts. These are copper bolts throughout the Netherlands which have a known (pre-measured) altitude. They are not sure whether to release the exact positions of these bolts due to several simple reasons: copper can be stolen for money, some of the bolt lie on private properties and are recorded as such (address, names).

Due to the nature of the institution, they do not have a lot of data about people. When information needs to be published about such data sets (e.g. information on road accidents) they ask CBS to correlate the data with other statistics (e.g. how many people actually went to the hospital) and release general statistics about the incident (region level, big city level).

From the discussion it can be noticed that there is no true process of reasoning about privacy at RWS. As such there is also no process to anonymize the data. Privacy is protected through secrecy - non-disclosure. The need for a framework is obvious.

A.2 Kadaster

At Kadaster we had an interview with Dick Eertink. Kadaster is the institution which manages the parcels location and size within the Netherlands. Each parcel has its own number, location, size, type (building space / agricultural space). To this there are transaction documents linked which define the last sell event. So who bought it from who, for what amount or what mortgage has been used. Also, from these documents one can retrieve the name, address, postal code and city of the buyer/seller. The BSN is also stored, but that is not released to the public.

The access to the data is allowed through the online interface. One can request information about one parcel at a time. The only limitations in place are the licence agreement (use the information just for yourself) and the fact that it costs 3.50 euro per inquiry. This for example does not protect an individual if an attacker acts in a targeted manner.

They also give big datasets for different purposes, mainly internally to the government. There are strict rules on usage (such as WBP - dutch privacy law) and they try to verify that people respect these rules, but in practice its not that easy.

There are three types of data: parcel information, transaction information and owner/former owner information. Only access to the latest information is given (no history given such as past owners). Yet, with some effort, most of the history can be reconstructed. The CBS (college bescherming persoonsgegevens) is currently debating on how to handle these categories. The desire is to eventually remove the costs of inquiry and make this Open Data.

Other information they manage includes:

- national topography
- address and building/land type for that address (mostly already public)
- act as an intermediary for information about cables and underground pipes

There is currently no process in place to anonymize the data and deciding on how to publish sensitive data sets is not easy. The laws are not very specific yet regarding the types of data they handle. They expect this year (2013) new versions for the Dutch and European privacy laws.

A.3 Statistics Netherlands (CBS)

At CBS we had an interview with Peter-Paul de Wolf. The goal of CBS is to publish relevant and trustworthy statistics. They gather a lot of information from individuals and companies and thus must handle it with great care. Most of the time they deal with microdata of people and sometimes of companies. They said protecting company microdata is not possible, in general, since companies are too easy to identify in the crowd.

They took part in the development of several tools in collaboration with other statistical departments in Europe. What they mostly use for anonymizing data here in the Netherlands are the generalization and suppression techniques (present in the μ-argus tool). Other methods include PRAM (post randomization method), which they tried a few times. The

problem with this method is the sanitized data. It is very hard to use and one needs all sorts of corrections to any statistical operation performed, in order to compensate for PRAM.

Their microdata can be released in three formats.

- public use files - accessible to all, the rules for this include having no less than 200000 respondents per region, and a k parameter of 10000.
- under contract data - given for research purpose only - somewhat less anonymized (k is in the range of 100 to 1000)
- data that stays on CBS - again, for research only - even less anonymized if at all.

The third type is very interesting. Researchers either come on-site and use the dataset based on the tools available on the premises (e.g. SPSS) or they access the data remotely, but only get to see what the tool outputs on the screen. There is no dataset transferred. This is becoming more and more popular and the request of anonymized datasets is becoming less popular. To manage this, a strict screening process has been put in place. The results are inspected and one must be able to show the steps performed to achieve those results. Transparency is key.

Upon requesting some data sets, it was suggested that it would be easier to work with synthetic data. The US has a reputation of generating quality synthetic data. It would be easier since requesting data is too complex for my goal - requests need to be filed, then approved, and at most, Type 2 data would be provided, which is not very useful for our research goal.

The Anonymization process

They use the notion of key attributes (QIDs) which can be used to re-identify individuals. Three categories of attributes can be distinguished: identifiable, m ore identifiable, the most identifiable. Based on these three categories, they try to make combinations (2-3 up to 10-20 in mu-argus) that meet a certain non-uniqueness criterion (e.g. no less than 100 per combination). It falls onto them to decide which attributes fall under which category. There are some standard attributes and others are simply agreed upon within CBS - experience/gut feeling plays a big role here.

Risk decision making is based on combination frequency. Usually, the data is based on a representative population sample. Sometimes combining this with information from GBA or some other administrative institution (not always possible) is required to be able to reason about the sample. If the sample is not big enough, they use different techniques to estimate the population. Once the population is known, it can be checked whether a certain combination is frequent or rare.

During the generalisation process, choosing which column to generalize first is done based on experience - their statistics department knows which columns are more important, in general, to researchers. Even so, researchers are never happy with the data they get (the data is always anonymized towards certain purposes). To prevent privacy breaches by sequential release, they only anonymize a dataset once.

Data types, thresholds and measurements

The data does not have to be numerical, since their program only looks at frequency of combinations. Threshold value, k is usually 100. It has been determined based on experimentation with the data. 10 is too little, 1000 is too much for the second type data.

When reasoning about utility, they do this more on feeling than on measurements. It is hard to measure utility if you do not know the purpose of the data. One possible idea would be to generate several utility measurements, aimed at different tasks (data mining, query answering etc).

One topic they have interest in is if there exists a better way to determine record/table risk, other than combination frequency?

A.4 Amsterdam Economic Board (AEB)

On behalf of AEB I had an interview with Ron van der Lans en Jasper Soetendal. AEB tries to improve economic growth by bringing together people from different institutions / organisations (CEO's, managers, scientists, researchers etc).

From the discussion it was clear that the most person related data that they have is in the Dienst Basis Informatie (DBI). As in learned from previous interviews, most of the time, they rely on O&S (research and statistics department) to publish their data by means of aggregation. The aggregation levels differ from regions to city regions to neighborhood combinations. Other rules that apply are for example that there must be at least 3 to 5 people in every aggregation. The only publicly available data sets are the ones about electricity and gas. The data is aggregated on building level.

They are in the process of opening up their data, but most of the time, one can simply access this data by requesting it at the local town hall.

In determining what data has privacy issues, they rely on common sense, experience and whether or not the data is about people. Usually, the data that AEB has is not so sensitive (e.g. trashbins, lamp posts etc). There are currently about 160 datasets published[1]. In the future, they will probably have more than thousands of datasets that will be published. Some examples for which some security has been taken is data on fire alarms - the street and approximate geo-coordinate has been released. They are also looking into how data on public works should be released. It contains information (phone number, address and other information) of the person in charge of the works.

A.5 IBM Ireland

Aris Gkoulalas-Divanis is one of IBM's researchers that are currently working on privacy. Other topics covered by his research include Dublinked and mobility data. Currently, he is doing a postdoc on medical data (anonymizing medical data). For Dublinked, they try improve on how to decide on the vulnerability of a data set. Currently, this is done based on experience. They are manually inspecting data and the only automated tasks are generating

[1] `amsterdamopendata.nl`

histograms for attributes and identifying unique combinations (e.g. {sex=male,age=55} is unique). On a macro level, he is researching "knowledge hiding", which is essence is preventing people to understand patterns in data by reducing frequency of the patterns.

He has been experimenting with different kinds of data: relational, transactional, sequential data, each requiring a different approach to protect.

Regarding Open Data, it only makes the problem more complex. With the opening of all the new data sets, the data could be used in context not foreseen (combination with other data sets).

They looked at the The Health Insurance Portability and Accountability Act (HIPAA) and other similar regulations, but they only provides minimum requirements; it is not enough to actually protect the data.

Talking about anonymization techniques, we have learned that an efficient anonymization technique will only lead to less utility since it will cut corners on utility to finish faster; the focus of such algorithms is on privacy. A good anonymization algorithm seems to be Mondrian (does k-anonymity) by recursively partitioning the space. It achieves a good balance between privacy and utility.

As far as run time goes, he noted that slower algorithms take hours. From his experience, a k value of 5 is enough for medical data. Most approaches can be parallelized which reduces the overall computation time.

Anonymity levels - how they decide on parameters and data sensitivity:

- measure re-identification risk by studying which elements are unique in the dataset
- identify outliers
- reason based on this how sensitive the data is

Three types of utility measures that they use:

- IL (information loss): each generalisation increases the IL (general measure)
- based on workload (for which goal is the data anonymized): provides more utility for specific tasks
- average aggregation query - what is the error on these queries (general measure)

Selecting the QID is done based on type of data. This reflects the need for prior experience. In the case of medical data, selecting a QID is relatively easy because the data and its sensitivity is well defined. In other areas this is more difficult.

For visualizing the trade-off between risk and utility, R-U confidentiality maps(Section 5.3) can be used. The metrics to be used to measure risk and utility vary, depending on the data type and publication goal. In the case of Mondrian, one measure for risk can be for example 1/k (most risky individual).

There are also two IBM products which handle data sanitization, but in a very simple way.

- Infosphere Optim: implements masking approaches to protect sensitive data. It uses auxiliary data dictionaries (for example to replace names). In essence, it generates a new data set. Data masking is not the same as anonymization, it is simpler.
- Infosphere Guardium works with reduction. It automatically identifies data, identifies sensitive words (patient names) and simply removes them. This may turnout to decrease utility too much.