university of groningen

faculty of mathematics and natural sciences

Risk of de-anonymization of anonymous demographical and questionnaire data

Bachelor's Thesis

July 15, 2015

Student: J.J. Hoving, Y.W. Tijsma, T.V. Verbeek

Primary supervisor: prof. dr. ir. M. Aiello

Secondary supervisor: F.J. Blaauw, MSc

Abstract

Privacy is increasingly becoming an issue in everyday life. Personal information is gathered from many sources and data leaks are not out of the ordinary. People are progressively aware of which information to share and which information to keep secret.

HowNutsAreTheDutch is an online questionnaire and diary study that collects anonymous data from people who want to determine their psychological wellbeing in comparison to others. The focus of this work is to analyze basic demographic information to determine how anonymous this information really is.

An analysis tool is developed to determine the risk of de-anonymization on demographic information coupled with general population data. The research performed in this work indicates that a single piece of personal information, postal code, has a high impact on a person's anonymity. Additionally, increasing the amount of demographic information leads to a higher risk of identification. Multiple options are available to increase anonymity by utilizing de-identification techniques and smart data management.

Foreword

This thesis is written as completion to the bachelor Computing Science, at the Rijksuniversiteit Groningen. This thesis is the result of a group of three students with a strong motivation to research a topic that concerns everyone. With this study we want to address the University Medical Center of Groningen and other institutions, to provide a guide in improving questionnaires, surveys and general data management.

Research was mainly done by Tim Verbeek, with Ynte Tijsma and Jelte Hoving focusing on analysis. The results were analyzed together and overall the work was done as a team. Many thanks to our supervisor Frank Blaauw who assisted us in our research and provided us with good directions for our project. We also want to thank Matthijs Koot for answering some of our questions and providing tips.

Jelte Hoving,
Ynte Tijsma,
Tim Verbeek

Groningen, July 2015

Contents

List of Figures

List of Tables

Listings

Abbreviations

NSA	**N**ational **S**ecurity **A**gency
HND	**H**ow **N**uts are the **D**utch
US	**U**nited **S**tates
ZIP	**Z**one **I**mprovement **P**lan, (Postal Codes)
QID	**Q**uasi-**ID**entifier
CBS	**C**entraal **B**ureau voor de **S**tatistiek, (Dutch registry office)
CSV	**C**omma **S**eperated **V**alues
CoB	**C**ountry **o**f **B**irth
YoB	**Y**ear **o**f **B**irth

Chapter 1

Introduction

Privacy is increasingly becoming an issue in everyday life. Since Edward Snowden leaked classified information from the NSA, people are progressively aware of which information to share and which information to keep secret. Organizations are often also a factor in sharing information about citizens. Data from organizations or governments like research results, questionnaires and demographics could also pose a threat to privacy, if personal data could be traced back to individual people.

A large Dutch population study is currently running known as HowNutsAreTheDutch (HND). In HND people can determine how their psychological wellbeing compares to the psychological wellbeing of other people by filling in a questionnaire, or join a specialized 'diary study', in which participants can measure how much their psychological wellbeing fluctuates over time and see what causes that fluctuation. In essence, the data collected with these studies is anonymous; merely basic demographic information is collected, such as birth year, birth month, gender, (Dutch) zip code and several other demographic variables.

The present work focuses on the degree of anonymity of HND variables, to determine the degree of anonymity that HND data actually holds, and what needs to be done to prevent situations in which personal identifiable information can be extracted from the data.

1.1 Research question

Various organizations, researchers, and governments are doing research to find out what needs to be done to protect the privacy of citizens. Several techniques have been developed to reduce the risk on unique identification.

The main goal of this project is to investigate the risk of de-anonymization in questionnaire and demographic data, and to identify the most important threats to anonymity, in order to create awareness of the privacy risk and to influence data-management procedures to increase anonymity. To realize this goal, the variables used by the HND questionnaires are scrutinized to find out which variables or combination of variables can potentially be (mis)used to uniquely identify an individual.

In order to achieve this goal we have formulated a research question and

various sub-questions.

What is the risk that a person's real identity be revealed through anonymous data?

1. *Which variables are responsible for revealing a person's identity?*:

   Research is done to filter questionnaire and survey questions for variables that could potentially lead to identification of an individual.

2. *What is the factor with which anonymity is reduced for each of the variables?*:

   All potentially dangerous variables are analyzed to find out to what extent they reduce the anonymity of a person.

3. *What is the effect of variable combinations on anonymity compared to single variables?*:

   Various variable combinations are investigated. This shows us if combined factors relate to the single variables and to measure the impact of combinations on a person's anonymity

4. *What can be done to reduce the risk of unique identification through anonymous data?*:

   It is important to find a balance between safeguarding anonymity and gathering sufficient data. Several options are explored to find this balance. An analytical tool is developed to search through available data and enable us to answer the questions regarding this topic.

In addition to our research question, a data-management plan is drafted based on the gathered results. The advice entails which data should be best kept separate from the other data and provides guidance to researchers on which variables to store where in order to keep anonymity of data intact.

1.2 Thesis organization

The remainder of the document is organized in the following way:

In Chapter 2 we present the related work in the field of anonymity and privacy, introduce the concept of k-anonymity to quantify anonymity and explain two common approaches to data management.

Chapter 3 focuses on the available data and the search for a suitable database. The HND questionnaires are thoroughly analyzed to select viable variables while the structure and contents of the chosen database is discussed.

Chapter 4 is dedicated to the design and architecture of the analytical tool we developed and the choices made during the design stage.

Key implementation elements of the analysis tool are at the core of Chapter 5, explaining their function. Some problems encountered during development are mentioned here along with their solutions.

Chapter 6 contains the evaluation of the results generated by the tool. We make a comparison of all single variables and introduce an identification factor to clearly differentiate between these variables. Investigation of variable combinations follow shortly after.

Chapter 7 presents the conclusions for this work. We propose some options for future work in chapter 8. Additionally, an advice on data management for the UMCG is added as an appendix based on the results discussed in chapters 6 and 7.

Chapter 2

Related Work

Research in the area of anonymity is ongoing and earlier work in this field is relevant to the goals set by this project. In this chapter, we discuss other work concerning analysis of anonymity and introduce the concepts k-anonymity and quasi-identifiers. In addition, we discuss de-identification of demographic information and its effect on the quality of a dataset, research and enquery results or similar areas.

2.1 Background

There have been numerous detailed articles covering the possible privacy threat posed by surveys and questionnaires. The foundation of our research has been laid down by Sweeney roughly 15 years ago (Sweeney, 2000). In her paper she tries to 'uniquely identify people using simple demographics' and found that, using 1990 United States (US) Census summary data, 'combinations of few characteristics often combine in populations to uniquely or nearly uniquely identify some individuals' (Sweeney, 2002). The experiments performed by Sweeney showed that 18% to 87% of the US population had identifiers that could uniquely identify them based on a combination of demographic variables. A combination in which 5-digit postal codes, gender and full date of birth proved to be the most revealing. A paper by Golle revisited the work done by Sweeney (Golle, 2006). Although Golle used more recent data, his results were similar to those of Sweeney. From these papers we can assume that disclosing demographic data is a possible threat to anonymity.

The works by Sweeney and Golle both discuss data from the US and are thus slightly different from our own research, as the Netherlands has a different way of using postal codes and is much smaller in size than the US. In light of this, the work done by Koot is very significant to our project as he conducted similar research on Dutch demographical data, providing a benchmark for our own research. Koot's dissertation describes research on the amount of (empirical) privacy and large differences therein, and threats to anonymity, depending on where a person lives and other demographical data (Koot, 2012).

There are a number of similarities that all papers share. One of the similarities is that all authors stress the danger of different databases being coupled. Certain demographic variables that do not occur very often have a much higher

chance of being coupled with other data in different databases. This could lead, for example, to a couple of databases being quite 'anonymous' on their own, but very revealing when combined with other sources. An example of a combination is given by Sweeney, where she coupled medical data with public voter list data, demonstrating the threat posed by re-identification on attributes shared over multiple databases, as shown in Figure 2.1 (Sweeney, 2000) where two datasets share the same attributes. These attributes link to additional information that can re-identify a person.

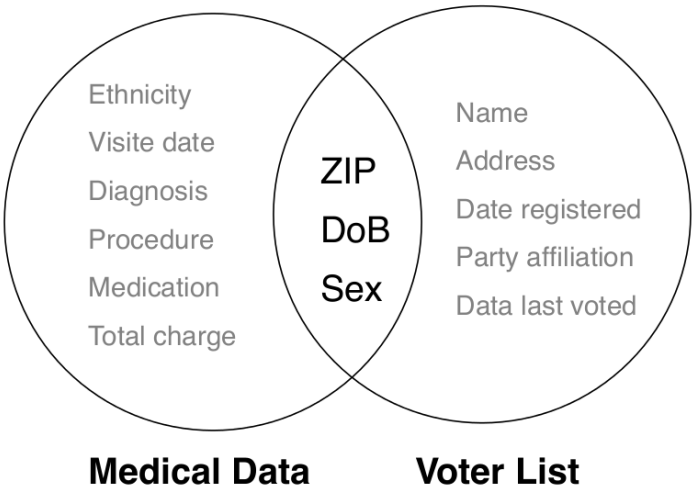

Figure 2.1: Database cross-reference

2.2 K-anonymity and quasi-identifiers

When discussing attributes and the risk of identification, it is useful to have a few terms at the ready for quantifying anonymity. The current research and the research done by others focus on data being perceived as anonymous, meaning it does not include direct identifiers such as a person's name or address. What is left are identifiers which, on their own, do not identify a person completely. We call these variables quasi-identifiers (or QID's) (Dalenius, 1986). To quantify the amount of anonymity an individual can expect, we introduce k-anonymity by Sweeney (Sweeney, 2002). Sweeney's interest is in re-identifiability of people based on their entries in databases (including linking between different databases).

A database provides k-anonymity protection if the information for each person contained within cannot be distinguished from at least $k-1$ other individuals who appear in the database (Koot, 2012; Sweeney, 2002). The protection provided by k-anonymity is meant for more than just indicating how many people are in a group sharing the same QID (or combination thereof), it is also a system model designed to lower the risk of re-identification. A simple example would be to have only five people sharing one postal code. This would mean

a k-value of five for those individuals, if there is no way of determining a more precise result. If data of a survey or enquery needs to be released to the public it can be modified in such a way that it guarantees (up to a certain amount) a k-anonymity for every individual contained in the release.

2.3 Prosecutor vs Journalist

The general idea of using the notion of k-anonymity is to reduce the risk of an anonymous individual being re-identified. There are two frequently encountered scenarios in the field of data management: the prosecutor and the journalist scenario. In the prosecutor scenario an 'attacker' will try to find a specific individual in a (anonymized) database. The other scenario being the journalist-scenario, where an 'attacker' does not intend to find a specific individual, but rather any individual. In this case the intruder wishes to re-identify a single individual to discredit the organization disclosing the data (Emam & Dankar, 2008). The 'attacker' in these scenarios can be classified as someone who has unlegitimate access to the data or tries to match data with his own data entry or database.

The amount of effort in anonymizing data also corresponds with a certain amount of data loss. When you de-identify entries in a database or survey to reduce the risk on unique identification, you will distort the data up to a certain point. An example of such dataloss can be shown with age, instead of storing the actual age of a person, one could also choose to group ages and list their information under that group, increasing the anonymity of the affected ages. The increase of anonymity comes at the cost of reduced accuracy in the dataset. Therefore, a greater anonymity results in a greater distortion. The important detail here is that de-identification protection has to be executed very carefully depending on the goals of the data holder (the person or organization in charge of, or responsible for the data) to minimize any loss of data or accuracy where it is not necessary. Numerous additions to the k-anonymity model have been suggested over time to decrease data loss as efficiently as possible (l-Diversity, t-closeness, m-invariance, etc) (Koot, 2012). Although these are relevant concepts for this research and because of these additions being very specific, we consider them out of scope.

Chapter 3

Data and Analysis

To effectively research the HND data, we analyze the HND questionnaires, the datasets we retrieved from the Dutch registry office (CBS), and other possible options. We have a look at the variables we want to scrutinize and what is actually possible. The contents and format of the CBS datasets are discussed as well as the accuracy and uniformity of the data.

3.1 HowNutsAreTheDutch

The aim of our research is to scrutinize the variables which people are asked to fill in the questionnaires of HND. We received an overview from our source and a page with generic QID's [1]. We then proceed to make a list of variables that we could potentially use for our research as shown in Table 3.1.

Table 3.1: List of extracted variables

age	relationship status	children
level of education	working status	postal code
country of birth	country of birth (father)	country of birth (mother)
household status	average income household	household size
household contents	number of brothers and/or sisters	twins+ or not
religion	left or right handed	hobby
weight	height	year of birth
gender		

On a side note, it is important to realize that a dataholder might have more personal information than he or she might realize. A great example is a question asking for one's email address or when a photocopy of a passport or ID card is required. The email address is usually asked for outside of an actual questionnaire, but can still be considered part of it. A photograph of a person could lead to gender, age or ethnicity being guessed even though it is not a part of a questionnaire or survey.

Although we do have the QID's that appear within the questionnaire, we do not have access to the actual data and values therein for privacy reasons. In

[1] http://www.hoegekis.nl/

order to determine the threat some of these QID's pose to a person's anonymity, we need a source that can provide us with enough actual data and sharing similar QID's to make an educated guess on their impact. A variety of datasets are compatible for this purpose. The most important criteria for us are ease of access, cost, compatibility, accuracy, and national coverage. We have looked into several sources of data and compared these to each other.

3.2 CBS

We considered a number of sources for the kind of data we need. We had the choice of either going for public or private data. Both have their advantages and disadvantages. A list of options:

- Municipal data: This is data from municipalities which only contains information about the population of that specific municipality. It is not very easy to access and will take a lot of time to receive. It will probably require a long procedure and obviously does not include national coverage. It is unlikely to be de-identified, but hard to acquire.
- Dutch registry office data (CBS): This data is publicly available in different formats, de-identified to a large extent and of considerable size. It does have national coverage.
- UMCG medical data: This type of data is extremely hard to acquire and it will take a long procedure to actually acquire it. It is very privacy sensitive and unlikely to have national coverage.
- Third party data: This data is probably very accurate but it is not free of charge.

After researching these different datasets it became clear that the CBS data would be the best option. We need a dataset free of charge and the CBS data allows for a high degree of freedom and instant access in a format of our own choosing. The national coverage is also very helpful for our research and it is similar to what other researchers have done before, allowing a measure of comparison later on. The major disadvantage is, of course, the data loss we suffer due to the de-identification procedure enacted upon public data by the CBS.

CBS (or Centraal Bureau voor de Statistiek) data is available on their website and in different file formats [2]. The CBS offers CSV, Excel, SPSS or HTML, and opted for CSV. CSV is generally preferred in this type of research and widely used for reading and writing data. Another reason for choosing CSV was familiarity and personal preference.

The (CBS) Statline system keeps track of records for multiple years. We decided to go for the most recent version available, but also the version which allowed us to cover most of our variables. The year 2013 turned out to be the most recent data we could get that got us the best variable coverage.

The CBS data we retrieved from the Statline website is quite comprehensive and benefits from an explanation in more detail. We show some of the contents, structure, and general aspects of the CSV files based on a few examples.

[2] http://statline.cbs.nl/

3.2.1 Dataset contents

From the seven datasets we use, the smallest dataset contains around 8,000 entries. The largest datasets contains just under 320,000 entries. The minimum number of QID's used in the datasets is two with a maximum of four QID's.

3.2.2 Dataset structure

In order to explain the structure of our datasets figure 3.1 and figure 3.2 show two datasets. We chose these two datasets, because they illustrate the two different structures a dataset can have. On a sidenote, the datasets shown below are not complete and they contain a lot more entries.

Huishoudens; personen naar geslacht, leeftijd en regio, 1 januari

			Perioden	2013	2013	2013	2013	2013
			Geslacht	Mannen	Mannen	Mannen	Mannen	Mannen
Regio's	Onderwerpen_1	Onderwerpen_2	Leeftijd	0 tot 5 jaar	5 tot 10 jaar	10 tot 15 jaar	15 tot 20 jaar	20 tot 25 jaar
Aa en Hunze	Personen in particuliere huishoudens	Thuiswonend kind	aantal	518	705	802	747	426
Aa en Hunze	Personen in particuliere huishoudens	Alleenstaand	aantal				15	57
Aa en Hunze	Personen in particuliere huishoudens	Partner in niet-gehuwd paar zonder ki...	aantal				1	47
Aa en Hunze	Personen in particuliere huishoudens	Partner in gehuwd paar zonder kinderen	aantal				0	3
Aa en Hunze	Personen in particuliere huishoudens	Partner in niet-gehuwd paar met kinderen	aantal				0	6
Aa en Hunze	Personen in particuliere huishoudens	Partner in gehuwd paar met kinderen	aantal				0	1
Aa en Hunze	Personen in particuliere huishoudens	Ouder in eenouderhuishouden	aantal				0	0
Aalburg	Personen in particuliere huishoudens	Thuiswonend kind	aantal	453	422	504	450	338
Aalburg	Personen in particuliere huishoudens	Alleenstaand	aantal				2	16
Aalburg	Personen in particuliere huishoudens	Partner in niet-gehuwd paar zonder ki...	aantal				0	15
Aalburg	Personen in particuliere huishoudens	Partner in gehuwd paar zonder kinderen	aantal				0	17
Aalburg	Personen in particuliere huishoudens	Partner in niet-gehuwd paar met kinderen	aantal				0	2
Aalburg	Personen in particuliere huishoudens	Partner in gehuwd paar met kinderen	aantal				0	4

Figure 3.1: Dataset example 1, household, gender, age and region

Bevolking; geslacht, leeftijd, burgerlijke staat en regio, 1 januari

		Onderwerpen	Bevolking naar geslacht	Bevolking naar geslacht
		Onderwerpen	Mannen	Vrouwen
Regio's	Perioden	Leeftijd	aantal	aantal
Aa en Hunze	2013	0 jaar	96	75
Aa en Hunze	2013	1 jaar	94	82
Aa en Hunze	2013	2 jaar	104	106
Aa en Hunze	2013	3 jaar	118	117
Aa en Hunze	2013	4 jaar	115	111
Aa en Hunze	2013	5 jaar	129	137
Aa en Hunze	2013	6 jaar	130	118
Aa en Hunze	2013	7 jaar	147	133
Aa en Hunze	2013	8 jaar	136	139
Aa en Hunze	2013	9 jaar	181	144
Aa en Hunze	2013	10 jaar	154	173
Aa en Hunze	2013	11 jaar	178	168

Figure 3.2: Dataset example 2, gender, age, marital status and region

Our datasets are large tables with rows and columns. The majority of datasets have multiple rows and columns giving meaning to the values. Figure 3.2 has two rows giving meaning to the values. The rows describe which variable is listed and the values for that variable. In Figure 3.2 this is the variable gender, and its values are male and female. The three columns in Figure 3.2, from left to right, contain the variables region, period, and age. Not all values of those three variables are shown here. Region contains all municipalities, period only contains the value (the year) 2013, and the values of age are all individual ages between 0 - 95. In Figure 3.2, for example, we can see that there are 96 males with the age of zero living in Aa en Hunze.

After analyzing the various datasets, we conclude that we can divide each dataset into five components. To illustrate these components, we make use

of figure 3.3 and Figure 3.4. The first block, bearing the title, is highlighted with a purple rectangle. The title describes the contents of the datasets. The next component is the column, or row, outlined in green. It does not hold any information and is there to align the tables properly. The remaining three components are important. The column QID's and its values, indicated by the red rectangle, and the row QID's and its values, indicated by the blue rectangle, give meaning to the actual values stored in the datasets. This distinction is important for the design and implementation of our program. We want to use these three components when running our tool. It is important to note that the definitions of the QID's in both the rows and columns cannot be used since they do not offer a useful description for the values of the row and column variables. Figure 3.2 shows an example of this. 'Onderwerpen' is not a proper description of the QID gender. Finally, we have the actual values highlighted with a black rectangle. These values can be empty(''), could contain a dash('-') or a positive integer value. A certain combination of variables is not possible when the value is empty. If a value contains a dash, it means that the combination of column and row variables is possible, but there are no people that match the criteria. When a value contains a positive integer number, that is the number of people having the local combination of variables.

Huishoudens; personen naar geslacht, leeftijd en regio, 1 januari

			Perioden	2013	2013	2013	2013	2013
			Geslacht	Mannen	Mannen	Mannen	Mannen	Mannen
Regio's	Onderwerpen_1	Onderwerpen_2	Leeftijd	0 tot 5 jaar	5 tot 10 jaar	10 tot 15 jaar	15 tot 20 jaar	20 tot 25 jaar
Aa en Hunze	Personen in particuliere huishoudens	Thuiswonend kind	aantal	518	705	802	747	426
Aa en Hunze	Personen in particuliere huishoudens	Alleenstaand	aantal				15	57
Aa en Hunze	Personen in particuliere huishoudens	Partner in niet-gehuwd paar zonder ki...	aantal				1	47
Aa en Hunze	Personen in particuliere huishoudens	Partner in gehuwd paar zonder kinderen	aantal				0	3
Aa en Hunze	Personen in particuliere huishoudens	Partner in niet-gehuwd paar met kinderen	aantal				0	6
Aa en Hunze	Personen in particuliere huishoudens	Partner in gehuwd paar met kinderen	aantal				0	1
Aa en Hunze	Personen in particuliere huishoudens	Ouder in eenouderhuishouden	aantal				0	0
Aalburg	Personen in particuliere huishoudens	Thuiswonend kind	aantal	453	422	504	450	338
Aalburg	Personen in particuliere huishoudens	Alleenstaand	aantal				2	16
Aalburg	Personen in particuliere huishoudens	Partner in niet-gehuwd paar zonder ki...	aantal				0	15
Aalburg	Personen in particuliere huishoudens	Partner in gehuwd paar zonder kinderen	aantal				0	17
Aalburg	Personen in particuliere huishoudens	Partner in niet-gehuwd paar met kinderen	aantal				0	2
Aalburg	Personen in particuliere huishoudens	Partner in gehuwd paar met kinderen	aantal				0	4

Figure 3.3: Dataset example 1 divided in components

Bevolking; geslacht, leeftijd, burgerlijke staat en regio, 1 januari

			Onderwerpen	Bevolking naar geslacht	Bevolking naar geslacht
			Onderwerpen	Mannen	Vrouwen
Regio's	Perioden	Leeftijd		aantal	aantal
Aa en Hunze	2013	0 jaar		96	75
Aa en Hunze	2013	1 jaar		94	82
Aa en Hunze	2013	2 jaar		104	106
Aa en Hunze	2013	3 jaar		118	117
Aa en Hunze	2013	4 jaar		115	111
Aa en Hunze	2013	5 jaar		129	137
Aa en Hunze	2013	6 jaar		130	118
Aa en Hunze	2013	7 jaar		147	133
Aa en Hunze	2013	8 jaar		136	139
Aa en Hunze	2013	9 jaar		181	144
Aa en Hunze	2013	10 jaar		154	173
Aa en Hunze	2013	11 jaar		178	168

Figure 3.4: Dataset example 2 divided in components

As mentioned earlier in this chapter, the CBS data we retrieved from their Statline website suffers from data loss. This means that the data we use has been de-identified to such an extent that any operations or searches done by us will not be as accurate as they could be. This attempt to make the data more anonymous happens all the time with this kind of data and is called de-

identification of macro-data. Macro, in being general registry data about all Dutch citizens. This procedure already hints at the threat of identification and is a result of research done in this field. If the data would have been more accurate or complete, the amount of people uniquely identifiable would most likely be significantly higher.

A good example of data loss is the age variable (or QID). In many of the datasets age groups are used instead of regular age being listed. As seen in Figure 3.1, in several datasets age has been grouped into clusters of five where we lose the accuracy on specific ages. Ages of 95+ are even all grouped together into one entry. This is, ofcourse, easy to explain as there are not many people alive at that age and they are therefore easy to uniquely identify without having specific information. Other examples are postal codes (only the four numbers or multiple postal codes grouped under their respective municipalities) and household (3+ children all grouped together). It might be interesting to try and find out what method the CBS utilizes to anonymize their data, but this is currently not in our scope.

3.3 Variable selection

After having a look at the CBS data and what kind of information is available to us, we can now decide on which variables to select from the HND questionnaire. These variables are shown in Table 3.2.

Table 3.2: List of usable variables

age	relationship status	children
gender	year of birth	postal code
country of birth	country of birth (father)	country of birth (mother)
household status	household contents	household size
municipality		

This list is a subset of the list in Table 3.1. Some of the questions in the questionnaire are of no use, as they ask for trivial information not tracked by CBS data. It could be relevant to compare our results to earlier work, thus it is important to have a look at what previous research has focused on with regards to certain combinations of QID's. Especially Sweeney (or Golle) and Koot have interesting and usable combinations that we can replicate and then compare (Sweeney, 2002; Golle, 2006; Koot, 2012). Their research should give us an indication on the accuracy of our research and validity of the results (compared to US and Dutch data).

3.4 Data distribution

Before we start scanning and searching through our datasets we need to highlight the distribution of our data. The uniform distribution of the actual data, or uniformity, affects the conclusions one can make based on this data. When our datasets are uniform over certain variables, you could draw a conclusion from some examples and formulate a theory or prove a connection. In our case, our

data is usually not uniform and very unlikely to be completely uniform over any variable.

An example of a lack of uniformity is the distribution of age over the total population or the non-uniform distribution of people over all postal codes. Amsterdam, for example, has over 700,000 citizens while Leeuwarden floats around 100,000, and many towns and areas only have a fraction of that amount. Date of birth is an important variable in this context, as there have been a number of studies that investigated its distribution.

Chapter 4

Program design

Before focusing on the actual implementation of the tool, that assists us in answering our research question, we take a look at the design. To answer our research question we need to make clear which variables from the questionnaire are more dangerous or less dangerous with regards to privacy than others. In other words, which variables result in a high k-anonymity and which result in a low k-anonymity. To determine which variables are dangerous, we can take the different variables and search for the variations we can create in the CBS data, and analyze the results we get. The tool makes this search process faster and more efficient.

The tool can be divided in the following components:

1. Importing of datasets
2. Creating a Person
3. Calculating k-anonymity value for a Person
4. Generating results

In this chapter we state our major design choices for each of these components in the process.

4.1 Importing of datasets

As mentioned in Chapter 3, the data we use, is stored within several (large) CSV files. The datasets are not really set up for efficient searching. Therefore, if we want to process the datasets in a correct way, we need to store them into a coherent form. Our goal is to ignore all data that has no use for us when storing the information. To do this, we look at several ways to store data, and this left us with two options. The first option is to store the datasets in internal models, which are built using (multidimensional) lists. The second option is to store the datasets in a database. The latter we could either do manually or we could use a program, like Stattransfer, to do this for us.

We decided to go with the internal models using lists, as it is a simple and transparent format that allows us to traverse and alter the data with little effort. Furthermore, we do not have to make use of an external program in order to

store the data in a database, maintain a database, and communicate with it. Another argument to not use a database is that this data is only used by us to answer our research question.

For every dataset a separate model is made. A model contains the three components discussed in chapter 3. This means, when searching, we have several internal models to go through. This is done every time the analysis tool is executed. We could have stored our internal models, so it would not need to perform this step every time. This delay is, however, insignificant to us. If we were to use a database, we would only have to do this once.

While it seems to be rather straightforward to implement this way of storing the data, it is more problematic than it seems at first glance. The data from the CBS is fairly complex and there are a lot of inconsistencies between the different datasets. This was highlighted in Chapter 3.

4.2 Creating a Person

Since we want to calculate someone's k-anonymity given some characteristics of that person, we need to store these characteristics. To store the characteristics of an actual person, we created a *Person* class. This Person class has various attributes that store the QID's we analyze. Those attributes are classes which we call *Identifier* classes.

One of the most important aspects of the Person class is that it should support the so-called 'empty' QID's. If no gender is specified, the 'gender' attribute will not contain a gender. We call this an 'empty' variable. This is necessary, since we need to determine the risk that every QID has on the anonymity of a person.

4.3 Calculating k-anonymity value for a Person

To visualize the difference between ages using some characteristics, we iterate over all available ages ($0-95$) for every search. This means that, for every search, we generate just under 96 Person objects. All these objects have their attributes provided by the user and an age in the range of $0-95$. The characteristics stored in a Person object are matched against the QID's in the internal models of the tool and this provides us with k-anonymity values.

In order to match the values of QID's found in the datasets and the QID's of our Person object, we created a mapping. This mapping has a string as key and an Identifier instance. For every QID value we encounter, we can look up the Identifier instance we need to compare the string to.

During the search process, for every internal model, we iterate over all available variables. After we have traversed all values of one variable, we delete the rows or columns that do not match the characteristics of a Person object, and continue with the next variable. We are left with values that apply to the Person object when every QID in the internal model has been processed. This can be a single value when the Person object has no 'empty' QID's for the QID's in the internal model. It is also possible to be left with multiple values as some characteristics might be unknown. For example, if gender is unknown, all values for men and women remain. When we find more than one value we take the sum

of all these values. After this has been done for every internal model, we are left with seven k-anonymity values. The minimum of those seven values is the final k-anonymity, which is equal to $\min(k_1, ..., k_7)$ where $k_x = \sum(value_1, ..., value_n)$ where $x \in datasets \wedge value$ matches the characteristics.

4.4 Generating Output

After we are done with searching, we have a k-anonymity for every age between 0 and 95. With these values we create a plot, which is shown in the user interface, and a spreadsheet. The spreadsheet contains the k-anonymity value for every age. This plot shows the different ages on the x-axis and the k-anonymity on the y-axis. An example is shown in Figure 4.1. Figure 4.1 shows the k-anonymity for children living at home for every age.

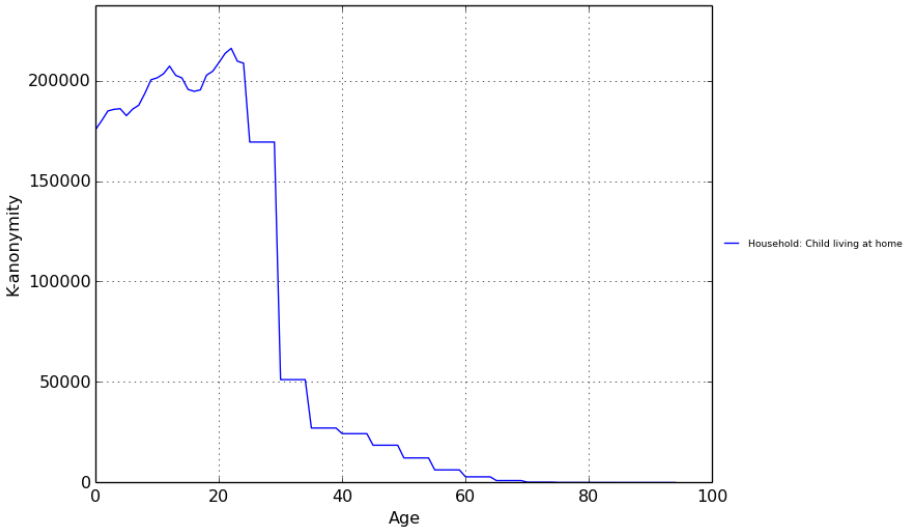

Figure 4.1: A generated graph

We have considered using different sorts of plots, but we decided to use a simple graph. A simple graph gives us an indication of our results. When we want to do more precise calculations we use the generated spreadsheet.

Chapter 5

Implementation

Our aim for the analysis tool is to extract data from the datasets discussed in chapter 3. With this data, the analysis tool can calculate the k-anonymity for people with specific identifiers and generate relevant graphs and tables. These graphs and tables can be used to analyze the identifiers. In chapter 4 we discussed our design decisions and this chapter covers the interesting parts of the implementation.

5.1 Language

The first choice concerns the programming language. Important properties of the language have to be ease of use, support for drawing graphs and being able to work with the datasets from the CBS. Furthermore, we prefer to use a language that we already have some experience with. With these reasons in mind there are two obvious candidates, namely Java and Python. Both languages fulfill the criteria that we are looking for. Additionally, both languages are very popular and widely used, resulting in extensive online support when necessary.

We decided on Python as our programming language. There is a wide range of useful libraries available but one important aspect of Python made the difference: lists. Lists, also known as arrays in other languages, are one of the compound data types that Python supports. Lists can be easily be indexed, sliced and manipulated with built-in functions. We feel that lists are easier to use than arrays in Java. Furthermore constructing a simple interface seemed more straightforward to us, because we had some previous experience with this.

5.2 Program

We developed our data analysis tool using the Model-view-controller pattern. Using this pattern we describe our tool.

5.2.1 Model

The model part consists of a *class Model*, a *class CSVMapping*, a *class TableMapping*, a *class Person*, and seven other classes describing a person's characteristics. We refer to these seven classes as *Identifier* classes.

The class Model is, given a filename of a dataset, able to read the dataset and process the dataset resulting in a model. Furthermore, the class has methods to calculate a k-anonymity for a given Person instance. In order to do calculate the k-anonymity, the class Model makes use of the class CSVMapping and class TableMapping. Among other things the class CSVMapping contains an associative array that maps values of QID's that can be found in our datasets to Identifiers instances. Figure 5.1 illustrates this associative array.

```
def fillMap(self):
        self.map["\d+ jaar"] = Age()
        self.map["\d+ tot \d+ jaar"] = Age()
        self.map["\d\d\d\d, "] = Location()
        self.map["Mannen"] = Gender()
        self.map["Vrouwen"] = Gender()
        self.map["Totaal buitenland"] = Origin()
...
```

Listing 5.1: Mapping of QID's to Identifier instances

The snippet of the class Person is shown below in figure 5.2. It has eight attributes. The attribute *self.result* stores the k-anonymity value and the other seven attributes are the Identifiers of that Person.

```
...

class Person():

    """Resembles a real life person. An instance of
       Person has the following attributes:
    an Age, Gender, Location, Origin, Marital and a
       Household instance."""

    def __init__(self, age=None, gender=None, postalcode=
        None, alien=None, father=None, mother=None,
                    status=None, household=None):
        self.result = sys.maxsize  # K-anonymity
        self.age = Age(age)  # Integer
        self.gender = Gender(gender)  # String
        self.location = Location(postalcode)  # String
        self.origin = Origin(alien, father, mother)  #
            Boolean, Boolean, Boolean
        self.marital = Marital(status)  # String
        self.household = Household(household)  # String
...
```

Listing 5.2: The class Person

In Chapter 4 we saw that each program QID should be able to match a value of a dataset QID to its own value. When we match these values, we should also take into account that we do not always have a value for a program QID. In practice, this is a trivial problem to solve. We illustrate a matching method by

giving an example. The match function in Figure 5.3 has four cases.

- *self.age* is empty. This means no value was given by the user, we can match everything.
- *self.age* is not empty. Now we need to check whether the input parameter of the function match *self.age*. Since the CBS datasets have three different formats for age, we need to check them all.

```
class Age():

    """Resembles the age of a Person."""

    def __init__(self, age=None):
        self.age = age  # Integer

    def match(self, string):
        """Returns True if string matches the value of
            self.age else it returns False. If self.age is
             an empty
        variable it returns True."""

        if self.age is None:
            return True
        if re.match(r'\d+ jaar$', string):
            return int(string.split()[0]) == self.age
        elif re.match(r'\d+ jaar of ouder$', string):
            return int(string.split()[0]) <= self.age
        else:
            return int(string.split()[0]) <= self.age <
                int(string.split()[2])
```

Listing 5.3: The class Age

To return a k-anonymity value we need to implement the searching method of the data analysis tool. Since we need to be able to search for specific values in the dataset, we have to come up with a reasonably efficient way to traverse the sets and return the correct values. Initially, we applied a linear search function to all datasets. The linear search funtion checked every row and column to see if it corresponded with the correct attributes of the Person attribute. If a match was found the indices of these rows/columns were stored in a list. The linear search function turned out to be a very inefficient search method. The more datasets we used, the more time it took to produce results. On average it took more than half an hour to produce results for a single search query. Since we want to run a lot of search queries, we need to have a more efficient search algorithm.

To try and reduce the running time of our search function we replace the linear search process with a binary search process. Since binary search has a time complexity of O(log n) and linear search a time complexity of O(n), it should be much faster than the standard linear implementation. We decided to

use binary search, due to the large amount of different values inherent to some QID's such as region or age, and since linear search was slow to the point that it was almost unusable. Before we can get proper results using binary search, we need to sort the data in our internal model, as the QID region in our datasets is not sorted correctly. Sorting is done in O(n log n), but since it is only performed once, when the program starts, it does not significantly influence the running time. The binary search algorithm for these identifiers is relatively standard, the only difference is that, once it has found a corresponding value, it searches around this result for values that are equal to the current one. After all, a value for a QID could occur multiple times in a dataset. Figure 3.1 serves as an example, as it returns a range with indices. These indices contain the right values in respect to the person we are searching for.

5.2.2 View

The view part consists of one class that handles the UI and related actions. It handles the input of the user, displays the graphs that are generated, etcetera. A screenshot of the UI is shown in Figure 5.1.

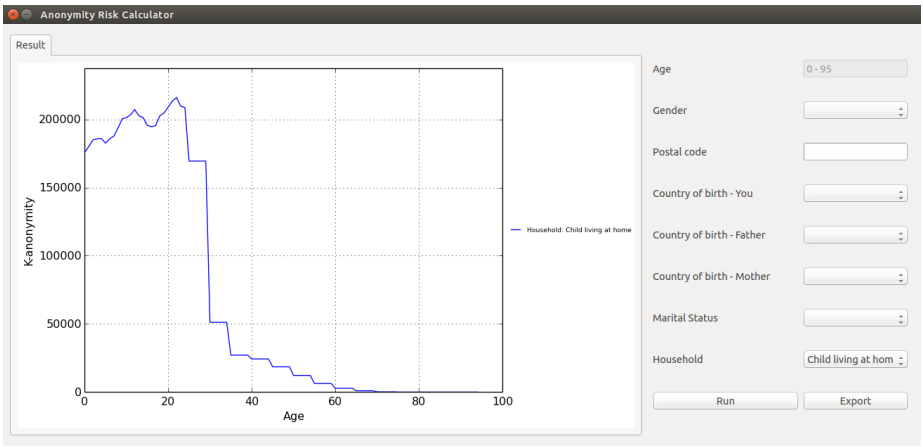

Figure 5.1: The UI of the tool

On the left-hand side of Figure 5.1 we can a generated plot given some input(QID's) from the user. In this case the given QID is household. The users input is shown on the right-hand side. The right-hand side has multipe fields with which one can give the characteristics of a Person and two buttons. These two buttons allow the user to run a search and to export both the graph and spreadsheet.

5.2.3 Controller

The controller component consists of one class that handles all actions related to the model and communication with the view.

Chapter 6

Evaluation

To answer our research question we need to evaluate the results we gathered from the analysis tool. This chapter focuses on our attempt to answer that question. Single quasi-identifiers are discussed first, with combinations following shortly thereafter.

6.1 Single QID's

In order to determine the reduction factor of every QID on a person's k-anonymity value, we need to compare the reduced k-anonymity value to the whole of the Dutch population.

A separate run of our program for each QID yields the smallest discernable k-anonymity value per QID. A trivial calculation, (dividing k-anonymity of the total population by the k-anonymity of the search query) then results in table 6.1. Additionally, output of the tool also provides us with graphs that assist in visualizing our results. The next two figures (Figure 6.1, Figure 6.2) are examples of these graphs.

Figure 6.1 clearly shows that the Gender QID is almost uniformly distributed over all ages. The most interesting information we can glean from this figure is the distance between the lines of the plots. Comparing either gender line with the total population line yields a value per age. This value averaged over all ages is what we call the identification factor. This value is also visible in Table 6.1 and implies the factor by which an individuals anonymity decreases when this QID is entered in a survey or questionnaire.

Figure 6.2 gives an example of data that is not uniformly distributed. This figure shows the marital status plotted against age. Again we see the total population and the effect of filling in one of the QID's. The identification factor can vary greatly over a QID and this can be seen clearly in the graph. The amount of married people, for example, only really starts increasing from the age of 20 (obviously) and then increases rapidly. This has an effect on the identification factor, as that factor will be close to zero during childhood yet significantly larger when age increases. This phenomenon also occurs for other QID's, although in different fashions. We can compute an average for a QID but this does not mean much on its own as it depends heavily on a person's situation. What we can conclude from these results is that there are a high

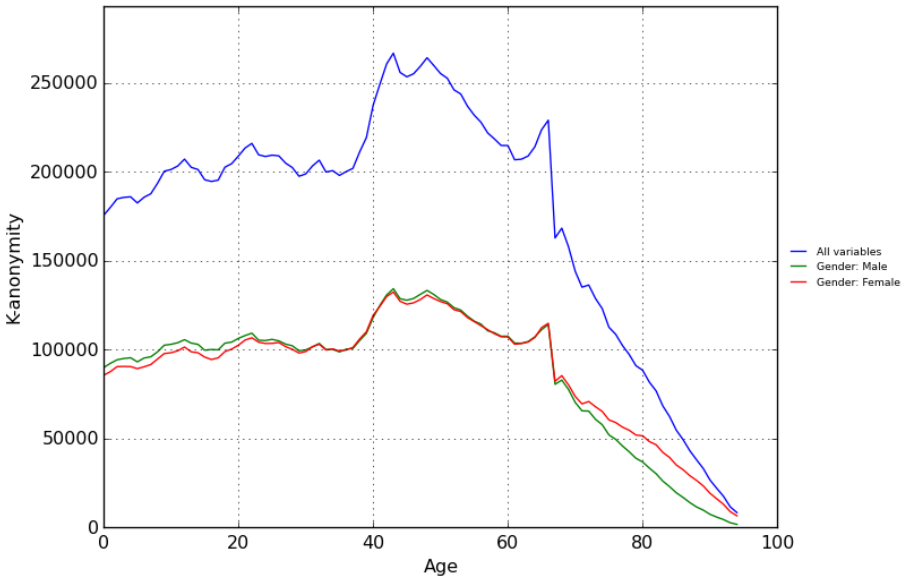

Figure 6.1: Gender(Male,Female) over total Dutch population

number of outliers stretched over all QID's with uneven data distribution.

The analysis of all single QID's results in table 6.1, where we can clearly see the selected variables from our analysis tool and their respective identification factor. An interesting QID to examine is postal code. We did not iterate over every available postal code in the Netherlands as this would take too long for the program to complete. Instead, a subset of thirty random postal codes were selected and analyzed to provide us with a rough estimate for the total population. This is also why the 'Population' entry is an average of the included postal codes and therefore not consistent with the total Dutch population.

The factors in Table 6.1 demand some attention as well as we see a clear distinction between QID's. The biggest factor, by far, is postal codes where we see an average identification factor of 13,282 which is about 700 times larger than the next biggest factor. From this we can conclude that revealing one's postal code is the largest threat to a person's anonymity. This is not even taking into account complete 6-symbol postal codes, where we expect an even higher value. Other notable factors are to be found in marital status and household, but these do not have the same impact as the postal codes QID. Even though average factors are not that high, we have to specify that outliers do have a very high factor as shown before in Figure 6.2 although they are obscured by the average.

6.2 QID combinations

Now that we know the identification factor of every single QID, we can check if a combination of QID's results in a combined factor that corresponds with multiplying single QID factors.

Figure 6.3 and Figure 6.4 show us a few QID combinations. The structure

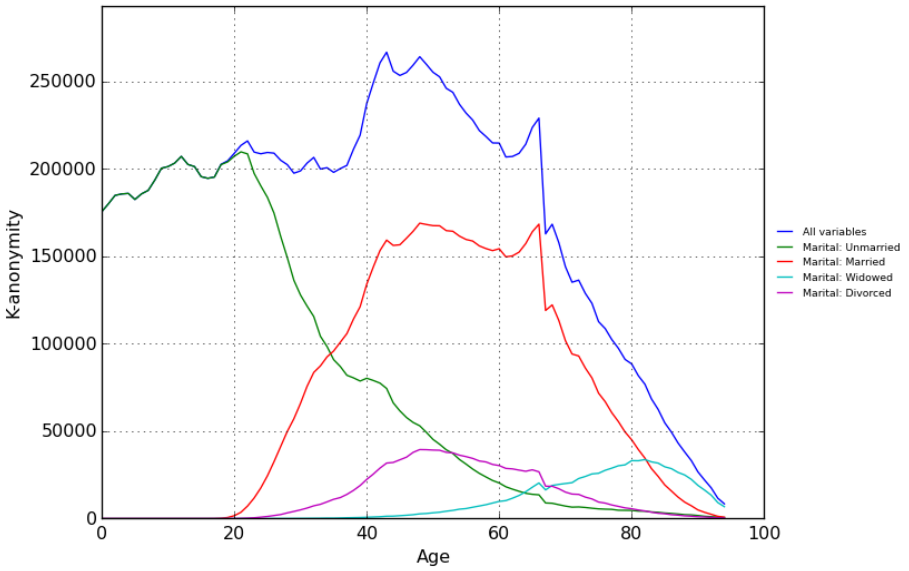

Figure 6.2: Marital status over total Dutch population

of the figures is just like Figure 6.1 and Figure 6.2. Figure 6.3 does have a few lines that overlap and thus not visible to the viewer. Although we do get the actual values in a table as shown in Table 6.2, the explanation is that these separate results are generated from the same dataset and do not yield any smaller k-anonymity value. This overlap affects the CoB Father/Mother QID set to 'Other'. Other than the overlap, we notice a pattern for some QID's in our results. From the figures we can tell if a certain QID is evenly distributed or not. In Figure 6.3 it can be seen that having the Dutch nationality is almost identical to the overall population line, suggesting that this identification factor is constant over almost all ages.

The values of several QID combinations can be viewed in Table 6.2 which shows mostly the Gender QID with a couple of other ones. Postal code, again, consists of the same 30 random postal codes we selected earlier. The other combinations cover the entire Dutch population. All combinations have their identification factor listed as before. The main result from this second table is that the identification factors cannot be multiplied directly with each other. They do add up to the correct value for some values and yet fall short for other ones. As an example, we look at the first combination and their respective values from Table 6.1. When we multiply these factors, the result is exactly the combined value. But when we take another combination that includes a QID like postal code we get a factor that is less than what we would expect from the multiplication. The reason for this discrepancy is the data distribution of the QID's, if all factors in the combination are uniform you can multiply the factors and get the same result. If one or more QID's have a non-uniform data distribution in a combination you could get a higher or lower combined factor. Finally, the combination containing postal code is again the highest and factors across the table are higher than they are in the previous table. This leads to the conclusion that more QID's (sometimes significantly) increases the identification

Table 6.1: Table containing the (average) identification factor for all analyzed QID's and their populations

QID	Input	Factor	Population
Gender	Total	2	16,759,854
	Male	2.02	8,303,834
	Female	1.98	8,456,020
Postal Code	Total	-	16,759,854
	Subset	13,282	24,981
Marital Status	Total	9.60	16,759,854
	Unmarried	2.12	7,894,264
	Married	2.46	6,822,724
	Widowed	19.77	847,869
	Divorced	14.03	1,194,997
CoB You	Total	5.24	16,759,854
	Netherlands	1.12	14,967,802
	Other	9.35	1,792,052
CoB Father/Mother	Total	2.89	16,759,854
	Netherlands	1.05	16,005,184
	Other	4.73	3,541,478
Household	Total	3.18	16,759,854
	Child living at home	2.58	6,492,910
	Single	1.44	11,649,136
	Living together with children	4.47	3,751,585
	Living together without children	3.37	4,967,160
	Married with children	1.93	8,705,608
	Married without children	2.04	8,224,393
	Single parent	6.44	2,604,695

factor.

6.3 Maximum and minimum factor

As an experiment we ran another combination that contained the highest factors we could find in our table for all possible QID's except for postal code. The result of this is shown in Figure 6.5.

The interesting part of this figure are the k-anonymity values across the graph. The peak does not even reach a value of 400 and there are quite a few results hitting 200 and less. The average k-anonymity value is about 150 on a total population of nearly 17 million. The maximum factor excluding postal code therefore passes $110,000$. A combination with the minimum factors yields a factor of around 960 and can be considered to be the largest group of people or an upper border.

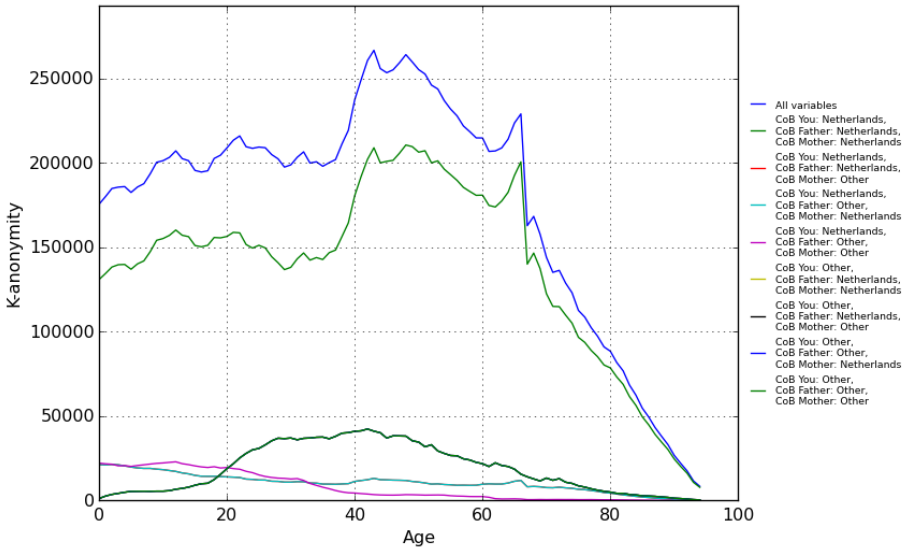

Figure 6.3: Country of birth of a person, (his, her) father and mother over total Dutch population

Table 6.2: QID combinations and their respective factors

QID's	Input	Factor	Population
Gender + CoB You	Total	10.49	16,759,854
Gender + CoB Father/Mother	Total	5.78	16,759,854
Gender + Marital Status	Total	24.08	16,759,854
Gender + Household	Total	7.69	16,759,854
Gender + Postal Code	Subset	19801.62	748,817
Gender + CoB You + Marital Status	Total	166.25	16,759,854
CoB You + CoB Father + CoB Mother	Total	12.18	16,759,854

6.4 Difficulties

During our research we encountered several difficulties. The lack of complete accuracy in our CBS datasets and the existence of outliers proved to be problematic.

6.4.1 Data loss

While generating results we encountered a few examples of data loss. As can be seen in Figure 6.6, the graph has a block-like flow where the smallest possible k-anonymity is being taken from a dataset that contains age groups. The block flow is not a real problem here, its the fact that the k-anonymity value for the whole group is counted five times instead of once. This is a problem, as the k-anonymity value of a group seems higher than it actually is. The cause of the problem is that the program looks through our datasets for every age and finds the exact same value five times in a row in the same age group. The result is that the graph shown is not an accurate figure. This effect does show

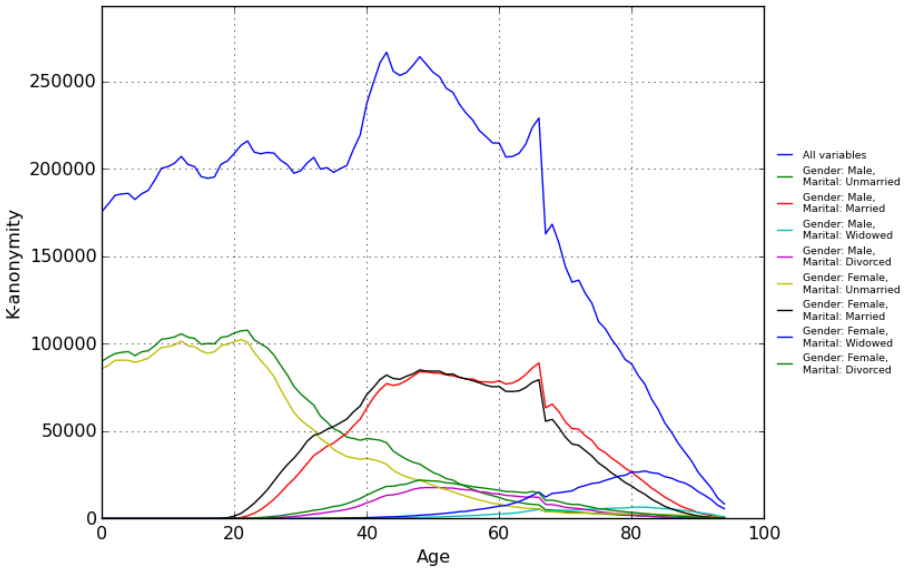

Figure 6.4: Gender(Male,Female) combined with marital status over total Dutch population

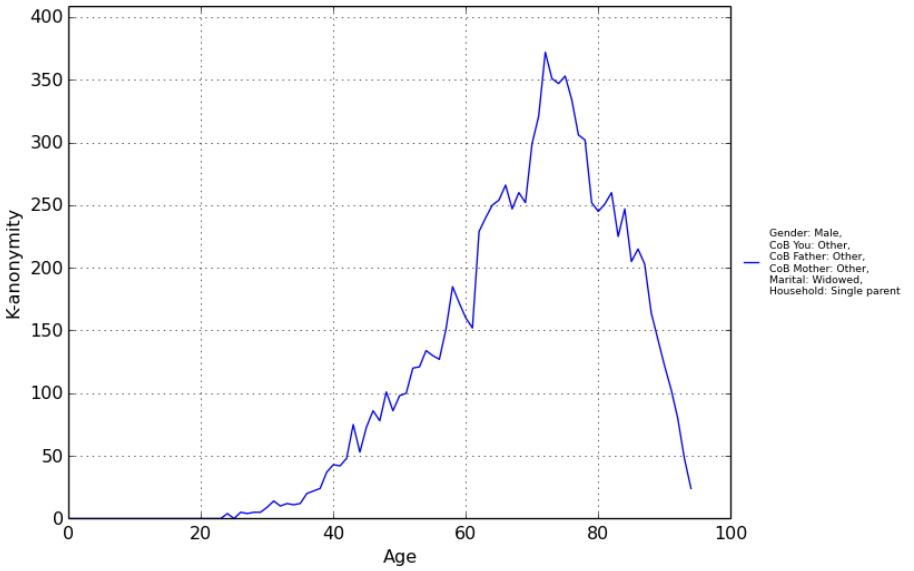

Figure 6.5: Maximum factor QID combination

that forming age groups in datasets is a good way of increasing anonymity and especially for outliers, as they can be 'hidden' inside these age groups.

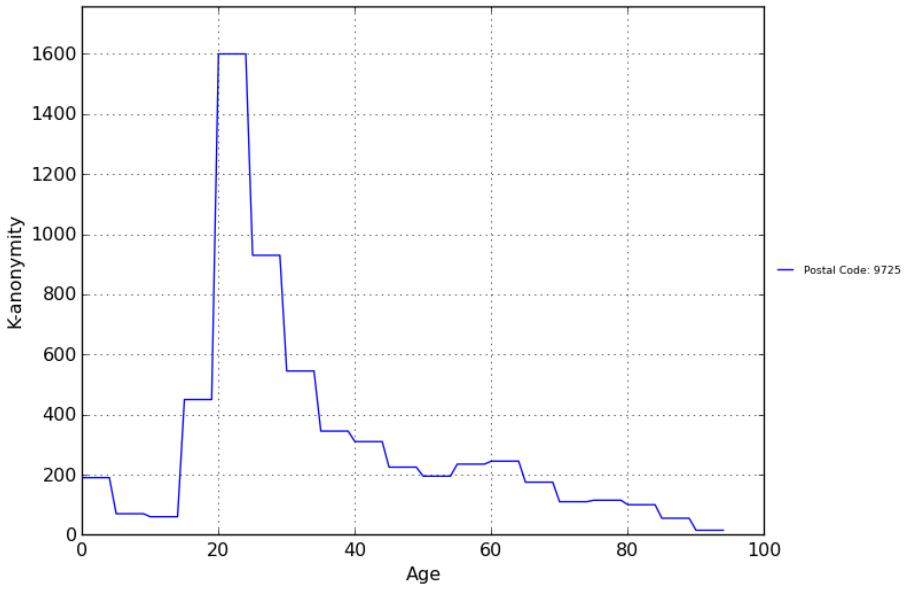

Figure 6.6: Age groups, data loss example

Table 6.3: Outliers. Note: Postal Code is a subset of 30 randomly selected postal codes

QID combination	$k = 1$	$k \leq 5$	$k \leq 10$	$k \leq 50$	$k \leq 100$
-	0	0	0	0	0
YoB	0	0	0	0	0
YoB + Gender	0	0	0	0	0
YoB + CoB You	0	0	0	0	0
YoB + Gender + CoB You	0	0	0	0	1
YoB + Gender + CoB You/Father	0	0	0	1	4
YoB + CoB All	0	0	0	2	6
YoB + Gender + Marital	3	7	15	30	42
YoB + Household	0	1	1	3	3
YoB + Marital Status	2	5	8	12	16
YoB + Postal Code	1	32	57	173	240

6.4.2 Outliers

One problem with the use of an identification factor is that it does not account for outliers. An outlier is, for example, a case where we find a very low k-anonymity value in comparison to the average of the search query. We show some examples of this phenomenon in Table 6.3. In this table we have some reasonably generic search queries which still result in a number of people with low k-anonymity values.

We define an outlier to have a k-anonymity value lower than five, while the average over all ages is higher than 500. While it is hard to calculate the exact number of outliers, by looking at Table 6.3 and extrapolating this

data, we estimate that it is below 0.1% of all possible results. While this is quite a small number it is still important to keep these outliers in mind when drawing conclusions from the identification factor. To conclude, it can be stated that when the number of outliers is larger, that particular identifier is more dangerous.

Chapter 7

Conclusion

From the results we can draw the following conclusions. When comparing the different identifiers it is clear that postal code is by far the most dangerous one and has the largest negative impact on a persons k-anonymity. This is especially true when it is taken into account that it has the largest number of outliers of all identifiers. We do need to mention that this identifier is only dangerous in specific cases, since it is clear that a postal code with a large number of inhabitants poses a significantly lower risk to the anonymity than a postal code with few inhabitants. The other single identifiers are not nearly as dangerous as postal codes and in general they do not have a very large impact on a persons anonymity.

When we consider combinations of identifiers, it becomes clear that combinations containing postal codes are the most dangerous. This reinforces our statement that postal code is by far the most dangerous identifier. There are some combinations which can be used to identify a number of people, but in comparison to postal code this number is quite small.

Summarizing, while some identifiers are more dangerous than others, the exact impact of an identifier on a persons anonymity depends greatly on the situation of that specific person and if he or she is an outlier or not. At the same time is it very hard to completely indentify an individual since the datasets that are used are de-identified to a large extent, which makes the identification process much harder.

In order to give an answer to the research question given in the introduction, we restate and answer the sub questions below.

1. Which variables are responsible for revealing a person's identity?

   All variables are responsible for revealing a person's identity, although the degree in which they do so differs greatly. We can conclude from our research that postal code has the largest identification factor. The other variables do reveal a person's identity, but very slightly in comparison to postal code.

2. What is the anonymity reduction factor of each demographic variable?

   In order to quantify the threat of each variable, we calculated the identification factor of all variables. The average factor of each variable is not very threatening except for postal code. The difference between postal

code and other variables is notable, although it is possible for a variable other than postal code to get an unusually high factor in an edge case.

3. What is the effect of variable combinations compared to single variables?

   The effect of combinations is directly multiplicative on the identification factor of the single variables. Unless the variables in question do not have an even data distribution. Outliers tend to distort the identification factor in scenarios where uneven data distribution is the case.

4. What can be done to reduce the risk of unique identification through anonymous data?

   During our research and analysis we suffered from de-identification of our CBS datasets. This is caused by CBS methods to limit the risk of unique identification. One of those methods is to group age together in groups of five. A notable effect of this method is that outliers tend to be hidden inside the age groups and receive a lower chance of unique identification. Another example of this is postal code, where postal codes are grouped by municipality in some datasets. Grouping of data is therefore a valid way of obscuring outliers and increasing overal anonymity.

To answer our research question:
What is the risk that a person's real identity be revealed through anonymous data?

The overall risk of unique identification is very small. The only way a person would really be uniquely identified using our data, is when a multitude of variables would be used or when such a person is considered to be an outlier. Removing postal code and hiding outliers proved to be enough to almost guarantee a high enough k-anonymity to be considered anonymous.

Chapter 8

Future Work

Based on the work in this research project, further research can be done on the topic of questionnaire and survey anonymity.

It is still unclear to us to what extent the CBS uses methods of de-identification on their data. Therefore, the de-identification process of the CBS and other organizations could be analyzed to try and find out which methods they use.

The data used for this thesis is not as accurate as we hoped for. More work could be done to look for alternative datasets or to set up an experiment with simulated data. A simulation of a full dataset is not a large amount of work and could yield precise results with full control over the data. The question is whether our research can be improved by utilizing more accurate data. A good scenario would be where one could compare results with previous research, as done by Koot, Golle, and Sweeney.

Another idea is to include additional variables and to see if the average k-anonymity could be brought down to significantly low levels. Instead of constructing a tool, one could have a look at Oracle database software to make use of MySQL. This might speed up searches and could be interesting to compare to our analysis tool, for example. Dedicated database software can assist in managing larger datasets, if a larger amount of variables will be analyzed.

References

Dalenius, T. (1986). Finding a needle in a haystack-or identifying anonymous census records. *Journal of the American Medical Informatics Association*, *2*, 329-336.

Emam, K. E., & Dankar, F. K. (2008). Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, *15*.

Golle, P. (2006). *Revisiting the uniqueness of simple demographics in the us population.*

Koot, M. R. (2012). *Measuring and predicting anonymity* (Unpublished doctoral dissertation). Universiteit van Amsterdam.

Sweeney, L. (2000). *Simple demographics often identify people uniquely.*

Sweeney, L. (2002). K-anonymity: a model for protecting privacy. *Internation Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*.

Appendix A

UMCG Advice

Maintaining privacy is one of the most challenging aspects of public research. Therefore, it is critical that seemingly anonymous research data is actually anonymous.

How does one protect the privacy of participants in a survey or questionnaire? What is the impact of specific demographic information on a person's anonymity? These questions stand at the core of the research we, as a group of three bachelor-students, performed for our bachelor's thesis. The focus of this document is to advise a research institution or organization how to preserve anonymity (and privacy) of survey and questionnaire subjects and still gather sufficient data.

In surveys and questionnaires, respondents are asked for basic demographic information about their person. Researchers aim for this information to be as detailed as possible. The threat to privacy is often overlooked when collecting personal information, especially when it concerns data devoid of personal information which can directly identify an individual. What is left, are identifiers which, on their own, do not identify a person completely. We call these identifiers quasi-identifiers (or QID's).

It is important to determine the weight of QID's on the anonymity of an individual. We can conclude from our research that a QID, or combination thereof, can result in a large decrease in anonymity. With location information being the largest offender (e.g. postal code) and long combinations to be another. Furthermore, special cases regularly occur in datasets containing demographic information which we call outliers. These outliers have a very specific set of characteristics and are often easy to identify

In order to prevent unnecessary threats to anonymity, we distinguish two main cases and provide advice to minimize the loss of privacy.

- A dataset to be published at some point:

  In this case, a survey or questionnaire has developed into a dataset which is set to be published. As the data will be publicly available, we need to make sure that outliers are sufficiently protected and reduce the impact from location information and QID combinations.

- A dataset that remains private:

The remaining case is where the dataset will not be published at all. The only real risk in this situation is a data leak.

Several solutions are available to protect privacy in both cases. An important tool is the grouping of information. Grouping is basically reducing the accuracy of a QID to such an extent that the possible threat to privacy is reduced to an acceptable level. Examples are age and postal code, where specific ages can be pooled together into ranges (0 to 5, 5 to 10) and where postal codes can be hidden inside their municipalities or even provinces. Grouping can be done when formulating questionnaires or afterwards, when data has been collected.

Another option would be to store QID's separately. This prevents QID's from being linked to each other and separates them from sensitive data in questionnaires and surveys. There is no need to store everything separately, only the most identifying QID's should be sufficient. Both options greatly reduce the occurence of outliers, as they are 'hidden' within the data.

The downside to both of these solutions is the loss of data. There will always be a tradeoff between anonymity and accuracy. It is important to strike a balance between having data that is accurate enough to be of use, yet provides enough anonymity to the people who are part of the survey or questionnaire.

When publishing at some point, we advise to use a combination of the above. It is not necessary to use grouping on all the data, or to store everything separately. Try and locate the most important threats to anonymity for a dataset and act accordingly.

When deciding not to publish, grouping is less ideal and a secure protection of dangerous QID's should be sufficient to reduce the anonymity risk. This situation is very similar to creditcard information, where important numbers are hidden in a separate secure environment.

The best advice we can provide is to think about the questions one might ask in a survey or questionnaire and to judge if these are really necessary for the research goal. Is this kind of personal information relevant? Or is the added risk to privacy acceptable for the increased accuracy of the dataset?

To provide an example, in order to increase the anonymity of the HowNutsAreTheDutch (HND) questionnaires:

- Instead of asking for full date of birth, just ask for year of birth or allow for selection of a range of ages.
- Instead of asking for full postal code, limit the question to a city, municipality or the first few digits of a postal code.
- Is it necessary to ask for the country of birth for either parent?
- An example of a less specific question already in the questionnaires is: 'Are you in a relationship?' Without specifying what kind of relationship.

These are but a few examples of ways to handle a questionnaire. Keep in mind that these are examples and we cannot guarantee to what extent these specific questionnaires need to be de-anonymized and still provide a useful dataset. Finding the balance is up to the researchers or institution.