Wetenschappelijk Onderzoek- en Documentatiecentrum
Ministerie van Justitie en Veiligheid

Cahier 2018-20

On statistical disclosure control technologies

For enabling personal data protection in open data settings

M.S. Bargh
R. Meijer
M. Vink

Cahier
De reeks Cahier omvat de rapporten van onderzoek dat door en in opdracht van het WODC is verricht.
Opname in de reeks betekent niet dat de inhoud van de rapporten het standpunt van de Minister van Justitie en Veiligheid weergeeft.

Alle rapporten van het WODC zijn gratis te downloaden van www.wodc.nl

Preface

To increase its transparency, accountability and efficiency through open data, the Dutch Ministry of Justice and Security (MJ&S) has set up an open data program so that the publicly funded datasets of the ministry can be shared with the public and with other organisations. To this end, protecting privacy has become a growing challenge, because both the amount of data and the threat of data-abuse are growing rapidly.

Sharing data responsibly is an important precondition for the MJ&S to share its data. Therefore, the Research and Documentation Centre (abbreviated as WODC[1] in Dutch) has studied the tools and methods that can support professionals in protecting privacy-sensitive data. Two important aspects of data protection are data minimisation and use of data only for the purpose in mind. These data minimisation and purpose limitation are two important principles of the General Data Protection Regulation (GDPR) for protecting privacy.

The study shows that the Statistical Disclosure Control (SDC) tools and methods studied can be used to support realising these two principles of the GDPR. The tools help the professionals make appropriate trade-offs between privacy and utility of data. Such SDC tools and methods can be used when sharing data within a limited group as well as when opening up data to the public. The results of this study are relevant for data analysts and data managers who want to learn about and use SDC technologies to protect personal data or to analyse the data sets modified by SDC technologies.

This study, the results of which are presented in this report, was funded by the Information and Purchasing Department (abbreviated as DII[2] in Dutch) of the MJ&S. The authors and I are grateful to the DII department for making this study possible. We are also indebted to the Police for the delivery of a realistic data set, which was used to experiment with the SDC methods.

I thank, also on behalf of the authors, the members of the advisory committee (prof.dr.ir. Marijn Janssen (chairman), dr.ir. Maurice van Keulen, Henk-Jan van der Molen CISSP/CISM/CISA, mr.dr. Marc van Opijnen and mr. Just Stam) as well as the research advisors (drs. Walter Schirm and drs. Fanny Wallebroek) for their valuable contribution to this study.Finally, I thank also on behalf of the authors, the reviewers of the report (dr.ir. Sunil Choenni and dr. Susan van den Braak) for their constructive criticism.

Acting director WODC
A.L. Daalder

[1] WODC: Wetenschappelijk Onderzoek- en Documentatiecentrum.

[2] DII: Directie Informatievoorziening en Inkoop.

Contents

Abstract

To enhance the transparency, accountability and efficiency of the Dutch Ministry of Justice and Security, the ministry has set up an open data program to proactively stimulate sharing its (public-funded) data sets with the public or with other organisations. Disclosure of personal data is considered as one of the main threats for data opening. This study, as one activity within the open data program, aims at investigating Statistical Disclosure Control (SDC) tools and methods for protecting personal data. More specifically, the main objective of the study is to provide insights in the main functionalities provided by SDC technologies so that data controllers can be supported in their decision-making processes related to storing, sharing, and opening the ministry's personal data. To this end, the study context is tuned to the ministry's settings and requirements. This deliverable presents the acquired insights, particularly for three selected open source SDC tools (namely: μ-ARGUS, ARX and sdcMicro).

Abbreviations

AVG	Algemene Verordening Gegevensbescherming
CISO	Chief Information Security Officer
CPO	Chief Privacy Officer
CRM	Customer Relationship Management
DII	Directie Informatievoorziening en Inkoop
DPA	Data Protection Act
DPIA	Data Protection Impact Assessment
EC	Equivalent Class
EID	Explicit IDentifier
EU	European Union
FLOSS	Free/Libre/Open Source Software
GDPR	General Data Protection Regulation
JSON	JavaScript Object Notation
NAT	Non-sensitive ATtribute
OGA	Open Government Act
PPDP	Privacy Protecting Data Publishing
PRAM	Post RAndomisation Method
PU plane	Privacy-Utility plane
QID	Quasi IDentifier
SAT	Sensitive ATtribute
SDC	Statistical Disclosure Control
SUDA	Special Uniqueness Detection Algorithm
XML	Extensible Markup Language

Management summary

Background, scope and research questions
Growth of data – in terms of, for example, their volume, variety and velocity – increases the threat of personal data disclosures (or data disclosures, in short). On the one hand, the growth (in size) of a data set makes it difficult to detect and deal with those data disclosure risks that are hidden in the data set (i.e., the intrinsic risk factors). On the other hand, the growth (in size or number) of other data sets (i.e., the increase of the *background knowledge* available to other parties) makes it difficult to assess and deal with the data disclosure risks that may arise when combining the data set with other data sets (i.e., the extrinsic risk factors). Consequently, it becomes difficult for data controllers to share their data with specific groups, individuals or the public – where the latter, i.e., sharing data with the public, means to open the data.

Disclosing sensitive information about individuals can occur when personal data are transferred, stored or analysed. Information security mechanisms, such as data encryption and access control, can be used to protect data in transit or storage. When data are already accessed (be it legitimately or illegitimately), it is still possible to disclose sensitive information about individuals illegitimately (i.e., unauthorised data usage). Even if directly identifying information (like names) is removed from the data, a legitimate or illegitimate data accessor may use statistical disclosure mechanisms to reidentify some data items, particularly by using other information sources. For example, the term 'mayor of Amsterdam' in a data set can reveal the identity of an individual if you already know who that mayor is or if you can find it out with a Google search. Data controllers in turn, can use *Statistical Disclosure Control (SDC) technologies* to mitigate the intrinsic and extrinsic data disclosure risks in such cases where the data are accessed either legitimately or illegitimately, but are analysed illegitimately.

SDC technologies aim at eliminating both directly and indirectly identifying information in a data set, while preserving data quality (i.e., the so-called *data utility* in SDC settings) as much as possible. Directly identifying information (like names and social security numbers) and indirectly identifying information (like the combination of birthdate, postal code and gender) in a data set contribute to its intrinsic and extrinsic risk factors, respectively. SDC technologies can be applied to microdata sets and aggregated data sets. Microdata sets, which may have (very) large sizes, are referred to structured tables with some rows, representing individuals and individual units like households, and a number of columns, representing the attributes of those individuals (like their age, gender and occupation). Aggregated data sets include frequency tables that contain the numbers of individuals in some groups (like the number of the residents in a district) and quantitative tables that contain the sums of individuals' attribute values (like the total income of the individuals who work in a specific department of a company).

The scope of this study is limited to the SDC technologies for protecting microdata sets. Within this study, we are particularly concerned with protecting justice domain data sets for open data purposes. Note that the scope of this study and the applicability domain of SDC technologies are wider than just open data. We pay special attention to data opening because the Dutch Ministry of Justice and Security intends to boost its open data initiatives for improving its transparency and accountability.

Within this context, the objective of the study is *to investigate SDC technologies for protecting microdata sets.* To this end, we define and address the following research questions:

1 What are the legal constraints relevant for SDC-based data protection, particularly for opening justice domain data?
2 What are the main functionalities of available SDC tools for protecting personal data and preserving data utility?
3 How can background knowledge be accounted for in SDC-based protection of personal data?
4 What are (other) promising SDC functionalities or methods (proposed in literature)?

Methodology and results

To answer the research questions, we have carried out an extensive desk research over the relevant topics such as privacy enhancing technologies, SDC methods, privacy impact assessment processes, (new) laws and regulations, and open data initiatives. Further, we have presented our intermediary results to various (expertise) groups such as data analysts, privacy experts, in job trainees, and (applied) university students to fine-tune the scope, select relevant topics, and to perform a sanity check on the results and approach.

For addressing the first research question, we have additionally carried out semi-structured interviews with three data protection experts experienced with privacy laws and regulations. Further, to answer the second research question, we have devised and carried out a number of experiments to obtain a preliminary indication of the usability and scalability aspects of the SDC tools.

In the following, we briefly describe the main results of the study per research question.

On legal constraints

In light of *General Data Protection Regulation (GDPR*; see GDPR, 2016), SDC technologies can be used to realise the data minimisation, purpose limitation, and proportionality principles of GDPR. Specifically, SDC tools can provide insights into and mechanisms for (a) transforming raw data, (b) assessing the utility of the raw and transformed data, (c) estimating the data disclosure risks of the raw and transformed data, and (d) making trade-offs between data utility aspects and data disclosure risks. These SDC-based insights and SDC mechanisms, we conclude, are necessary for data controllers to become GDPR compliant when sharing and opening their data nowadays.

Pseudonymisation and anonymisation are two important terms within the domain of SDC technologies. These terms are not defined uniformly and are used differently in legal and technological domains. We note that, for example, most data anonymisation mechanisms in the technological sense can be regarded as data pseudonymisation mechanisms in the GDPR sense. As part of our study context, we elaborate on these terminological differences.

Justice domain data are mainly concerned with sensitive personal data (for example, criminal justice and law enforcement data). Not including personal information plays an important role – if not to say a necessary role – for opening privacy-sensitive justice domain data. Therefore, we also investigate when a data set can be considered as being without personal information (or anonymous) according to GDPR.

For data being considered as anonymous, we propose the notion of a threshold to mark the boundary of data anonymity. This threshold is basically context (and time) dependent (i.e., depending on, for example, available technologies and their advancements, other available data sources, and the motivations for and costs of reidentifications). Therefore, data disclosure risks may increase in the future, i.e., the currently anonymous data may become non-anonymous personal data, as the anonymity threshold level rises over time. Sometimes, on the other hand, the threshold level may subside, for instance, in case that the current background knowledge does no longer exist.

On main functionalities of SDC tools
In this study, we investigated three non-commercial open source software SDC tools, namely: μ-ARGUS, ARX and sdcMicro. On the one hand, the investigation of the tools enabled us to (a) obtain an insight into main SDC functionalities (by the virtue of being developed/deployed in these existing tools), (b) obtain hands-on experience about SDC technologies (by experimenting with these SDC tools), and (c) learn from the experiences of the research community and academia (as they incline towards easy and free to learn, use, and extend software tools).

On the other hand, the investigation of the SDC tools (together with our literature study) led us to characterise SDC technologies with a generic functional model, which comprises four components of

- data transformation to transform an original microdata set to a transformed microdata set by using SDC methods and models;
- data disclosure risk measurement to quantify the data disclosure risks in the transformed microdata set by considering data disclosure scenarios and linkage types;
- data utility measurement to quantify the data quality of the transformed microdata set; and
- trade-off evaluation to make trade-offs between the data disclosure risks and data utility aspects of the transformed microdata set.

This SDC functional model includes also a feedback loop to indicate systemically the underlying process when using SDC tools for data anonymisation.

Using the functional model, we provide an insight in the main functionalities of the SDC tools, i.e., per component of the functional model. The data transformation component includes SDC methods (such as removal, suppression, pseudonymisation, generalisation, permutation, perturbation and anatomisation) and SDC models (such as k-anonymity, l-diversity, t-closeness, k-map and δ-presence). Generally, a combination of SDC methods are used to realise an SDC model and a combination of SDC models are realised within an SDC tool. The data disclosure risk measurement, which considers the disclosure scenarios and the uniqueness aspects of data items, includes two risk measurement categories: elementary measures (like the values of k and l in k-anonymity and l-diversity) and advanced measures (which, in turn, rely on defining data disclosure scenarios such as prosecutor, journalist, and marketer attackers). The data utility measurement component includes general-purpose measures (like discernibility measure and special-purpose measures (like classification measure and classification performance measures). The data privacy-utility evaluation component relies on human expertise mainly to make a trade-off between the disclosure risks and utility of the transformed microdata set based on the corresponding measurements.

Further, we propose a framework to examine the non-functional aspects of these SDC tools, based on a usability perspective relevant to our study (i.e., for data analysts who want to learn about SDC technologies). This framework comprises the following criteria:

1 ease of access or availability, for instance, being open source, being free of charge, and being platform independent;
2 ease of use, for instance, ease of data import, ease of data processing, ease of data export, and having user-interface/GUI;
3 ease of learning, for instance, availability of documentation, quality of the documentation, community support, and intuitiveness of the tool;
4 ease of extension, for instance, integration capability with other software, number of active developers, recent maintenance activities, and developer support.

Finally, we describe an experiment for testing the execution time (i.e., a specific aspect of performance) of the three SDC tools investigated. To this end, we have considered the differences in the functionalities provided by the three SDC tools in order to set up a uniform way of testing these tools as much as possible. In other words, the devised experiment aims at (a) being practically feasible and (b) delivering as much similar tests as possible for these tools. We designed our experiments in the following way:

- use ARX to find a number of generalisation settings, ordered according to their data utility measures as calculated by ARX;
- pick up the first generalisation setting from the list above;
- run ARX, μ-ARGUS and sdcMicro for the chosen generalisation setting, measure their execution times.

Our investigation of the functional aspects of the SDC tools show that ARX appears to be more accessible for newcomers and adopters comparatively. In other words, μ-ARGUS and sdcMicro are suitable for more experienced experts relatively.

On background knowledge

Increasingly being available to intruders, background knowledge is a key extrinsic risk factor. Background knowledge includes the information in publicly available databases or directories (like electoral registers, telephone directories, trade directories, registers of professional associations), in personal and informal contacts (due to or via, for example, co-locality and being neighbours), in social media; or in organisational databases (available to, for example, government agencies and commercial companies). During the attribute mapping activity of an SDC process, some attributes of microdata sets are designated as *Quasi Identifiers* (QIDs). QIDs refer to those attributes that intruders may use to link the identities of some data subjects, which are available in the other information sources, to the data items in the transformed microdata set. In protecting microdata sets via SDC tools, therefore, the background knowledge available to intruders is captured by appropriately defining the QIDs. We note that there is no universal way of attribute mapping, e.g., defining QIDs. Therefore, data controllers should carefully carry out this attribute mapping within an SDC process in order to contain disclosures risks and maintain data utility at acceptable levels.

On promising SDC functionalities

Investigating the range of SDC functionalities, which is based on studying the three SDC tools and the literature, has enabled us to develop a vision for joining forces of these tools and/or for extending these SDC tools in the future. We identify a number

of SDC functionalities that are useful to be included in (future) SDC tools, especially for protecting justice domain data sets. Examples of these functions are:

- risk assessment based on actual population microdata set;
- semiautomatic data transformation together with user involvement; and
- data anonymisation based on the characteristics of justice domain data (to deal with, e.g., continuous publishing and location dependency)

Discussion and follow-up research

Data protection technologies, in general, and SDC tools, in particular, cannot give a 100% guarantee against data disclosure risks. Having no 100% guarantee can particularly be attributed to the extrinsic risk factors in the data environment. Therefore, one should be realistic about the potentials of data protection technologies and applying them should not give a false sense of privacy. As there is generally no single solution to deliver guaranteed privacy, many practitioners advocate adopting a risk-based data protection approach, instead of a strictly guaranteed data protection one. This requires perceiving data protection as a continuous risk management process, not as a onetime operation with a binary outcome (i.e., resulting in being anonymous or not being anonymous forever). We think that SDC tools are an essential ingredient of such a risk management process. Enabling data controllers to become GDPR complaint when sharing and opening their data, SDC tools should be included in the Data Protection Impact Assessment (DPIA) process to identify and deal with data disclosure risks via data minimisation while maintaining data quality acceptable for a given purpose. To this end, we further argue that the role of SDC tools is to support (thus not to replace) domain experts. *In summary, we see applying SDC technologies as a necessary step for realising the due diligence principle that asks for putting sufficient efforts to protect personal data in a given context.*

SDC tools provide a wide range of functionalities, features, and configuration options for data controllers. In practice, however, it is not trivial to use and configure these tools when there are so many options to choose from. Use and configuration of these tools become even more cumbersome and complex when one considers also the variety of the data to be protected and the diversity of the data environment in/for which the data protection must be carried out. Further, one needs to be able to interpret and finetune the parameters of SDC tools and methods in order to appropriately support the decision-making process of data minimalisation. Therefore, *we recommend conducting further research on how to apply SDC tools to justice domain data, particularly by conducting a number of case studies with real data from the justice domain*.

Finally, based on the insight gained in this study, we provide a short list of research directions:

- to investigate the necessity and consequences of anonymity in the GDPR sense, also at the data controller and for open data initiatives;
- to devise a workflow for using an SDC tool in practice;
- to provide a guideline for configuration and interpretation of SDC parameters and results;
- to devise a methodology for effective collaboration among various stakeholders involved in the data anonymisation process so that SDC tools can effectively be used in practice;
- to carry out a number of case studies to characterise the SDC requirements of justice domain data sets for any data sharing (including data opening) purposes; and

- to devise complimentary (legal) measures needed before, during and after protecting data with SDC technologies.

1 Introduction

Data sharing with the public or specific groups must comply with, among others, the privacy rights of individuals. There are various technologies for protecting privacy-sensitive data (i.e., personal data). Statistical Disclosure Control (SDC) technologies refer to a subset of personal data protection mechanisms, developed for minimising personal data while sharing useful data. In this report, we present the results of our study of SDC technologies, particularly in the context of sharing or opening data from the justice domain.

In this introductory chapter, we present the study's motivations and objectives (Section 1.1), scope (Section 1.2), research objectives and questions (Section 1.3), and research methodology (Section 1.4). Finally, we present the outline of the report (Section 1.5).

1.1 Motivation and objective

Growth of data – in terms of, for example, its volume, variety and velocity – increases the threat of personal data disclosures (or data disclosures, in short). Consequently, it steadily becomes difficult for data controllers to share or open their data. On the one hand, with the growth of a data set, it becomes more difficult for data controllers to detect and deal with the risks of data disclosures hidden in the data set (i.e., the intrinsic factors of personal data discloser risks). On the other hand, the growth of other data sets (i.e., the background knowledge) makes it more difficult for data controllers to assess and deal with the risks of data disclosures that may arise when combining the data set with other data sets (i.e., extrinsic factors of personal data discloser risks). The amount of background knowledge available to intruders increases due to, for example, sequential data releases, multiple data releases, continuous data releases, collaborative data releases, big data infrastructures, social network applications and open data initiatives.

Consequently, one needs to augment the toolset of data controllers who have traditionally applied specific rules, often predefined in laws and legislations, for data protection. This augmentation requires developing and using state-of-the-art methods, metrics and software tools for gaining insight into potential intrinsic and extrinsic privacy (and information sensitivity) issues before (and perhaps after) data release.

SDC technologies reduce (or, ideally, eliminate) the personal data in a data set to be released. One important aspect to consider in this approach is to maintain the utility of the released data as much as possible after applying such technologies to the original data. Data utility relates to the quality of the released data, which can be defined (or, ideally, determined) based on the purpose for which the data are released. There are some metrics defined in literature for measuring data utility (e.g., for those metrics characterizing data quality, see Bargh et al., 2016) and the references therein). Besides the extent to which the personal data are indeed reduced or eliminated, a fair comparison of different data protecting technologies requires accounting for data utility after applying such technologies.

Studying SDC technologies becomes also highly relevant in light of EU General Data Protection Regulation (GDPR; see GDPR, 2016), which asks for a systematic reali-

sation of data minimisation and purpose limitation principles when processing personal data. To this end, GDPR asks for adopting a data protection by design/default approach and, in case of high privacy risks, for executing a Data Protection Impact Assessment (DPIA). The results of this study, as such, will enhance the knowledge-base and expertise within the Dutch Ministry of Justice and Security, needed for bridging the gap between the privacy by design approach and privacy engineering practice. In this way, eventually, the study contributes to the development of a socio-technological methodology for privacy engineering in the future.

The study, results of which are presented in this report, is financed by the Information Services and Purchasing Department[3] within the Dutch ministry of Justice and Security. To enhance its transparency, accountability and efficiency, the ministry has set up an open data program to proactively stimulate sharing its (public-funded) data sets with the public or with other organisations. Disclosure of personal data is considered as one of the main threats for data opening. This study, as one activity within the open data program, aims at investigating SDC technologies for protecting personal data. To this end, the study context is tuned, as much as possible, to the ministry's settings and requirements.

1.2 Scope

Disclosing sensitive information about individuals can occur when the data are being transmitted, stored or analysed. Information security mechanisms, such as data encryption and access control, can be used to protect data in transit or storage. These mechanisms protect personal data against so-called 'unauthorised access', as mentioned in Choenni et al. (2015). When data are accessed, either legitimately or illegitimately, it is still possible to disclose some sensitive information about individuals via statistical data disclosure mechanisms (e.g., via information inference). These situations are referred to as 'unauthorised-use' in Choenni et al. (2015). The scope of this work is limited to the latter category, which can be mitigated by using SDC technologies. Therefore, information security issues and mechanism are out of our scope.

The results of this study are relevant for data analysts and data managers who use SDC technologies to protect personal data or analyse the data sets modified by SDC technologies. As such, the target audience of this work, i.e., the aforementioned data managers and analysts, fall between cyber security experts (like CISOs – Chief Information Security Officers), privacy lawyers (like traditional CPOs – Chief Privacy Officers) and data analysts/scientists.

The study provides an overview of the main functionalities of SDC technologies in detecting and resolving data disclosure risks, particularly for opening and sharing the data sets coming from the justice domain. The term 'justice domain data sets' in this report denotes all the data that pertain to the justice branch of the Dutch government. The data range from the data of court proceedings and judgments to the data that are gathered within the administration and registration processes and procedures of the whole justice branch. These data are generally gathered by a number of independent organisations that are involved in the Dutch justice system (i.e., the organisations within the administration scope of the Dutch Ministry of Jus-

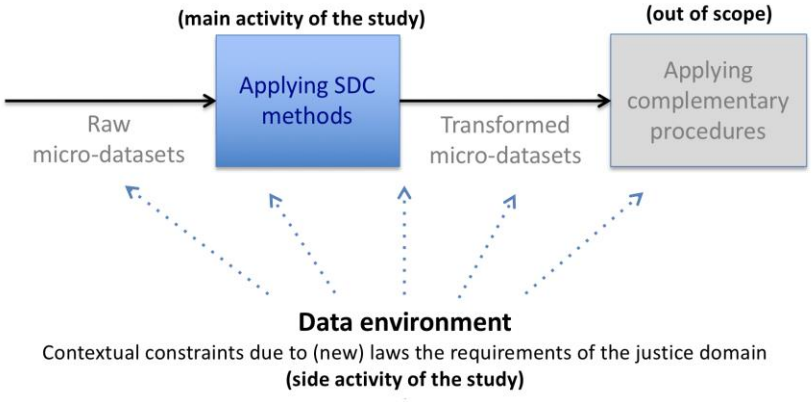
[3] In Dutch: "Directie Informatievoorziening en Inkoop" (DII).

tice and Security like the Public Prosecution Service, the courts, the Central Fine Collection Agency (CFCA) and the Police).

In this study we consider only microdata[4] sets, which refer to structured tables containing some rows, representing individuals or individual units like households, and a number of columns, representing some attributes about those individuals (like their age, gender and occupation). Frequency tables (containing, e.g., the numbers of individuals in some groups), quantitative tables (containing, e.g., the sums of incomes of the individuals in some groups), replies to statistical queries (containing answers to, e.g., the queries about the average, maximum, median, etc. of an attribute from a database), and semi-structured/unstructured documents (containing texts in natural languages partially/fully) are out of our scope.

There is no silver bullet in protecting personal data. The results of this work, therefore, should be considered as a means of enabling the due diligence principle when processing personal data. The objective of applying SDC technologies is to push the frontiers of data protection from applying simple methods, like data removal, to applying advanced methods, like data generalisation – which has limited adverse impacts on data quality compared to data removal. This paradigm shift is indicated in Figure 1. Furthermore, in this study we do not consider those complementary procedural solutions that link between the technological and non-technological (e.g., legal and governance) mechanisms when opening or sharing justice domain data sets. Developing a techno-procedural (or socio-technological) approach and its validation in practice are left for our future research. Nevertheless, we shall slightly elaborate upon the main legal constraints, which are particularly relevant for the justice domain. Note that we discuss the non-technological aspects as far as they are relevant for SDC technologies, as indicated by 'side activity' in Figure 1.

Figure 1 An illustration of the scope of the study, indicated by its main and side activities



[4] Note that the term 'micro' in microdata does not refer to the size of the data sets. Microdata sets may have a large number of records or attributes.

1.3 Research objective and questions

The objective of the study is to investigate SDC technologies, particularly for opening justice domain data sets. Therefore, we shall start with addressing the following research question:

> Q_1: *What are the legal constraints relevant for SDC-based data protection, particularly for opening justice domain data?*

As the main activity of the study, i.e., the technological aspects of the study, we shall continue with addressing the following three research questions:

> Q_2: *What are the main functionalities of available SDC tools for protecting personal data and preserving data utility?*

Regarding the data extrinsic factor of background knowledge, we shall address the following research question:

> Q_3: *How can background knowledge be accounted for in SDC-based protection of personal data?*

The intention is also to explore those state-of-the-art SDC mechanisms or functionalities that are not yet (widely) integrated in the SDC tools studied. In order to provide some guidelines for developing (new) SDC tools, we investigate also the following research question:

> Q_4: *What are (other) promising SDC functionalities or methods (proposed in literature)?*

1.4 Research methodology

For this study, we have carried out an extensive desk research over the relevant topics such as privacy enhancing technologies, SDC methods, privacy impact assessment processes, (new) laws and regulations, and open data initiatives. Further, we have presented our intermediary results to various (expertise) groups such as data analysts, privacy experts, in job trainees, and (applied) university students to fine-tune the scope, select relevant topics, and to sanity check the results and approach.

In order to position the study within the context of data sharing/opening in the justice domain (particularly for addressing research question Q_1), we have carried out semi-structured interviews with three data protection experts experienced with privacy laws and regulations. Further, we have devised and carried out a limited number of experiments to obtain an indication of the scalability and usability aspects of the SDC tools. For designing these experiments, we have tried to consider the differences in functionality of these tools in order to examine these tools as fairly as possible.

1.5 Outline

The report is organised as follows. The legal principles and constraints relevant for the study and the positioning of the study within the context of open data are presented in Chapter 2. Subsequently, we turn our focus on the technological aspects of SDC-based data protection. The theoretical background of SDC technologies (e.g., SDC concepts, methods and models) are presented in Chapter 3. Chapter 4 presents a generic functional model of SDC tools as well as the core components of the model for measuring data disclosure risks, for measuring data utility, and for making privacy-utility trade-offs. The functionalities and some non-functional aspects of the software tools studied are examined in Chapter 5. Chapter 6 discusses the study results and finally, our conclusions and recommendations for future research are presented in Chapter 1.

2 Study context with a reflection on legal aspects

In this chapter, we describe the context within which this study is initiated and carried out. This context can mainly be characterised by recent Dutch government policies to boost its open data initiatives as well as by recently coming into effect GDPR, to which, among other laws, such open data initiatives must comply. Understanding this context is crucial to define the scope and direction of the study and interpret its results. This chapter, as such, aims at answering the research question Q_1 (i.e., the legal constraints relevant for SDC-based data protection, particularly for opening justice domain data).

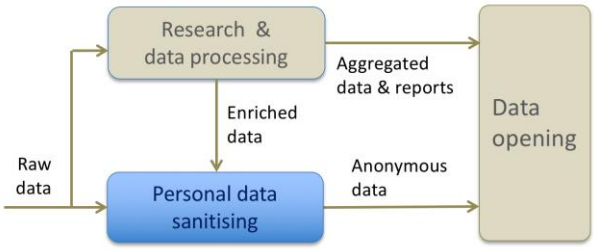
We start the chapter with sketching our vision of the open data infrastructure for the justice domain in Section 2.1. In Section 2.2, the main characteristics of GDPR and when it applies to justice domain data are briefly described. We shortly review the legal requirements of opening justice domain data sets in the Netherlands in Section 2.3. In Section 2.4, we describe two data protection concepts of GDPR that are particularly relevant for this study, i.e. pseudonymisation and anonymous information. Subsequently in Section 2.5, we explain the DPIA process, noting that DPIA is required by GDPR, and elaborate on the role of SDC within the DPIA process. Finally, we draw some conclusions in Section 2.6.

2.1 Data opening process

To improve its transparency, accountability and efficiency, the Dutch Ministry of Justice and Security seeks to open its (public-funded) data sets – containing registration data, research data and processed/aggregated data – to the public proactively. In order to share these justice domain data with the public, the data should in principle contain no privacy-sensitive data, as we shall explain in the following sections. Protecting personal data in this context asks for making trade-offs between contending values such as data privacy (representing rights of individuals) and data utility (representing the rights of the society), given the knowledge and insights available on the expected data privacy issues and threats.

In Figure 2 we illustrate a view on the process for opening and sharing justice domain data sets schematically. The raw data, which contain personal information potentially, are used for (scientific) research and data processing. This activity results in aggregated data and reports, which do not contain personal information anymore, as well as enriched/processed data, which may contain some personal information. The aggregated data and reports are shared with the public. The raw data and enhanced data are also good candidates for being shared with the public as open data (as well as with specific groups such as scholars, scientists and data-journalists). These data, nevertheless, should be protected against privacy risks and the required trade-offs should be made and evaluated. The component called 'personal data sanitising' in Figure 2 contains all such data protection activities. The scope of this study falls within the personal data sanitising component in Figure 2, noting that the study does not cover all data protection activities therein.

Figure 2 An illustration of the process for opening justice domain data



2.2 General Data Protection Regulation

Data processing has to be compliant to privacy laws and regulations. Although the focus of this study is not the legal aspects of privacy, we are going to describe GDPR highlights below in order to sketch the legal context of the study.

On the 25th of May 2018, GDPR came into force. From that moment on, the Dutch Data Protection Act (DPA), or in Dutch: Wbp[5] (see Wbp, 2000), stopped to be in effect. In Article 5 of GDPR eight data protection principles are mentioned. We focus here on (parts of) those principles that are relevant for the scope of this study, i.e., the SDC technologies and the corresponding aspects of data utility and data privacy.[6]

- Purpose limitation principle: personal data may only be collected for specified, explicit and legitimate purposes and not further be processed in a manner that is incompatible with those purposes, see Article 5(1-b) of GDPR.
- Data minimisation principle: personal data should be adequate, relevant and limited to what is necessary in relation to the purposes for which they are collected and processed, see Article 5(1-c) of GDPR.
- Data accuracy principle: data should be accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay, see Article 5(1-d) of GDPR.

In this study the focus is on justice domain data. For criminal justice data, being a subset of justice domain data, the new Directive EU 2016/680, see (Directive EU 2016/680, 2016), complements GDPR. Directive EU 2016/680, in full Directive Data Protection Law Enforcement[7], aims at processing personal data by law enforcement and supervisory authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties. However, when such personal data are processed for other purposes than those mentioned above, like archiving in the public interest or using them for scientific, statistical or historical work, in principle GDPR applies (see Article 4(3) of Directive EU 2016/680; and see WP29, 2017a).[8]

---

[5] In Dutch: 'Wet bescherming persoonsgegevens' (Wbp).

[6] The other principles are 'lawfulness, fairness and transparency', 'storage limitation', 'integrity and confidentiality' and 'accountability' (see Article 5 of GDPR).

[7] In Dutch: 'Richtlijn gegevensbescherming opsporing en vervolging'.

[8] Note that also the National GDPR Implementation Law of the Netherlands has recently come into force (since 16 May 2018). This GDPR Implementing Law determines the national rules to execute the GDPR. We will not include this law in our study. In addition to legal and organisational technological matters concerning the Dutch

2.3 Legal aspects of opening justice domain data in the Netherlands

Opening justice domain data, depending on their type, has to be compliant not only to GDPR, but also to other generic and specific (local) privacy laws and regulations. The open data policy of the Dutch government aims at opening data whenever this is compliant to privacy laws and regulations. The generic Open Government Act, or in Dutch: Wob[9] (see Wob, 1991), is seen as the pivotal law for deciding which data may (not) be opened according to the National Open Data Agenda of the Netherlands (NODA letter, 2015) and to our interviews. Wob contains the exceptions and limitations for opening of government data. Article 10(1-d) of Wob forbids in particular the opening of sensitive personal data, which include 'criminal justice and law enforcement data'[10]. Note that this forbiddance is in an absolute prohibition way (Memorandum, 1986), i.e., without taking, for example, the access to information rights of the public into consideration. However, in Article 10(1-d) of Wob there is an exception to the rule of not opening sensitive personal data, namely: 'unless this opening evidently does not lead to a breach of personal privacy'[11]. We argue that, in the context of open data, 'criminal justice and law enforcement data', as they are in their original form, do not qualify for the criterion 'evidently does not lead to a breach of personal privacy'. Therefore, we suspect this exception does not hold for opening of 'criminal justice and law enforcement data', as they are in their original form. It is, nevertheless, out of the scope of this study to further elaborate on the contention between 'unless this opening evidently does not lead to a breach of personal privacy' and 'forbiddance is in an absolute prohibition way' existing in the context of open data within Article 10(1-d) of Wob.

In addition to Wob, there are two important Dutch laws related to protecting personal data within the justice domain (especially for protecting the data pertaining to crime and criminal offences). These laws are the Law on Police Data, or in Dutch: Wpg[12] (see Wpg, 2007), and the Law on Judicial Information and Criminal Records Act, or in Dutch: Wjsg[13] (see Wjsg, 2002). Like Wob (see Articles 10 and 11), both Wpg (see Article 22) and Wjsg (see Article 15) allow opening criminal justice domain data, especially the data related to crime and offences, if the data imperatively do not contain any personal data nor lead to identifying persons.

From the discussion above we conclude that not including personal data plays an important role, if not to say to be a necessary condition, for opening justice domain data sets. In the following, therefore, we investigate when a data set can be considered as being without personal data according to GDPR. To this end, we shall look at when data are considered as anonymous and pseudonymised according to GDPR. Focusing on GDPR, we will not investigate the other abovementioned laws and regulations.

Data Protection Agency, this law mainly specifies additional rules about personal data processing, which are too detailed for the scope of this study.

9 In Dutch: 'Wet openbaarheid van bestuur' (Wob).

10 In Dutch, 'strafrechtelijke persoonsgegevens en persoonsgegevens over onrechtmatig of hinderlijk gedrag in verband met een opgelegd verbod naar aanleiding van dat gedrag', see Article 10 (1-d) of Wob and its reference to Article 16 of Wbp.

11 In Dutch: 'tenzij de verstrekking kennelijk geen inbreuk op de persoonlijke levenssfeer maakt', see Article 10 (1-d) of Wob.

12 In Dutch: 'Wet politiegegevens' (Wpg).

13 In Dutch: 'Wet justitiële en strafvorderlijke gegevens' (Wjsg).

2.4 Data protection according to GDPR

In this section we elaborate on data protection as perceived from the viewpoint of GDPR. We first describe which data items should be protected (Subsection 2.4.1) and then address how GDPR envisions data protection in general (Subsection 2.4.2). We focus on the concepts of pseudonymisation and anonymous information from GDPR viewpoint (Subsections 2.4.3, 2.4.4 and 2.4.5) due to their relevancy to SDC technologies and to open data.

2.4.1 Data items to protect

GDPR is applicable only when personal data are involved. According to GDPR, personal data refer to any information relating to an identified or identifiable natural person (so-called 'data subject'), as defined bellow.

Definition of an identifiable natural person*:* 'An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person' (see Article 4 of GDPR).

GDPR discerns several types of personal data in terms of identifiability and data types. These types are described in the following.

- *Directly identifiable data* relate to a person in a straightforward way, for instance someone's name or address.
- *Indirectly identifiable data* do not relate to a person, but they may still be considered personal data if they influence the way in which a certain person may be considered or treated in society. An example is the type of house or car of a data subject, because it may be a proxy of the income or wealth of that data subject. Further, the data that in combination with other data may lead to identifiability are to be seen as indirectly identifiable data.
- *Sensitive data* are particularly sensitive in relation to the fundamental rights and freedoms of individuals. They deserve specific protection as their processing could inflict significant risks to the fundamental rights and freedoms of individuals, see Recital 51 of GDPR. These personal data may be processed only when the data processing complies with strict data protection measures. According to GDPR, sensitive personal data include:
  - *Special categories of personal data,* which are about natural persons' racial or ethnic origins, political opinions, religious or philosophical beliefs, trade union memberships, genetic data, biometric data for the purpose of uniquely identifying a natural person, health data, or sex-life or sexual orientation data.
  - *Personal data related to criminal convictions and offences.* Although these are not labelled as a special category (see the previous bullet), they are also seen as sensitive data.

2.4.2 Data protection methods

To start with, it is worthwhile to note that the term privacy is not used in GDPR. This is because privacy is a wide concept and includes also non-data-related aspects like physical privacy (Verheul et al., 2016). Thus, GDPR deals with the concept of (personal) data protection. GDPR globally mentions a number of data protection principles, concepts, methodologies and technologies like purpose limitation, data minimi-

sation, limited storage periods, data quality, and data protection by design/default, as the legal basis for processing personal data, see Article 47(d) of GDPR.

In the following subsections, we focus on the concepts of pseudonymisation and anonymous data[14] as defined or used within GDPR. These two terms are relevant for our study because, on the one hand, anonymous data have an important role in opening justice domain data, as described in Section 2.3. On the other hand, the terms pseudonymisation and anonymisation are widely used within the domain of SDC technologies. In the SDC domain, these terms have a different scope and/or meaning than the definitions of their counterparts in the GDPR domain. It is, therefore, important to clarify their differences, particularly for studies like ours that aim at using SDC technologies for protecting personal data according to GDPR.

Further, note that, data protection according to GDPR is more than just applying SDC technologies and it also includes applying other technological measures such as data encryption and access control. These technologies are not related to SDC and, therefore, their counterpart concepts within GDPR are omitted from our discussion below.

2.4.3 Pseudonymisation

GDPR defines pseudonymisation as follows.

Definition of pseudonymisation: It refers to 'the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person' (see Article 4 of GDPR).

GDPR considers pseudonymisation as an appropriate technological and organisational data protection measure – besides other measures like encryption (see Articles 25 and 32 of GDPR) and access control (see Recital 39 of GDPR) – 'designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects' (see Article 25 of GDPR). Pseudonymisation is seen as a measure which may contribute to data minimisation, i.e., data being 'adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')' (see Article 5(c) of GDPR). Moreover, according to GDPR, pseudonymisation is apt to ensure a level of security appropriate to the risk (see Article 32 of GDPR).

We find that GDPR definition of pseudonymisation covers a large scope of data processing technologies – including data anonymisation in its technological sense (to be defined in the following chapter) – whenever the resulting transformed data can somehow be attributed to an identified or identifiable person. In such cases the transformed data have to be seen as personal data according to GDPR.

[14] For convenience and as we focus on microdata, we use the term 'anonymous data' from this point on to refer to the term 'anonymous information' used in GDPR (see Recital 26 of GDPR).

2.4.4 Anonymous data

In GDPR (and Directive EU 2016/680) the term 'anonymisation' is not used. GDPR, however, defines anonymous information as follows.

Definition of anonymous information: It refers to the 'information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable' (see Recital 26 of GDPR).

In order to determine the possibility of a natural person being identifiable, we must consider 'all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments', see Recital 26 of GDPR. The Working Party 29, (see WP29, 2014) on the identifiability of a natural person mentions: 'importance should be attached to contextual elements: account must be taken of *all the means likely reasonably to be used* for identification by the controller and third parties, paying special attention to what has lately become, in the current state of technology, likely reasonably (given the increase in computational power and tools available)'.

According to GDPR, the term anonymous is used to denote the status of data, thus being anonymous refers to a state and not to a process. In defining this term, GDPR clearly demarcates its scope, i.e., anonymous data fall out of GDPR scope. Pseudonymisation, on the other hand, refers to a process. This is also due to, we suspect, the large scope of the GDPR definition of pseudonymisation, which leaves little room for an independent definition of anonymisation as a process other than, among others, deleting all informational content of the data (and thus reducing data utility enormously).

In GDPR, data being anonymous means that it is anonymous for everybody, even for the data controller. Otherwise, the data have to be seen as pseudonymised (i.e., pseudonymised in GDPR terms). Therefore, for attaining anonymous data, the data controller must take extra measures to make the data not identifiable also for itself.

Finally, in Section 2.3 we concluded that having anonymous data is particularly relevant for opening (justice domain) data. This relevancy is because anonymous data are without, i.e., cannot be associated with, personal data. Here we elaborate further on this conclusion. For processing criminal justice data, the following statement in Directive EU 2016/680 is important: 'In principle, personal data should be processed until they serve the purpose for which they were collected and when they are no longer necessary for that purpose, they should be deleted, unless subsequent processing is foreseen by law and is deemed relevant for a purpose which is not incompatible with the original purpose for processing. Alternatively, the Directive (and GDPR) allow for retention in a form that does not allow identifying the data subjects. Both options should be considered.' (WP29, 2017a). On the other hand, GDPR allows for processing personal data for archiving purposes in the public interest, for scientific or historical research purposes, or for statistical purposes. This processing, however, should be subject to appropriate safeguards by ensuring that technical and organisational measures are in place (particularly with respect to the

principle of data minimisation). In our opinion, opening data can be seen within the scope of personal data processing, which can be fulfilled by 'further processing which does not permit or no longer permits the identification of data subjects', see Article 89(1) of GDPR.[15] Revisiting the conclusion of Section 2.3, we argue that both this GDPR instruction (i.e., 'not permit or no longer permits the identification of data subjects') as well as the statement of Directive EU 2016/680 (i.e., 'retention in a form that makes data subjects unidentifiable') may imply that justice domain data should be made anonymous in the GDPR sense before being opened. In other words, pseudonymisation might not be enough for opening such data because it is potentially possible to reidentify some data subjects by linking the pseudonymised data with other data. We consider this as a topic of future research.

2.4.5 On achieving data anonymity

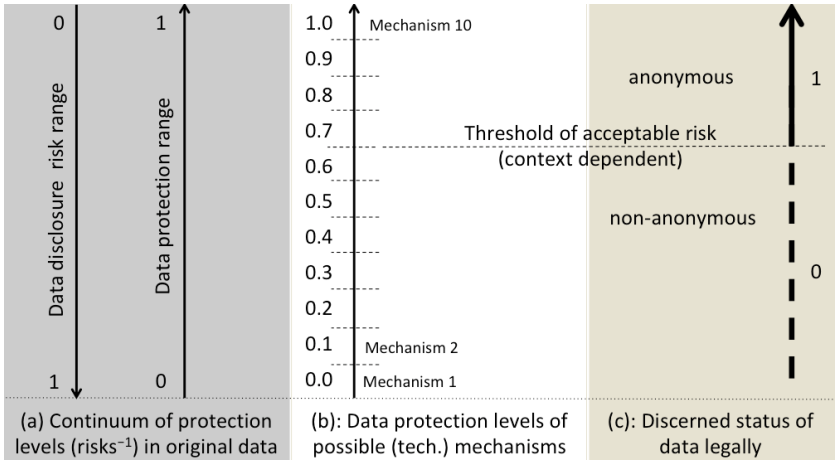
Mitigating data disclosure risks (i.e., the impact severity and likelihood of data disclosures), while maintaining data utility can be enabled by using SDC technologies. When the risks are mitigated such that individuals are no longer identifiable, then the transformed data are anonymous in the GDPR sense and GDPR does not apply to the transformed data. However, there is a risk factor inherent to SDC-based data protection, i.e., data anonymisation in the technological sense (WP29, 2014). According to our understanding and interview results, this means that either:

1 It is not 'truly' possible to attain 'anonymous' data in the GDPR sense because the inherent risks of data disclosures cannot be mitigated,
2 We can have 'anonymous data' in the GDPR sense if the risks are contained within an acceptably negligible level, considering, among others, available technologies, other data sources, and the costs of re-identification at the time of data anonymisation/processing.

The first option, i.e., never having anonymous data, seems for us to be too restrictive and against the GDPR spirit (otherwise the term 'anonymous' should not have been mentioned). The second option, i.e., having anonymous data via applying appropriate safeguards (e.g., SDC technologies and perhaps non-technological procedures) when the corresponding risks are below a certain threshold value, appears to be plausible for us. Figure 3 illustrates this view schematically, where part (a) indicates a continuum range of data protection levels imaginable for the data, part (b) illustrates a countable number of mechanisms that can be used to protect the data incrementally in practice, and part (c) illustrates the range of data protection mechanisms that result in anonymous data, considering their acceptably negligible risks.

[15] See also the National GDPR Implementation Law UAVG, Article 24(a), which – following Article 9(2-j) of GDPR – allows the processing of special categories of personal data in concordance with Article 89(1) of GDPR. Note that here only the processing of special categories of personal data is referred to. Processing of personal data relating to criminal convictions and offences is not mentioned.

Figure 3 An illustration of data protection and anonymous data concepts



Note that data disclosure risks may increase over time, and the currently anonymous data (as we define it by means of the threshold) may become personal data in the future (WP29, 2014). This dynamicity and change of anonymity status are captured by making the value of the threshold for being anonymous dependent on context in Figure 3. This implies that an applied SDC mechanism, which results in an anonymous data set as defined by means of the current threshold, may not do so in the future due to shift of the threshold value upwards in time.

We note that the threshold level does not necessarily get lifted. Also, a correction downwards is thinkable, for instance in case that the identifying background knowledge (e.g., the corresponding data) becomes no longer available. For example, according to GDPR, a necessary condition for the transformed data to be considered as anonymous (i.e., to cross above the threshold level in Figure 3) is that the data are anonymous for everybody including the data controller. Therefore, when a data controller maintains the original (identifying) data, then the transformed data (for example after removing or masking the identifiable data) are not anonymous in the GDPR sense but they are still personal data because the controller can identify individuals from the transformed data with the help of the original data. It is interesting to note that when data controllers erase the original data (due to, for example, maintenance or database clean-up operations), then the corresponding transformed data may become anonymous. In case of achieving anonymity for open data purposes, it is for future research to investigate the necessity and/or consequences of anonymity at the data controller.

On the impact of data controller on anonymous data: 'To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly', see Recital 26 of GDPR. The 'means likely reasonably to be used to determine whether a person is identifiable' are those to be used 'by the controller or by any other person' (WP29, 2014).

Once being decided that data are anonymous in the GDPR sense, GDPR and its data protection principles do not apply anymore (see Recital 26 of GDPR). This holds for the anonymous data and for any processing of the anonymous data like using them for research or statistics (see Recital 26 of GDPR). The currently anonymous data

may become non-anonymous in the future due to, for example, increasing background knowledge or new technological developments as one cannot foresee such advancement at the time of data publishing. Consequently, the transformed data may fall within the scope of GDPR again. This dynamicity, we argue, may be considered as an Achilles heel of GDPR data protection in open data settings. Once the transformed data are opened and published on the Internet, the data can no longer be removed (or only with great difficulty). Therefore, it becomes unrealistic to expect that GDPR can successfully be enforced to the transformed data worldwide at all times (as the transformed data might have reached some regions outside of GDPR jurisdiction).

Although GDPR does not apply to the transformed data rendered as anonymous data in the GDPR sense (see Recital 26 of GDPR), such anonymous data may still have adverse impact on individuals leading to privacy loss (WP29, 2014). We argue that this hurting of individuals may arise when using sensitive data. In such cases Article 8 of ECHR and Article 7 of EU Charter of Fundamental Rights protect the sphere of an individual's private life. The Working Party 29 refers specifically to the case of profiling. As such, 'even though data protection laws may no longer apply to this type of data, the use made of data sets anonymised and released for use by third parties may give rise to a loss of privacy. Special caution is required in handling anonymised information especially whenever such information is used (often in combination with other data) for taking decisions that produce effects (albeit indirectly) on individuals.' (WP29, 2014).

2.5 DPIA and the role of SDC therein

DPIA is required by GDPR (as well as Directive 2016/680). On the other hand, SDC technologies can play an important role within the DPIA process. Therefore, we elaborate here on the role of SDC technologies (thus this study) within the DPIA process. We start with defining a DPIA process in the following.

DPIA process: It is a process 'designed to describe the processing, assess its necessity and proportionality and help manage the risks to the rights and freedoms of natural persons resulting from the processing of personal data by assessing them and determining the measures to address them' (WP29, 2017b).

DPIA is important, as it enables data controllers to define appropriate measures to comply with GDPR requirements. Moreover, DPIAs demonstrate that appropriate measures have been taken to ensure compliance with GDPR (WP29, 2017b). For opening data, DPIA is essential to determine and evaluate the threshold of acceptable risk, to define measures to mitigate data disclosure risks, and to make the data protection process and the decisions taken therein transparent.

2.5.1 When to have a DPIA

Conducting a DPIA is mandatory when data processing is likely to result in a high data disclosure risk. DPIA is not required when the processing is not likely to result in a high risk or when there exists a similar DPIA. Neither is a DPIA mandatory when the processing has been authorised prior to May 2018, has a legal basis, or is in a list of processing operations for which a DPIA is not required (see Article 35(5) of GDPR).

Article 35(3) of GDPR provides three examples of processing operations that are likely to result in high data disclosure risks. The first example of a high-risk concerns systematically and extensively evaluating the personal aspects of natural persons, based on automated processing such as profiling. The second example involves processing special categories of data on a large scale or processing personal data relating to criminal convictions and offences. The third example concerns systematically monitoring a public area on a large scale.

In addition, the Working Party 29 has developed nine criteria to recognise those cases of personal data processing that require conducting a DPIA. These criteria are (see WP29, 2017b; Article 22 and Recital 91 of GDPR):
1 evaluation or scoring, including profiling and predicting;
2 automated-decision making with legal or similar significant effect;
3 systematic monitoring;
4 sensitive data or data of a highly personal nature; this includes special categories of personal data, as well as the personal data related to criminal convictions or offences;
5 data processed on a large scale;
6 matching or combining data sets;
7 data concerning vulnerable data subjects;
8 innovative use or applying new technological or organisational solutions; and
9 when the processing in itself 'prevents data subjects from exercising a right or using a service or a contract'.

Working Party 29 advises when two or more of the abovementioned criteria hold, a data controller should carry out a DPIA. In some cases, a data controller can even consider conducting a DPIA when the intended data processing meets only one of these criteria (WP29, 2017b). In the process of making data sets open, we suspect, conducting a DPIA may be necessary, particularly when criminal justice data are concerned.

2.5.2 Use of SDC within DPIA

Although there are different DPIA methods, four functions can be recognised that are required minimally in a DPIA (WP29, 2017b), namely:
1 describing the envisaged data processing operations and the purposes of the data processing;
2 assessing the necessity and proportionality of the data processing;
3 assessing the risks to the rights and freedoms of data subjects; and
4 envisioning measures to address the risks and demonstrate compliance with GDPR.

A DPIA model[16] has been developed for use by the national government organisations of the Netherlands (e.g., the ministries). This DPIA model has four parts.
- The first part describes the characteristics of the data processing. This part encompasses ten sections to describe, among others, the project and its context, the data processing itself and its goals, and the personal data types being processed.
- The second part, having five sections, reviews the legality of the data processing. This part presents the legal basis and the necessity, finality, proportionality and

[16] In Dutch called 'Model gegevensbeschermingseffectbeoordeling rijksdienst (PIA)'.

subsidiarity principles of the data processing. One section addresses the legal ground(s) for processing special categories of personal data.
- The third part describes and evaluates the privacy risks, in particular, the risks related to (a) the possible negative risks of the data processing on individuals' fundamental rights and freedoms, (b) the origins of these risks, (c) the likelihood of these risks could occur and the impact of these risks on the persons involved.
- The fourth part describes the measures (i.e., technological, organisational and legal measures) needed to mitigate these risks.

SDC technologies can be an important instrument for realising DPIA. They can be relevant for the first part of the DPIA because they are sometimes part of, or required for, the data processing. Moreover, SDC technologies can play a key role in the third and fourth parts, in particular, for developing measures to prevent or minimise data disclosure risks.

We envision that the role of SDC technologies in DPIA is to support (thus not to replace) domain experts in identifying data disclosure risks in (large) data sets and in mitigating those risks appropriately before opening/sharing data. Note that we shall not devise or develop a comprehensive technological-procedural method in this study. Thus, how to exactly embed SDC technologies within DPIA processes in practice is out of the scope of this study.

2.6 Conclusion

In this section we draw a number of conclusions from this chapter, which are widely used and relied upon in the following chapters.

The open data policy of the Dutch government aims at opening data whenever this is compliant to privacy laws and regulations such as GDPR and Wob (as well as Wpg and Wjsg for criminal justice domain data). Briefly reviewing these laws, we concluded that not including personal data plays an important role, if not to say to be a necessary condition, for opening justice domain data sets, particularly those data sets that are related to criminal justice and law enforcement.

We concluded that SDC technologies, on the other hand, are important data protection technologies for enforcing GDPR requirements. Offering a means for making trade-offs between data privacy-utility, SDC technologies are particularly relevant to the GDPR principles of purpose limitation, data minimisation and data accuracy, and they are necessary for realising these principles within DPIA process.

We investigated two GDPR data protection concepts of pseudonymisation and anonymous data. This is because the terms pseudonymisation and anonymisation are two important terms within SDC technologies. Moreover, being anonymous according to GDPR, plays an important role in opening of justice domain data. As these terms are defined differently in the GDPR and technological domains, we noted that, for example, most data anonymisation mechanisms in the technological sense can be regarded as data pseudonymisation mechanisms in the GDPR sense. Further, to determine when data are considered as anonymous, the notion of a context dependent threshold turned to be a useful concept to mark the boundary of data anonymity.

3 Foundations of SDC technologies

In this chapter we present the foundations of SDC technologies (like their definitions, principles and concepts). These foundations include also a number of SDC methods, SDC models and SDC tools, where generally a combination of SDC methods are used to realise an SDC model and a combination of SDC models are realised within an SDC tool. Further, we elaborate on the concepts of data anonymisation and data pseudonymisation, as used in the technological domain. We explain that, for example, data anonymisation in the technological domain means applying SDC technologies to data sets in order to protect personal data. One of the contextual factors that impact SDC-based data protection is the background knowledge available to intruders. In this chapter, we investigate also how the impact of background knowledge can be considered when protecting personal data.

This chapter provides the theoretical foundations needed for answering research questions Q_2 (investigating the main functionalities of available SDC tools for protecting personal data and preserving data utility) and Q_3 (accounting for background knowledge in protecting personal data). To this end, we describe the scope of the data protection considered in this study (Section 3.1), main concepts of SDC-based data protection (Section 3.2), the microdata characteristics that are related to SDC (Section 3.3), SDC methods and models for microdata protection (Section 3.4). Finally, we summarise the main topics discussed in this chapter in Section 3.5.

3.1 Specifying the scope

Data disclosure can occur due to a wide range of undesired phenomena, one of which can be attributed to statistical disclosures, which in turn can be dealt with SDC mechanisms. Further, SDC mechanisms can be applied to various data types. In this section, we shall further specify the scope of the study in regard to the data protection type and the data type considered in this study.

3.1.1 Beyond information security

While personal data protection, in general, and GDPR, in specific, are also concerned with information security mechanisms, in this report we only focus on SDC mechanisms to protect data against statistical data disclosures. Such disclosures occur when the data, which have already been accessed, are analysed illegitimately to derive personal information. These personal data disclosures are example of so-called *unauthorised-use* (Choenni et al., 2015), which can be realised via, for example, information inference.

Example of privacy sensitive information inference: Assume we release a data set about the crimes committed in large cities of the Netherlands, as well as the occupations of the suspects. If there is one specific crime in the data set with the suspect's occupation as 'mayor', then everyone can know who the suspect is with a high probability, as there are a few large cities in the Netherlands and there is a unique person as mayor per city. While the data set might be shared (and accessed) securely, a legitimate (or, of course, illegitimate) data receiver can still infer the identity of the suspect with a high probability, against our intention.

As mentioned above, the intruders in SDC settings have already access to the data either legitimately (i.e., by internal parties) or illegitimately (i.e., by external parties). The intruders in SDC settings, and throughout this report, are defined as follows.

Definition of an intruder in SDC settings: It is a party who has either a legitimate or an illegitimate access to some personal data (i.e., internal intruder or external intruder, respectively), and applies (statistical) data analysis (e.g., data linkage and information inference methods) to derive privacy sensitive information from the accessed data.

3.1.2 Data types

The scope of the study can also be narrowed down, based on the type of data. From the viewpoint of SDC, one can identify the following data types at a high abstraction level (see also De Haan et al., 2011).

- *Structured data* with a predefined and formal structure that specifies, for example, the type of data (e.g., name, date, address, numbers, and currency) and other restrictions on the data like range, number of characters, and categories (e.g., Mr., Ms. or Dr.). Relational data sets and spreadsheets are examples of structured data sets, which can be characterised as tables of rows (i.e., records[17]) and columns (i.e., attributes[18]).
- *Semi-structured data* do not have the formal structure mentioned above. Nevertheless, they have a self-describing structure through tags or markers to separate semantic elements and to form data field hierarchies in the data. XML (Extensible Markup Language), JSON (JavaScript Object Notation), and RDF (Resource Description Framework) are typically used to disseminate semi-structured data sets.
- *Unstructured data*[19] do not have any of the above-specified structures. Such data sets are typically in the form of natural language texts with some dates, numbers, and facts.

In this study we consider only structured data, which constitute a significant part of the administration and registration data gathered and stored within large organisations, particularly those in the justice domain. Structured data, in turn, can be categorised according to the following types.

- *Microdata*, which include the information about respondents, who can be individuals and individual units (like households) in the context of, for example, survey and census data (Hundepool et al., 2012; Willenborg & De Waal, 1996, 2001; El Emam & Malin, 2014). Microdata can be seen as relational tables with some rows, representing individuals, and a number of columns, representing some attributes about those individuals (like their age, gender and occupation).
- *Frequency-tables*, where the value of every cell is the number of contributors to that cell (Hundepool & Wolf, 2011).
- *Quantitative-tables*, where the value of every cell is summation of a continuous attribute over all the contributors to that cell (Hundepool & Wolf, 2011).

---

[17] Also called 'tuples'.

[18] Also called 'variables'.

[19] Some may argue that most so-called unstructured data are structured in one way or another. In this section we use have adopted a definition widely used in the technological domain.

In this study we consider only data disclosures of microdata. This type of data disclosure arises in two situations:

- *interactive information dissemination* through replying to the queries of data consumers about the microdata set; and
- *non-interactive information dissemination* through sharing (a transformation of) the whole microdata set with data consumers.

In this report we consider only the non-interactive dissemination of microdata sets, an example of which is given below.

Example of a microdata set, adopted from adaptation from Fung et al. (2010): The table below indicates a typical relational data set, i.e., a typical microdata set, where every row corresponds to one individual and a column corresponds to an attribute about those individuals.

Name	Job	Gender	Birthdate	Disease	Height (cm)
Bob	Engineer	Male	05/12/1982	Hepatitis	184
Fred	Engineer	Male	03/05/1983	Hepatitis	145
Doug	Lawyer	Male	04/09/1984	HIV	142
Alice	Writer	Female	17/03/1987	Flu	172
Cathy	Writer	Female	04/08/1985	HIV	170
Emily	Dancer	Female	08/01/1987	HIV	169
Gladys	Dancer	Female	28/02/1986	HIV	171

.

3.2 Basic SDC concepts

In this section we describe some basic concepts that are relevant for SDC-based data protection. Note that we do not elaborate on non-SDC-related aspects of these concepts, should the scope of such concepts, like data protection, span beyond the SDC domain.

3.2.1 Intrinsic and extrinsic aspects

Personal data protection goes beyond answering the traditional question of how risky the data by themselves are for release, i.e., to look at just the *intrinsic* characteristics of data. Personal data protection, instead, requires answering a more critical question of how data disclosure might occur (Elliot et al., 2016). To this end, one should consider, among others, the motivation, means, opportunity, and impacts of data disclosure attacks, i.e., one should look at the *extrinsic* characteristics of data. A key extrinsic data characteristic is the availability and use of external data resources (i.e., so-called *background knowledge*) to and by the intruder.

According to Elliot and Dale (1999), the background knowledge available for intruders can range from:

- publicly available information in databases or directories (like electoral registers, telephone directories, trade directories, registers of professional associations);
- personal and informal information due to or via, for example, co-locality (e.g., being neighbours), personal contact, and social media; and
- organisational databases available to government agencies and commercial companies.

Background knowledge also includes any information that is available on the Internet (like websites, social media, knowledge bases, etc.) and other publicly published data sets (like open data and big data), which can be *harvested* and used.

3.2.2 Data anonymisation and pseudonymisation

In Chapter 2 we elaborated upon the concepts of pseudonymisation and anonymous data in the legal domain of GDPR. In this section we explain how the terms data pseudonymisation and data anonymisation are used in the technological SDC domain predominantly. This explanation aims to highlight the differences, existing between the semantics of these concepts in legal and technological domains.

According to Fung et al. (2010), data anonymisation is a Privacy Protecting Data Publishing (PPDP) approach that aims at hiding the identity and/or the sensitive data of data subjects, while retaining sensitive data for the purpose of data analysis. Elliot et al. (2016) characterise data *anonymisation* as a process for ensuring the risk of somebody being identified in the data to become negligible.

There are direct and indirect personal data in microdata sets, which can lead to disclosure incidents due to intrinsic and extrinsic characteristics of the microdata set, respectively. Examples of direct identifiers are names, social security numbers and digitised unique biometrics. Examples of indirect identifiers are (the combination of) gender, postal code and birthdate. Elliot et al. (2016) distinguish four anonymisation types from literature, namely:

- *guaranteed* anonymisation: aims at delivering zero data disclosure risk, regardless of any conditions that we can assume;
- *formal* anonymisation: refers to eliminating direct identifiers in the released data set in order to protect the data set against data intrinsic threats;
- *statistical* anonymisation: strives to use SDC methods to reduce data disclosure risk to an acceptable level, while preserving the utility of the data at an acceptable level;
- *functional* anonymisation: aims at addressing also the contextual factors that affect the disclosure risks of a data set.

The guaranteed anonymisation can be achieved when the anonymised data provide little or no utility (Elliot et al., 2016), therefore, it is out of our scope in this study.

The formal anonymisation, also referred to as data *de-identification*,[20] is done by

- *replacing* direct identifiers with pseudo identifiers;
- *suppressing* (also called masking) all direct identifiers with a certain value (like with three specific characters); or
- *removing* direct identifiers.

Replacing direct identifiers with pseudo identifiers, i.e., the first method mentioned above, is called *pseudonymisation* in the technological domain. More specifically, pseudonymisation is a method whereby direct identifiers are replaced with fictitious names/codes[21] that are unique to individuals but do not directly (i.e., of themselves)

[20] Note that in North America this term is used in a wider sense than the one used here, which is based on its usage in European countries. In North America the term de-identification is used similarly to the term anonymisation used in the other regions including Europe (see El Emam & Malin, 2014).

[21] Note that one individual may have more than one pseudo identifier and a pseudo identifier 'must' refer to one individual (the term 'must' here in practice means the likelihood that a pseudo-identifier refers to different individuals is (extremely) negligible (Bargh et al., 2018).). The latter is called pseudo identifier uniqueness, which

identify individuals (Elliot et al., 2016). Often a mechanism is needed inside a microdata set (or among a set of related microdata sets) to relate different records that belong to the same entity so that data utility can be increased. Via pseudonymisation it is possible to link the records of an individual within a (set of related) data set(s) with the corresponding pseudo identifiers and, thus, without using their direct identifiers. The three de-identification methods mentioned above alone are not enough for data protection and they should almost always be used in conjunction with other anonymisation methods (Elliot et al., 2016).

Statistical anonymisation aims at transforming the data set mainly based on the statistical properties of the data set itself. Functional anonymisation push the frontiers of statistical anonymisation further by adding contextual considerations into the framework. Elliot et al. (2016) mention: 'our view has always been that anonymisation is a heavily context-dependent process and only by considering the data and its environment as a total system (which we call the data situation), can one come to a well-informed decision about whether and what anonymisation is needed'. Moreover, on the other hand, contextual considerations are also determinant of data disclosure risks. These contextual considerations are collectively referred to as *data environment* by Mackey and Elliot (2013), who recognise the following four data environment components:

- *data* to denote the other data present in the data environment (i.e., background knowledge). One should know the relation of the other data to and their overlap with the data in question;
- *agency* to denote those actors (like intruders) who are able to act upon the data in question;
- *governance process* to denote the way for managing users' relationships with the data, like formal governance norms as defined in laws, policies and licenses and users' practices to, for example, avert or accept risks;
- *infrastructure* to denote the way that the infrastructure (like information storage, data security, authentication and data exchange systems) and wider social and economic structures shape the data environment.

Data environment captures, among others, motivations of intruders, impacts of data disclosures, background knowledge, and data governance aspects.

In this study we consider those approaches that yield the abovementioned statistical anonymisation, as the main focus area of the study, or the functional anonymisation, as the marginal focus area of the study. (A detailed study of the latter is for our future research.) Compared to de-identification (i.e., formal anonymisation), statistical and functional anonymisation requires further altering of a data set (i.e., altering other attributes than the direct identifiers) in order to hinder any (statistical) linkage of a released data set with background knowledge.

3.2.3 Impact of background knowledge

Traditionally data protection has been considered stringently. For example, Dalenius mandates that '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' (Fung et al., 2010). Here, *having access to a data set* is the main trigger for possible personal information disclosure. It is, however, shown that it is impos-

---

can be local or global (i.e., a pseudo identifier uniquely referring to one individual within a data set or within a set of related data sets, respectively).

sible to enforce the stringent definition of data protection, as proposed by Dalenius, when the intruder has an arbitrary amount of background knowledge (Dwork, 2006)[22].

Example of the impact of background knowledge, adopted from (Dwork, 2006): Suppose that individuals' age is sensitive information. Further assume that, as background knowledge, an intruder knows Alice's age is five years younger than the average age of American women. If we disseminate a microdata set about the ages of American women, then the intruder can calculate the average age of American women from the released microdata set and infer Alice's age. According to Dalenius' definition, the 'release of the data set' has violated Alice's privacy (even if Alice is not American and thus her record is not in the released data set).

This example shows that background knowledge may have a more dominant role in revealing sensitive personal information than the released data set itself does.

With increasing background knowledge, as we witness nowadays, the discloser risk of personal data increases. It may be fine to publish two data sets individually but publishing both data sets can lead to increased data disclosure risks (as one data set serves as background knowledge for the other).

3.2.4 Data disclosures

A data disclosure can occur directly from a released data set alone or indirectly by linking the released data set with background knowledge. As explained above, these direct and indirect data disclosures can be attributed to intrinsic and extrinsic characteristics of the released data set, respectively. Actually, the extrinsic factors that influence data disclosers include the motivation, means, opportunity, and impacts of data disclosers attacks in addition to availability of background knowledge (Mackey & Elliot, 2013; Elliot & Dale, 1999). A key step to understand (and subsequently deal with) data disclosure attacks is to develop data *disclosure scenarios* (Elliot et al., 2016), whereby the intrinsic and extrinsic characteristics of data disclosures can be captured.

When publishing a (micro) data set, *statistical disclosures* may occur according to two following processes (Elliot et al., 2016):

- *Re-identification* (or identity disclosure), which is a process of attaching an identity to some data (e.g., a record in a microdata set).
- *Attribution* (or attribute disclosure), which is a process of associating a piece of information with a population unit (a person, a family, …). Via attribution we learn something *new* about a person or some persons.

[22] Based on the observation mentioned above, Dwork (2006) proposed the notion of differential privacy, where one aims at making a negligible/small difference between a data subject being in a data set or not. Unlike Dalenius' definition, where having access to a data set is the main trigger for possible data disclosures, in differential privacy being in a data set or not is the main trigger for possible data disclosures. When the difference between the likelihoods of being and not being in a data set is made small enough, the disclosure of personal data can become negligible regardless of (the availability and magnitude of) background knowledge. This independency from the background knowledge makes differential privacy appealing from data protection viewpoint, although its impact on data utility is severe (ref), which makes it unappealing for some applications.

Elliot et al. (2016) mention that formally a statistical disclosure occurs via attribution and not necessarily via reidentification, as illustrated in the following two examples. According to Elliot et al. (2016), reidentification 'typically' results in attributions. Note that reidentification does not always result in attribution, as seen in the following example.

Example of reidentification without attribution: Consider a table of five records, each record having two attributes: the job-function and nationality, where the latter is Dutch for all five records. If the job-function of the first record is 'mayor of Amsterdam', then everybody can identify the person corresponding to the first record (because, being world knowledge, Amsterdam has one mayor) and can know her/his nationality (being Dutch). Learning the nationality of the mayor of Amsterdam is not a statistical disclosure because the derived information is already well known (i.e., not being new).

Attributions, on the other hand, can occur without reidentification, see the following example.

Example of attribution without reidentification: Consider a table of ten records, each record having two attributes of 'job-function' and 'actual income' (and perhaps some other attributes). If the job-function of the first record is 'mayor of a large city in the Netherlands', then everybody can learn how much the mayor of, e.g., Amsterdam earns annually. Assuming that there are five large cities in the Netherlands, that the annual incomes of the corresponding five mayors are the same, and that the annual income of the corresponding mayors is not public knowledge, then releasing the table can cause attribution for the mayor of Amsterdam without associating the identity of the mayor of Amsterdam to the first record.

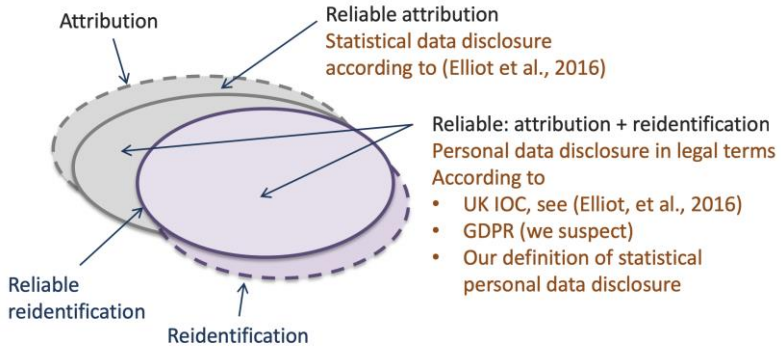
Both reidentification and attribution can occur at various levels of certainty. When the statistical disclosure is not 100% certain, one talks about *inference*. Inference is 'the capability of a user of some data to infer at high degrees of confidence (short of complete certainty) that a particular piece of information is associated with a particular population unit' (Elliot et al., 2016).

In legal terms, reidentification is considered as a breach of privacy rights. In the UK, a reliable attribution (i.e., with enough certainty of degree of confidence) is considered as also reidentification, according to the UK-IOC's interpretation of DPA, as mentioned in Elliot et al. (2016). We suspect that the same holds within GDPR, i.e., a reidentification or a reliable attribution is considered as a breach of privacy rights. Regardless of its legal ramifications, we define statistical data disclosure as follows and leave out the legal ramifications of this definition for future research.

Definition of statistical personal data disclosure (or, in short, statistical data disclosure or data disclosure): It refers to a reidentification or to an attribution that occurs certainly or at a high-enough degree of confidence/certainty.

The scopes of the concepts related to statistical disclosures are illustrated in Figure 4.

Figure 4 An illustration of the scopes of the concepts related to statistical data disclosures (e.g., attribution and reidentification)



3.2.5 Establishing statistical data disclosures

Statistical disclosure of personal data from a data set is established via:

- *Information gain*, where sharing the data set changes the belief of the intruder about an individual's personal information, and
- *Data linkage*, where sharing the data set allows the intruder to associate some *new* data items in the data set with an individual. This association can be probabilistic or deterministic.

Note that information gain and data linkage are not completely different as in both cases the intruder learns more information about the individual with respect to what (s)he already knew (i.e., the existing background knowledge of the intruder).

Example of information gain: Assume that everybody knows that a Dutch person has committed a specific crime from the press/media. Statistically, from the viewpoint of an intruder the likelihood that X, being a specific Dutch person, is the criminal person is almost 1/17,000,000. Further, assume that the intruder infers from a released data set that that criminal person resides in Rotterdam. Then, statistically the likelihood that one specific person in Rotterdam is the criminal personal becomes almost 1/600,000. The information gain can be measured by the difference:

$$- \log_2(1/17{,}000{,}000) - (- \log_2(1/600{,}000)) = \log_2(17/0.6) = 4.82 \text{ bits.}$$

If we further we learn from another data set that the criminal is born in a small city, of which two persons are living in Rotterdam, the information gain with respect to knowing the first released data set becomes:

$$- \log_2(1/600{,}000) - (- \log_2(1/2)) = \log_2(600{,}000/2) = 18.19 \text{ bits.}$$

Thus, the two data sets combined provide 18.19 + 4.82 = 23.01 bits of information.

Example of data linkage: Assume that we release a data set about crimes that have occurred in Rotterdam. The attacker infers from the released data set that Bob, who lives in Rotterdam, has committed the aforementioned crime (deterministic knowledge) or Bob is one of the two persons that could potentially have committed that crime (probabilistic knowledge, i.e., there is a 50% likelihood that Bob is that criminal).

Note that both information gain and probabilistic data linkage rely on empirical probabilities or on believes. While information gain depends on the difference between the empirical probabilities before and after a microdata release, the probabilistic data linkage captures the probability after the data release. Determining the amount of information gain or data linkage certainty that can lead to a statistical personal disclosure can be a situation-specific matter.

Those methods that aim at containing the value of information gain (i.e., making the difference between the intruder's prior and posterior probabilistic beliefs on the sensitive information of a data subject small) try to achieve the *uninformative principle*, which states 'the published table should provide the adversary with little additional information beyond the background knowledge. In other words, there should not be a large difference between the prior and posterior beliefs' (Machanavajjhala et al., 2007).

Four classical types of data disclosure via data linkage are mentioned in Fung et al. (2010), namely: Record linkage, attribute linkage, table linkage, and probabilistic linkage. These data linkage types are illustrated in the following example.

Examples of privacy threat types: The following table is a transformed version of the a microdata set in Section 3.1, where the values of the name attribute are transformed to pseudo identifiers N_i or suppressed by a specific value –.

Assume that the intruder knows Bob and Alice are in the table. If the intruder knows that Bob is a male engineer, then he can infer that Bob has hepatitis. This is an *attribute linkage*, as the intruder cannot know certainly which of row 1 or row 2 belongs to Bob (but both rows have the same value for attribute 'disease'). If the intruder knows Bob's birthdate as well, then he can reidentify Bob in the microdata set, knowing that row 1 corresponds to Bob. This is a *record linkage*.

Name	Job	Gender	Birthdate	Disease	Height (cm)
N_1 (or -)	Engineer	Male	05/12/1982	Hepatitis	184
N_2 (or -)	Engineer	Male	03/05/1983	Hepatitis	145
N_3 (or -)	Lawyer	Male	04/09/1984	HIV	142
N_4 (or -)	Writer	Female	17/03/1987	Flu	172
N_5 (or -)	Writer	Female	04/08/1985	HIV	170
N_6 (or -)	Dancer	Female	08/01/1987	HIV	169
N_7 (or -)	Dancer	Female	28/02/1986	HIV	171

Now let's assume that the table includes all patients in the town, who were sick within the last week. Further assume the intruder knows that Alice lives in that town, the town has 4 female artists, and Alice is an artist. Then the intruder can infer that Alice's record must be in this table (thus, she was ill last week) and her record should be one of the last 4 records in the table. This is a *table linkage*. Further, assume that the intruder knows Alice's birth year is 1987. Then the intruder can infer that Alice is HIV positive with a 50% chance. This is a *probabilistic linkage*.

3.3 Characteristics of microdata

In this section we start with describing the types of attributes in microdata sets from the viewpoint of SDC technologies in Subsection 3.3.1. These types refer to the extent to which attributes (uniquely) identify individuals and/or to the sensitivity of attributes. Subsequently we explain how to assign these attribute types to attributes in a typical microdata set in Subsection 3.3.2.

3.3.1 Attribute types

A relational microdata set can be specified by its records and attributes. Let's consider an original microdata set $D_N(A_1, A_2, \ldots, A_M)$ with N records and M attributes A_1, A_2, …, A_M. Further, we assume that every record corresponds to an individual, called data subject.[23] In the literature in the area of, for example PPDP (Fung et al., 2010) and SDC (Elliot et al., 2016), the set of attributes $\{A_1, A_2, \ldots, A_M\}$ are divided into four disjoint sets called: Explicit identifiers, quasi identifiers, sensitive attributes, and non-sensitive attributes[24].

Explicit Identifiers (EIDs), also called 'direct identifiers' in Elliot et al. (2016), refer to the set of attributes in the original data set D that structurally and on their own could uniquely identify an individual, i.e., a data subject. Examples of explicit identifiers are a data subject's name, home address and unique personal numbers (like the 'social security number', 'national health service number', 'voter card identification number', or 'permanent account number'). Often the set of explicit identifiers is removed (i.e., filtered), replaced with an unrecognisable value (i.e., masked/suppressed), or replaced with a unique and unrecognisable value (i.e., pseudonymised), as a first step of data disclosure control.

Quasi Identifiers (QIDs) refer to the set of attributes in the original data set D that could 'potentially' identify individuals, i.e., data subjects. This identification is achieved through using the QID set to link the records of data set D with the other data sets and knowledge bases wherein both explicit identifiers and QIDs are present for some individuals. The QIDs in data set D, therefore, represent/capture the background knowledge that intruders have with respect to data set D.

Inspired by (Elliot et al., 2016), we distinguish the following QID types.

- *Indirect identifiers* refer to any attribute or set of attributes that are not structurally unique but are likely to become unique for at least some individuals in the data set and in the population.[25] The combination of attributes birthdate, postal code and gender can be an example of indirect identifiers, as shown in Sweeney (2000, 2002a).
- *Indirect identifier values* refer to the case where certain values of some attributes may be rare in the population (like a widower of 18 years old) whilst the corresponding attributes cannot be regarded as obvious identifiers (like 'age' and 'marital status' in this case). As these attribute values are rare in the population, someone from the communities of the data subjects can identify these individuals easily due to having demographics knowledge (Elliot et al., 2016). Another exam-

[23] 'Data subject' is also called 'record owner'. We adopt the former term throughout the report.

[24] Note that also GDPR discerns similar types of personal data in terms of identifiability and data types, see Subsection 2.4.1.

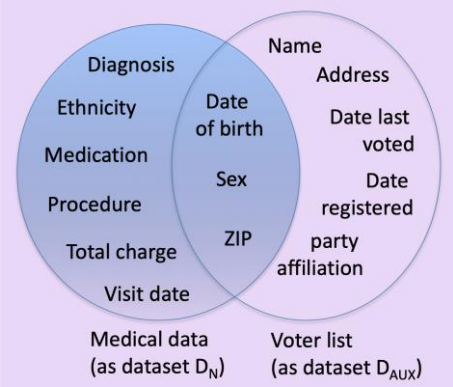
[25] In the following chapter we define the concept of population (or population data set) clearly.

ple is the case of 'mayor of Amsterdam' as the specific values of attributes 'job' and 'work place' (Hundepool et al., 2014).
- *Key variables* refer to those attributes in data set D that are specific to data set D and a data intruder has some auxiliary information about (together with the explicit identifiers of the corresponding data subjects). For example, the combination of attributes job, gender and education might identify data subjects in a specific data warehouse setting uniquely.

Note that being *specific to a setting or being generic for all settings* makes the difference between key variables and indirect identifiers. Key variables are specific to a particular scenario or a specific combination of data sets (like in a data warehouse, where the combination of attributes job, gender and education might identify some data subjects in that data warehouse uniquely). Indirect identifiers, however, refer to a generic set of attributes that could enable identification of some individuals in any scenario or any combination of data sets, like attributes birthdate, postal code and gender as found out in Sweeney (2000, 2002a).

Example of data linkage via QIDs: Sweeny (2002a) used an assumingly anonymous patient data set that was made available to researchers. The data set (GIC) included about one hundred attributes (see the left circle in Figure 5 some of these attributes) for about 135,000 state employees and their families in Massachusetts. The released attributes included patients' postal code (ZIP), birth date and gender. As background knowledge, Sweeney obtained the voter registration list for Cambridge, Massachusetts. The right circle in Figure 5 shows some attributes in this data set, which included the name attribute as well as the postal code, birth date, and gender of the voters. She used the attributes postal code, birth date, and gender as QIDs and could link the names to the medical information. In this way, she inferred the diagnosis, procedures, and medications of (famous) individuals. 'For example, William Weld was governor of Massachusetts at that time and his medical records were in the GIC data. Governor Weld lived in Cambridge, Massachusetts. According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit postal code' (Sweeney, 2002a). For a similar example from the justice domain see Choenni et al. (2010).

Figure 5 An example of using QID to link an anonymous data set with auxiliary data set[a]



[a] See Sweeney (2002a).

Sensitive Attributes (SATs) refer to those attributes that capture privacy-sensitive information about data subjects who (possibly) do not want to disclose them. Ex-

amples of sensitive attributes are disease, salary, loan, disability status, and crime type. These sensitive attributes are sometimes important for data consumers for data analytics purposes. Therefore, SATs are (often) shared without any (or with minor) alteration. Unlike QIDs, SATs are not known outside of the original data set D and, therefore, they cannot be characterised as background knowledge of intruders.

Example of SATs: As mentioned in Subsection 2.4.1, GDPR defines some data as sensitive data, as their processing could inflict significant risks to the fundamental rights and freedoms of individuals (Recital 51 of GDPR). The sensitive personal data include:

Special categories of personal data that are about natural persons' racial or ethnic origins, political opinions, religious or philosophical beliefs, trade union memberships, genetic data, biometric data for the purpose of uniquely identifying a natural person, health data, or sex-life or sexual orientation data.

Personal data related to criminal convictions and offences. Although these are not labelled as a special category, they are also seen as sensitive data.

Non-sensitive Attributes (NATs) refer to all the other attributes that are not direct-identifying, quasi-identifying or sensitive attributes.

Note that there is no universal definition of which attribute is (non-)sensitive. Some legal frameworks have specified some attributes as sensitive, for example, the UK's DPA considers racial or ethnic origin, political opinions, religious beliefs, trade union membership, physical or mental health or condition, sexual life, and some aspects of criminal proceedings as 'sensitive personal data' (Elliot et al., 2016). For another example, see the case GDPR (Subsection 2.4.1).

Furthermore, the *situational context*[26] and personal preferences (of data subjects) influence an attribute in being considered as sensitive or not. In some situations, the attributes related to one's income, wealth, credit record and financial deals can be sensitive. Religion might be considered as a sensitive attribute in some countries and as non-sensitive in others.

3.3.2 Attribute mapping

Given a microdata set $D_N(A_1, A_2, \ldots, A_M)$ one should map attribute A_m, where m=1, …, M, to one[27] of the types EID, QID, SAT and NAT. Defining the EIDs is straightforward and it is based on the intrinsic aspects of data set $D_N(A_1, A_2, \ldots, A_M)$. Defining NATs is also trivial, i.e., once the other three types are defined, the remaining attributes can be regarded as non-sensitive. In this subsection we focus on the non-trivial task of defining QIDs.

Assume that microdata set D'_N is a transformation of microdata set D_N, in which the attributes of EIDs are removed, suppressed, or pseudonymised (as illustrated on top of Figure 6). On the other hand, the identities of the subjects can somehow be avail-

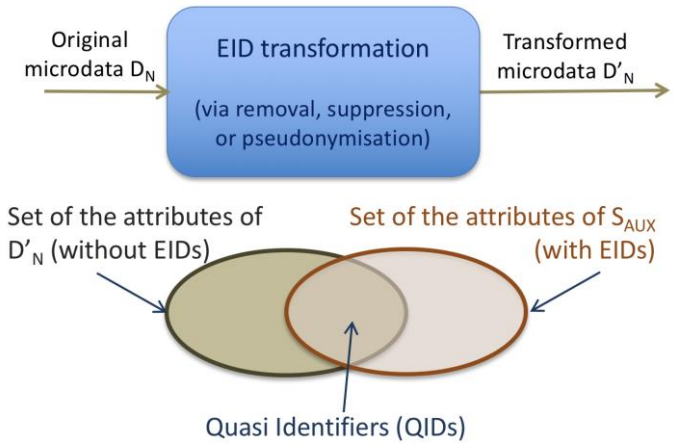
[26] In Mackey and Elliot (2013) the term 'data environment' is used to refer to these contextual factors, which we are going to elaborate upon in the following sections.

[27] We have not seen a case where an attribute having more one type so far. We do not, nevertheless, rule out a case where an attribute is both QID and SAT.

able to the intruder(s) through *auxiliary information sources* as background knowledge. Such auxiliary information sources encompass some QIDs of types 'indirect identifiers', 'key variables', or both (as mentioned in Subsection 3.3.1), and the EIDs of the corresponding data subjects. Elliot et al. (2016) mention four types of auxiliary information sources:

1 Those data sets that contain the same information for the same (or sufficiently similar) data subjects. For example, consider the case where the original data set D_N corresponding to the published data set D'_N is available to the data controller.
2 Those information sources that are publicly available via open data, in public registers, or on social media[28].
3 The information obtained from proximity knowledge. For example, the intruder obtains household information from acquaintances, an estate agent, or via own physical observation.
4 The information that the intruder obtains through personal knowledge. For example, he hears from the data subject's neighbours or colleagues.

Figure 6 An illustration of QID mapping



Considering a given disclosure scenario, one should identify all possible auxiliary information sources as far as possible. Subsequently, as illustrated on the bottom of Figure 6, one should mark those non-EID attributes that are present in the transformed data set D'_N and in those auxiliary information sources as QIDs. In other words, let set $\{A_1, A_2, \ldots, A_M\}$ be the set of those attributes of data set D_N that are not EIDs. Further, let $\{S_1, S_2, \ldots, S_{M'}\}$ be the set of attributes of an auxiliary information source S_{AUX} which also include (some) EIDs corresponding to (some of) the records in data set D_N. The set of QIDs is the intersection of sets $\{A_1, A_2, \ldots, A_M\}$ and $\{S_1, S_2, \ldots, S_{M'}\}$, as illustrated in Figure 6.

Intruders can use these QIDs as a bridge to link the EIDs available in the auxiliary information sources to the corresponding records in the transformed data set D'_N. Thereby, an intruder can reidentify some data subjects in data set D'_N through this linking.

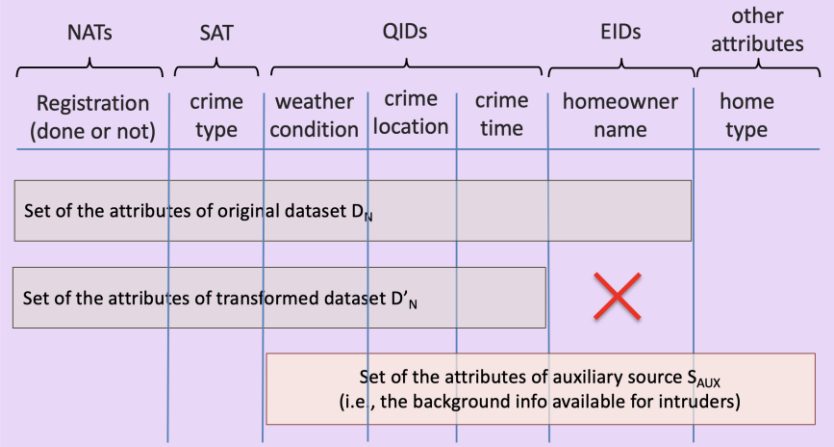
Note that a QID may seem innocent (like weather condition) at first sight, i.e., if it is considered intrinsically with respect to a data set. But such an attribute can enable

[28] In El Emam and Dankar (2008) such auxiliary information sources are called 'dentification databases' (or as we call them as identification data sets).

data linkage, when it also appears in an auxiliary information source with EIDs as illustrated in the following example.[29]

Example of attribute mapping: Figure 7 shows the set of attributes in the original data set D_N and those in the transformed data set D'_N that is without explicit identifiers. The QIDs are weather condition, crime location and crime time that can be found in auxiliary sources S_{AUX} together with the identities of the data subjects.

Figure 7 An example of attribute mapping



Considering the background knowledge available in auxiliary information sources is an important step of functional data anonymisation (see Subsection 3.2.2, where one considers the contextual conditions and constraints surrounding the data sharing process, next to the data set to be anonymised).

3.4 SDC technologies

In this section we describe main SDC (or anonymisation[30]) technologies for anonymisation of microdata sets. These technologies are categorised in three hierarchical levels of data SDC methods (Subsection 3.4.1), SDC models (Subsection 3.4.2) and SDC tools (Subsection 3.4.3).

[29] Further, even if such an attribute (i.e., weather condition) does not appear directly in auxiliary information sources with EIDs, but its values are highly correlated with some QIDs (like crime location and crime time), then data linkage may occur. For example, when only crime location and crime time are properly protected because they are QIDs, then knowledge about exceptional weather conditions (i.e., also background knowledge) can be used to infer the crime location and crime time (thus the values of those QIDs) for the corresponding records. Background information in the form of such exceptional attribute values should be accounted for in a different way than the QID approach discussed in this section.

[30] Note that the term 'anonymisation' here is used in a technological sense. Accordingly, the SDC technologies, SDC methods, SDC models and SDC tools used in this section may alternatively be called as anonymisation technologies, anonymisation methods, anonymisation models and anonymisation tools, respectively.

3.4.1 SDC methods

In this subsection we provide an overview of a number of core components (i.e., main methods) for transforming the attributes and records of microdata sets. Some of these methods have already been presented in previous sections, nevertheless we briefly restate them here for a self-contained overview. Note that this overview is by no means exhaustive and particularly includes those methods that are deployed in the studied SDC tools.

Removal is a method whereby EIDs (and other unnecessary attributes or attribute values) are omitted from the data set.

Suppression is a method whereby some values are replaced with a specific value in order to clearly indicate that the replaced values are present but are not disclosed (Fung et al., 2010). Suppression schemes include

- *Record* suppression whereby an entire record is suppressed (for the references see Fung et al., 2010);
- *Value* suppression whereby every instance of a given value in a table is suppressed (for the references see Fung et al. 2010);
- *Cell* suppression or local suppression whereby some instances of a given value in a table are suppressed (for the references see Fung et al., 2010).

Sometimes it is necessary to maintain records in a form that some statistical properties of the data are preserved. Therefore, instead of suppressing records with semantically irrelevant values, we can replace them with statistically meaningful values to ensure certain statistical properties of the data set (for example, to preserve the statistical averages of some attributes). See, for example, the permutation and perturbation methods mentioned in the following.

Pseudonymisation is a method whereby direct identifiers are replaced with fictitious values (i.e., pseudo identifiers) that uniquely specify or refer to individual records. Referencing to individuals can be locally unique (i.e., in a data set) or globally unique (i.e., in a set of related data sets). For example, in a data set which associates locations to individual, one may generate different pseudo identifiers for persons (e.g., one per day). In this way, a disclosure of a person's pseudo identity on some day, remains local to that day and does not propagate to the whole data set.

Note that in an SDC process the methods of removal, suppression and pseudonymisation make parts of the data (often the EIDs[31]) unavailable to the data consumers and intruders. These methods are almost always used in conjunction with the anonymisation methods, described below so that the other attributes (i.e., QIDs and SATs) can also be protected.

Generalisation[32] is a method whereby some values of an attribute are replaced with a parent value in the taxonomy-tree of the attribute (Fung et al., 2010). For example, the year values of the attribute age are changed to the corresponding decade values.

There are a number of schemes proposed in literature for generalisation. Fung et al. (2010) mention the following five generalisation schemes.

---

[31] Being the target of pseudonymisation and most often removal.

[32] The reverse operation of generalisation called specialisation (Fung et al., 2010).

- *Full domain* generalisation, whereby all values of an attribute are generalised to the same level at the attribute's taxonomy-tree (for the references see Fung et al., 2010).
- *Sub-tree* generalisation, whereby either all child values at a non-leaf node of an attribute's taxonomy-tree are generalised, or none are generalised (for the references see Fung et al., 2010).
- *Sibling* generalisation, whereby, unlike the sub-tree generalisation, not all siblings at a non-leaf node of an attribute's taxonomy-tree are generalised. In other words, only those siblings that are needed are generalised (for the reference see Fung et al., 2010).
- *Cell* generalisation, whereby the value(s) of an attribute are generalised for some instances, i.e., for some records and not for all records (for the references see Fung et al., 2010).
- *Multi-dimensional* generalisation, whereby the generalisation is carried out in multiple dimensions/attributes simultaneously, instead of per dimension/attribute as the way is done in the abovementioned schemes (for the references see Fung et al., 2010).

Note that cell generalisation is also called *local recoding* because it is not applied to all records. The other schemes are called *global recoding*, where the values of an attribute are generalised for all records. Approximately, one can claim that the generalisation schemes mentioned above produce less distortion (i.e., more data utility) in descending order of appearance on the list above.

Often considered as a sort of generalisation method, *top/bottom coding* is used to suppress the extreme values of a (typically) numerical attribute to a maximum or minimum value, or to some alphanumeric values.

Permutation is a method whereby one partitions a set of records into *groups*[33] and then replaces or *shuffles* their sensitive values within each group. In this way, permutation disassociates the relationship between a QID and a numerical SAT (for the reference see Fung et al., 2010). Numerical *micro aggregation* is a specific permutation method, whereby at first the groups of a certain size are made and then the values of the numerical attributes within every group are replaced with the mean value of those values of the group members. Actually, in micro aggregation the attribute values in a group are not shuffled but are replaced with their mean values in the group.

Perturbation is a method whereby one replaces the original values with some other (synthetic) values in a way that there is no significant difference between the statistical information derived from the perturbed data set and the statistical information derived from the original data set. Unlike permutation, perturbation applies to all records and does not divide them into groups. According to Fung et al. (2010) three perturbation schemes are:
- *Noise addition*[34] whereby a random value r drawn from some random distribution is added to the original sensitive value s (like the amount of income), resulting in

[33] In our opinion, grouping of records is the distinctive aspect of permutation, where values are adapted within groups of records, compared to perturbation, where values are adapted in all records.

[34] Note that technological methods such in micro-aggregation, top/bottom coding, and per group rounding up/down can be seen as noise addition, i.e., adding the so-called quantisation noise. We do not adopt the concept of noise addition for such methods in this report.

value s + r. The degree of preserving privacy depends on how closely one can estimate s from s + r (for the references see Fung et al., 2010).

- *Data swapping* whereby one exchanges the values of SATs of individual records. Data swapping aims at maintaining the low-order frequency counts (or marginal) for statistical analysis. It can be used for protecting both numerical and categorical attributes (for the references see Fung et al., 2010). For example, there are the following data swapping schemes:
  - *Rank swapping* where the possible attribute-values of an attribute A_m are set in an ascending order in, for example, list $OL(A_m)$. Then each attribute value $v \in OL(A_m)$ is swapped with another value $u \in OL(A_m)$ such that u is randomly chosen within a restricted range (i.e., the p% range of v) within $OL(A_m)$. This scheme was originally proposed for ordinal attributes and later applied to any numerical attributes (see also Hundepool et al., 2014, p. 15).
  - *PRAM* (Post RAndomisation Method) where the values of a categorical attribute are swapped based on a predetermined probability distribution (i.e., attribute value i is changed to attribute value j with probability $p_{i,j}$, where $\Sigma_j\, p_{i,j} = 1$ for all i's). This value swapping is done independently for every record. As the enforced probability distribution is known, one can estimate the (statistical) characteristics of the attribute in the original data set from the perturbed data set (Hundepool et al., 2014).
- *Synthetic data generation* (or condensation) whereby one builds a statistical model from the original data, creates sample data items from the model (resulting in the synthetic data set), and publishes the synthetic data set instead of the original data set.

Anatomisation is a method whereby one defines the joint frequency values of SATs per *Equivalent Class* (EC). An EC refers to every pattern of values of QIDs, see the example below. As the result of anatomisation, two tables are produced, namely:

- *Quasi-Identifier Table* (QIT) that includes the ECs of the QIDs, extended with a new attribute that holds the group IDs of those ECs;
- *Sensitive Table* (ST) that, per group ID created above, includes the joint frequencies of the possible values of the SATs.

Equivalent Class: An EC refers to every pattern of values of QIDs. For example, consider the case where there are two QIDs 'gender' and 'marital status'. Attribute 'gender' can assume one of the values 'male' and 'female'. Attribute 'marital status' can take one of the values 'married' and 'single'. Then there are 4 possible EC's, namely: (a) male, single, (b) male, married, (c) female, single, and (d) female, married.

Note that applying these SDC methods to preserve privacy reduces data utility inherently. Each method, nevertheless, impacts data utility differently and in a particular way. Therefore, an SDC method can be chosen depending on the forms of data utility that need to be retained.

3.4.2 SDC models

In this subsection we provide an overview of the main SDC models, which are realised by employing the aforementioned SDC methods. This overview, which is inspired by the list in Almasi, Siddiqui and Mohammed (2016), is by no means exhaustive and covers those SDC models that are widely deployed in the SDC tools investigated in this study.

The *k-anonymity* model requires that for every possible combination of the values of QIDs (i.e., for every EC), there are at least k records in the transformed microdata set. Such a microdata set is called *k*-anonymous. In this way the *k*-anonymity model aims at preventing record linkage through QIDs. Therefore, the probability of linking a data subject, whose data is known to be in the microdata set, to a specific record (and thus to the SATs in that record) is 1/k (see Sweeney, 2002a, 2002b).

Example of the *k*-anonymity model: Let D_7 be a microdata set, corresponding to the patient records in a hospital.

Original microdata set D_7

Name	Job	Gender	Birthdate	Disease	Height (cm)
Bob	Engineer	Male	05/12/1982	Hepatitis	184
Fred	Engineer	Male	03/05/1983	Hepatitis	145
Doug	Lawyer	Male	04/09/1984	HIV	142
Alice	Writer	Female	17/03/1987	Flu	172
Cathy	Writer	Female	04/08/1985	HIV	170
Emily	Dancer	Female	08/01/1987	HIV	169
Gladys	Dancer	Female	28/02/1986	HIV	171

Let us define attributes 'job', 'gender' and 'birthdate' as QIDs. To obtain transformed microdata set D'_7, we generalise attributes 'job' and 'birthdate'; and suppress the EID 'name'. Considering QIDs 'job', 'gender' and 'birthdate', we have 3-anonymity in the transformed microdata set D'_7 (with k=3 anonymity) as seen in the following. The resulting microdata set contains two ECs of (professional, male, 1980-1984) and (artist, female, 1985-1989).

Transformed microdata set D'_7 (with k=3 anonymity)

Name	Job	Gender	Birthdate	Disease	Height (cm)
***	Professional	Male	1980-1984	Hepatitis	184
***	Professional	Male	1980-1984	Hepatitis	145
***	Professional	Male	1980-1984	HIV	142
***	Artist	Female	1985-1989	Flu	172
***	Artist	Female	1985-1989	HIV	170
***	Artist	Female	1985-1989	HIV	169
***	Artist	Female	1985-1989	HIV	171

The *l-diversity* model aims at preventing the disclosure of sensitive attributes in the *k*-anonymity model. Although *k*-anonymity guarantees that there are at least k records in every EC, it may be possible that the value of a SAT for all the records in that EC is the same. Therefore, the intruder can learn the value of that SAT for all the corresponding data subjects. The *l*-diversity model aims at preventing such attribute linkage through QIDs, by requiring that the records in every EC have at least l *well-represented* values for each of their SATs (Machanavajjhala et al., 2006; 2007).

Example of the *l*-diversity model: Consider the transformed microdata set D'_7 (with k=3 anonymity) in the previous example. Given three QIDs of 'job', 'gender' and 'birthdate', there are two ECs of EC_1 = (professional, male, 1980-1984) and EC_2 = (artist, female, 1985-1989). Further assume that attribute 'disease' is a SAT. Considering EC_1 the SAT of disease assumes two values of 'hepatitis' and 'HIV'; and considering EC_2 the SAT of disease assumes two values of 'flu' and 'HIV'. Consequently, the intruder would have uncertainty between two values of the SAT value, should (s)he know the EC to which a victim (i.e., a data record) belongs. In this case, we say that D'_7 has distinct 2-diversity as for all ECs (i.e., EC_1 and EC_2) the SAT assumes at least two distinct values.

There are several ways proposed in literature to operationalise the requirement of well-presented SAT values. Some important variants of the *l*-diversity model are:

- *Distinct l*-diversity model, which ensures that there are at least l distinct values of every SAT in the records of every EC. (Note that if there are at least l distinct values of the SAT in every EC, then there is *k*-anonymity with k≥l.)
- *Entropy l*-diversity model, which ensures that the distribution of the frequencies of the values of each SAT in the records of every EC is close to the uniform distribution (i.e., the values of the SATs occur uniformly within every EC).[35] According to this definition, a sensitive attribute that has a more even distribution of its values in ECs, results in a larger value of l.
- *Recursive* (c, l)-diversity ensures that the most frequent value of a SAT does not occur too frequently, and the least frequent values occur adequately.[36] In this way, if the intruder excludes some possible sensitive values, the remaining values still remain hard to infer.

The *t-closeness* model ensures that the distribution of the values of a SAT in every EC is close to the distribution of the values of the SAT in the whole data set (for all ECs together). In this way, one deals with the skewed distribution of sensitive attribute values in ECs. In order to measure the closeness of the distributions in the *t*-closeness model there are, for example, the Earth Mover Distance (EMD) and Kullback-Leibler divergence function proposed in (Li et al., 2007). The *t*-closeness model requires the values of closeness of the distributions to be within (at least) t for all ECs (Li et al., 2007).

---

[35] Let f(i, j) be the number of the SAT with value j in the EC i with N_i records. Then, p(i, j) = f(i, j) / N_i is the empirical probability of the SAT value j in EC i. Entropy *l*-diversity means that the maximum value of l that is smaller than or equal to $\min_i$ H(p(i, j)), where H(.) is the entropy function.

[36] Let f(i, j) be the number of the SAT with value j in the EC i with N_i records. Then, p(i, j) = f(i, j) / N_i is the empirical probability of the SAT value j in EC i. Let f'(i, j) be the representation of f(i, j) in descending order in EC i, i.e., the frequency f'(i, j) denotes the j^{th} most frequent sensitive attribute value in EC i. Assume M is the number of attribute values for the SAT. Given l from the *l*-diversity model and a constant value c, the recursive (c, l)-diversity for EC i means that $f'(i, 1) < c \sum_{j=l}^{M} f'(i, j)$, i.e., f '(i, 1) which is the most-frequent attribute value, is smaller than c times the sum of the M-l+1 least-frequent attribute values f '(i, j), for j=l, …, M. If this holds for all i's, then the data set has recursive (c, l)-diversity.

Example from Fung et al. (2010): Consider a patient microdata set where 95% and 5% of the records have Flu and HIV, respectively. Suppose that 50% of the records in an EC are Flu and the rest are HIV. Therefore, both (i.e., the original data set and the EC) satisfies 2-diversity. However, the EC presents a probabilistic attribution threat because for any data subject in the EC the intruder could infer that the subject is HIV positive with 50% confidence, compared to 5% in the original data set.

The $\hat{k}$-map model (El Emam and Dankar, 2008) assumes that microdata set D_N is a sample of a so-called population microdata set P_L, where $N \leq L$. This sampling is defined with respect to QIDs. In other words, considering QIDs, D_N contains a subset of the records of P_L.

Example of sample and population microdata sets: Let P_7 be a population microdata set, containing 7 patient records of a hospital (note that P_7 is the same as D_7 mentioned in the previous example box).

Population microdata set P_7

Name	Job	Gender	Birthdate	Disease	Height (cm)
Bob	Engineer	Male	05/12/1982	Hepatitis	184
Fred	Engineer	Male	03/05/1983	Hepatitis	145
Doug	Lawyer	Male	04/09/1984	HIV	142
Alice	Writer	Female	17/03/1987	Flu	172
Cathy	Writer	Female	04/08/1985	HIV	170
Emily	Dancer	Female	08/01/1987	HIV	169
Gladys	Dancer	Female	28/02/1986	HIV	171

Microdata set D_3 with three records, see the following, is a sample microdata set of the population microdata set P_7, when considering the QIDs of 'job', 'gender' and 'birthdate'. Sample microdata set D_3 is recorded by the accounting department of the hospital for those patients who were released from the hospital a week ego.

Sample microdata set D_3

Name	Job	Gender	Birthdate	Treatment fee
Bob	Engineer	Male	05/12/1982	550 €
Cathy	Writer	Female	04/08/1985	2.300 €
Gladys	Dancer	Female	28/02/1986	1,500 €

Note that the sampling is defined without considering the EID of 'name' and also that the 'treatment fee' attribute is known only to the accounting department (thus being absent in the population microdata set P_7).

The assumption in $\hat{k}$-map is that the data controller has access to also the population microdata set P_L and therefore (s)he can apply the $\hat{k}$-anonymity method to the population microdata set directly to obtain the transformed population microdata set, denoted by P'_L (with $\hat{k}$-anonymity). Note that the $\hat{k}$-anonymity model is applied to P_L by considering the QIDs that are defined based on microdata sets D_N and P_L. Then, looking at the resulting P'_L (with $\hat{k}$-anonymity), the data controller maintains those ECs and the corresponding data records of sample microdata set D_N that also appear in P'_L (with $\hat{k}$-anonymity). The resulting $\hat{k}$-map sample microdata set is denoted by D'_N (with $\hat{k}$-map). This SDC model is illustrated in the following example.

Examples of *k*-map: Considering QIDs 'job', 'gender' and 'birthdate', we generalise attributes 'job' and 'birthdate'; and suppress the EID 'name' in the population microdata P_7 in the previous example box to obtain transformed population microdata set P'_7. Considering the three QIDs, we have 3-anonymity in transformed population microdata set P'_7, denoted by P'_7 (with k=3 anonymity), with two ECs of (professional, male, 1980-1984) and (artist, female, 1985-1989).

Transformed population microdata set P'_7 (with k=3 anonymity)

Name	Job	Gender	Birthdate	Disease	Height (cm)
***	Professional	Male	1980-1984	Hepatitis	184
***	Professional	Male	1980-1984	Hepatitis	145
***	Professional	Male	1980-1984	HIV	142
***	Artist	Female	1985-1989	Flu	172
***	Artist	Female	1985-1989	HIV	170
***	Artist	Female	1985-1989	HIV	169
***	Artist	Female	1985-1989	HIV	171

If we want to have 4-anonymity in population microdata set P'_7, the EC of (professional, male, 1980-1984) should be suppressed to obtain transformed population microdata set P'_7 (with k=4 anonymity).

Transformed population microdata set P'_7 (with k=4 anonymity)

Name	Job	Gender	Birthdate	Disease	Height (cm)
***	***	***	***	***	***
***	***	***	***	***	***
***	***	***	***	***	***
***	Artist	Female	1985-1989	Flu	172
***	Artist	Female	1985-1989	HIV	170
***	Artist	Female	1985-1989	HIV	169
***	Artist	Female	1985-1989	HIV	171

For 3-map, we should look at the sample microdata set D_3 (see the previous example box) and P'_7 (with k=3 anonymity). To obtain the transformed sample microdata set D'_3 (with k=3 map), we maintain those records of D_3 that also appear in transformed population microdata set P'_7 (with k=3 anonymity). Moreover, we apply the EC's defined for microdata set P'_7 (with k=3 anonymity) to obtain the transformed sample microdata set D'_3 (with k=3 map) as follows (where also the EID name is suppressed).

Transformed sample microdata set D'_3 (with k=3 map)

Name	Job	Gender	Birthdate	Treatment fee
***	Professional	Male	1980-1985	550 €
***	Artist	Female	1985-1989	2.300 €
***	Artist	Female	1985-1989	1,500 €

For 4-map, we should look again at sample microdata set D_3 (see the previous example box) and P'_7 (with k=4 anonymity). To obtain the transformed sample microdata set D'_3 (with k=4 map), we maintain those records of D_3 that also appear in the transformed population microdata set P'_7 (with k=4 anonymity). Moreover, we apply the EC's defined for the transformed population microdata set P'_7 (with k=4 anonymity) to obtain the transformed sample microdata set D'_3 (with k=4 map) as follows (where also the EID name is suppressed).

Transformed sample microdata set D'_3 (with k=4 map)

Name	Job	Gender	Birthdate	Treatment fee
***	***	***	***	***
***	Artist	Female	1985-1989	2.300 €
***	Artist	Female	1985-1989	1,500 €

In this way, the data controller creates sample data set D'_N with $\acute{k}$-map from sample data set D_N. In $\acute{k}$-map, some ECs in the resulting data set D'_N may have fewer than k records, while in $\acute{k}$-anonymity every EC should have had at least k records. This is acceptable because $\acute{k}$-anonymity is preserved in the population data set. As a result, the extent of information loss can be reduced significantly when using $\acute{k}$-map instead of $\acute{k}$-anonymity (El Emam and Dankar, 2008). In the $\acute{k}$-map model, actually, the structure of $\acute{k}$-anonymity is mapped from the population data set P_L to the sample data set D_N.

δ-presence model ensures that the probability of inferring the presence of a data subject's record in a transformed data set D^* is within the range of $\delta = (\delta_{min}, \delta_{max})$ from a population data set P. More specifically, let assume that the data set D^* is a transformed form of the private original data set D by generalisation of its QIDs. In this model one assumes that the original data set D is a subset of an externally known data set P in the sense that (some of) the data subjects in D have also data records in the external data set P. Hereby the δ-presence model 'can indirectly prevent record and attribute linkages because if the attacker has at most δ% of confidence that a data subject's record is present in the released table, then the probability of a successful linkage to her record and sensitive attribute is at most δ%' (Nerqiz et al., 2007).

Example of δ-presence: Let P_7 be a population microdata set, containing seven records of individuals.

Population microdata set P_7

Name	Job	Gender	Birthdate
Bob	Engineer	Male	05/12/1982
Fred	Engineer	Male	03/05/1983
Doug	Lawyer	Male	04/09/1984
Alice	Writer	Female	17/03/1987
Cathy	Writer	Female	04/08/1985
Emily	Dancer	Female	08/01/1987
Gladys	Dancer	Female	28/02/1986

Microdata set D_3 with three records, see the following, is a sample microdata set of the population microdata set P_7, when considering the QIDs of 'job', 'gender' and 'birthdate'. Assume that D_3 contains those patients hospitalised last year and the amount they paid.

Private sample microdata set D_3

Name	Job	Gender	Birthdate	Treatment fee
Bob	Engineer	Male	05/12/1982	550 €
Cathy	Writer	Female	04/08/1985	2.300 €
Gladys	Dancer	Female	28/02/1986	1,500 €

Now assume that the following transformed microdata set D^*_3 is made public.

Public sample microdata set D^*_3

Job	Gender	Birthdate	Treatment fee	
Professional	Male	***	550 €	(Corresponding to Bob)
Artist	Female	***	2.300 €	(Corresponding to Cathy)
Artist	Female	***	1,500 €	(Corresponding to Gladys)

Knowing that *set D^*_3 is a generalized sample of* P_7 and considering the QIDs of 'job', 'gender' and 'birthdate', the intruder can conclude that Bob is in D^*_3 with a probability of 1/3. *Cathy or Gladys is* in D^*_3 with a probability of 2/4 = 1/2. Therefore, $\delta = (\delta_{min}, \delta_{max}) = (1/3, 1/2)$.

3.4.3 Data anonymisation tools

There are various software tools, from commercial companies or open source initiatives, each of which offer a set of SDC functionalities for specific application domains. Examples of open source and non-commercial tools are:

- ARX, which is designed for protecting medical microdata sets by Munich University of Technology;
- sdcMicro, which is designed for protecting statistical microdata sets by Statistics Europe;
- μ-ARGUS, which is designed for protecting statistical microdata sets by Statistics Netherlands.

Two other freely accessible data anonymisation tools like Cornell Anonymisation Toolkit (CAT)[37] and UT Dallas Anonymisation Toolbox[38]. Further, a commercial data anonymisation tools available is as Privacy Analytics Eclipse from Privacy Analytics.[39]

ARX, sdcMicro and μ-ARGUS are the three non-commercial software tools covered in our detailed study in the following two chapters. The criteria and motivations for selecting these SDC tools are described in Section 5.1.

3.5 Summary

The SDC technologies studied in this report are concerned with protecting microdata sets. In this chapter, we characterised microdata sets by having a number of records (corresponding to individuals or individual units) and a number of attributes (being of four types: explicit identifiers, quasi identifiers, sensitive attributes, and non-sensitive attributes). For determining quasi identifiers, one must consider the background knowledge that is (going to be) available for intruders.

In order to protect microdata sets against statistical disclosures (i.e., reidentification and attribution at (high-enough) certain levels), one needs to understand the data environment and identify data disclosure scenarios. To this end, we presented four types of data linkage attacks, namely: record linkage, attribute linkage, table linkage, and probabilistic linkage.

The notions of anonymisation and pseudonymisation are two important concepts of SDC-based data protection. We explained these concepts in the technological domain and in relation to similar concepts (like de-identification and anonymisation types). In Table 1 we provide an overview of some of these concepts discussed in this chapter.

Table 1 Summary of the main SDC-based data protection concepts

Trigger of data disclosure	*Intrinsically* (regardless of background knowledge)	*Extrinsically* (in regard to background knowledge)
Sharing a data set (or not)	De-identification (via pseudonymisation, suppression & removal) ≈ formal anonymisation	Guaranteed anonymisation Statistical anonymisation Functional anonymisation
Being in a data set (or not)	Differential privacy, which is done independent from the background knowledge	

As mentioned in Chapter 2, in GDPR the notion of data anonymisation as a process is not mentioned and instead the term anonymous is used to denote the status of a data set without personal data, given 'all the means reasonably likely to be used' (Recital 26 of GDPR, 2016). Further, according to GDPR, pseudonymisation refers to all those data transformation mechanisms that somehow make it possible to reverse the data transformation operations. According to this interpretation, most of technological data anonymisation mechanisms can be regarded as data pseudonymisation mechanisms in legal terms of GDPR.

Protecting microdata sets against disclosures relies on a number of methods as the cornerstone of SDC technologies. We provided a brief overview of some main SDC

[37] See https://sourceforge.net/projects/anony-toolkit/, retrieved on 12 June 2018.

[38] See http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=home, retrieved on 12 June 2018.

[39] See https://privacy-analytics.com/software/privacy-analytics-eclipse/, retrieved on 12 June 2018.

methods, namely: removal, suppression, pseudonymisation, generalisation, permutation, perturbation and anatomisation. In practice, a subset of these methods is used to realise a specific data protection model. Subsequently, we gave a brief overview of main SDC models, namely: k-anonymity, l-diversity, t-closeness, k-map and δ-presence. Finally, SDC tools realise a number of these SDC models to enable data protection against statistical disclosures. We are going to study three of these SDC tools in the following chapters.

In summary, the objective in this chapter was to provide a theoretical foundation for understanding and describing SDC mechanisms and tools. As such, the chapter provides partial answers to research questions Q_2 and Q_3.

4 A functional model of SDC tools

The data anonymisation process in the technological domain is enabled by SDC-based data anonymisation tools (or SDC tools in short). These SDC tools are software systems with a number of functions, which are partly based on the SDC methods and models described in Chapter 3. In this chapter we provide a high-level functional model of the SDC tools. This functional model, which comprises four functional components, is not tool-specific, i.e., it is applicable for all the tools studied in this report. Furthermore, in this chapter we describe a number of new theoretical principles and mechanisms of SDC technologies that were not described in Chapter 3. These new principles and mechanisms (for example, data utility measures, data disclosure risk measures, and the evaluation of data utility and data disclosure risk measures) are related to the components of the high-level functional model. All these aspects are approached from and within the technological domain.

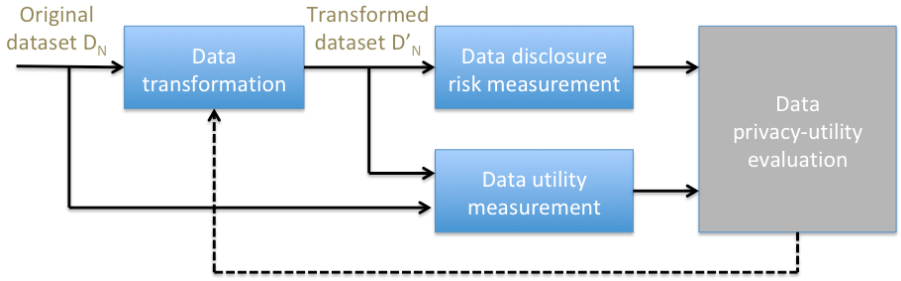
This chapter presents a theoretical foundation for answering research questions Q_2 (investigating the main functionalities of available SDC tools for protecting personal data and preserving data utility) and Q_3 (accounting for background knowledge in protecting personal data). In this chapter, we present a generic functional model of SDC tools in Section 4.1. After explaining the data transformation component shortly in Section 4.2, we provide an overview of common data disclosure risk measures and data utility measures in Sections 4.3 and 4.4, respectively. Subsequently, we elaborate on making privacy-utility trade-offs in Section4.5. Finally, we summarise the main topics discussed in Section 4.6.

4.1 A generic model

The functional model of an SDC tool can be characterised by four components, namely: (1) data transformation, (2) data disclosure risk measurement, (3) data utility measurement, and (4) data privacy-utility[40] evaluation. The relations among the aforementioned components are illustrated in Figure 8. The model is an abstraction of the more detailed models presented in the literature like (Templ, Kowarik & Meindl, 2015). In the following, each component of the functional model is described shortly. In Chapter 5 the functional model shall be specified for each of the studied SDC tools.

[40] Alternatively, data risk-utility. We prefer to use the term data privacy-utility, as it is more commonly used in technological domain literature.

Figure 8 A generic model of data anonymisation process



The data transformation component executes a number of operations on the original microdata set D_N to obtain a transformed microdata set D'_N, where N denotes the number of records in both data sets. Ideally, transformed microdata set D'_N should be safe from data disclosure risks, given the data environment. The operations executed include: attribute mapping (i.e., specifying EIDs, QIDs, SATs and NATs) given the data environment, applying a subset of the data anonymisation methods described in Subsection 3.4.1 to realise a combination of the data protection models described in Subsection 3.4.2. For example, the EIDs are removed, QIDs are generalised and risky records are suppressed/removed, and certain values of SATs are swamped, permutated and/or perturbed in order to enforce a certain data anonymisation model (e.g., k-anonymity and ℓ-diversity) on the original microdata set D_N.

The data disclosure risk measurement component includes those operations needed for quantifying the data disclosure risks (or privacy risks) of the transformed data set D'_N. In the domain of risk management, the term risk is defined as a function, generally the product of the likelihood[41] of occurring a harmful incident (or threat) and the impact of that incident. In our setting, these factors correspond to the likelihood of data disclosures and the impact of the disclosed personal data. In practice, however, the frequentistic likelihood of a data disclosure incident is regarded as the data disclosure risk in SDC literature and we adopt this viewpoint throughout the report.

The data utility measurement component includes those operations needed for quantifying the quality of the transformed microdata set D'_N. Often, the transformed microdata set faces some degree of quality degradation due to the applied transformations to the original microdata set D_N. For some usages, however, the data quality degradation may be unnoticeable. For this component, the main challenge is to measure the data utility, given the data usage context.

The data privacy-utility evaluation component includes those operations needed for making trade-offs between the disclosure risks and utility of the transformed microdata set D'_N based on the corresponding measurements, which are inputs to this component as shown in Figure 8. Quite often, this evaluation leads to making a trade-off between data disclosure risks and data utility. Often, experts make the privacy-utility trade-offs based on their domain knowledge. If the transformed microdata set D'_N is not satisfactory, then the data transformation should be repeated with new and improved parameters. This repetition is illustrated by the feedback link in Figure 8.

[41] Note that there are different notions for this term, like frequentistic, logical, and personalistic probabilities (Vlek, 2013), which we abstract from their differences and their distinguishing characteristics in this report.

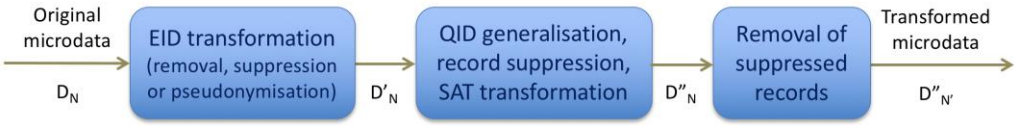
4.2 Data transformation

In the data transformation component in Figure 8 one can apply a combination of various SDC methods. For example, to achieve k-anonymity, one can apply generalisation only or generalisation in combination with record suppression. Applying only generalisation to achieve k-anonymity (i.e., making the sizes of all ECs larger than k) may require significant generalisations of QID values due to outliers in the microdata set D_N. Such large amounts of generalisation often result in an unacceptable data utility, due to information loss (Bayardo & Agrawal, 2005). Applying generalisation in combination with record suppression (i.e., of the outlier records) may alleviate this problem.

For convenience, from this point on we use the following convention to denote the microdata sets produced within the data transformation component, see also Figure 9. This notation is based on the most common operations applied within the data transformation component.

1 From microdata set D_N, the EIDs are removed, suppressed or to obtain microdata set D'_N.
2 Then, to achieve k-anonymity for microdata set D'_N with these combination methods, one may apply a certain generalisation scheme (i.e., generalise the values of each QID to a predetermined higher level) to get a microdata set in which the sizes of some ECs might be smaller or larger than k. Subsequently, one can suppress those ESs (i.e., their records) in the generalised microdata set that have sizes smaller than k (i.e., the outlier records) to obtain a fully k-anonymous microdata set D''_N.
3 Optionally, those records that are entirely suppressed in microdata set D''_N are removed, resulting in a fully k-anonymous microdata set $D''_{N'}$, where $N' \leq N$.

Figure 9 The notation convention used from this point on in the report



4.3 Measures of data disclosure risks

Assessing the required degree of data anonymisation can be related to assessing the degree of data disclosure risks.[42] In this section, we describe two categories of data disclosure risk measures, called elementary measures and advanced measures in this report, in Subsections 4.3.1 and 4.3.2, respectively.

4.3.1 Elementary measures

The parameters of the anonymisation models, described in Subsection 3.4.2, provide a means of measuring the privacy of transformed microdata sets. Specifically, the values of the parameters k, l, c, t and δ – corresponding to data anonymisation models k-anonymity, l-diversity, recursive (c, l)-diversity, t-closeness and δ-presence, respectively – can be used as measures of data disclosure risks, which relate

[42] In this report, we prefer not to use the term 'data anonymisation risk' as the term 'risk' has a better association with the term 'data disclosure' than it has with the term 'data anonymisation'.

inversely to data anonymisation (also known as privacy) measures. One can approximately assume that the higher the value of parameters k, l, c, t and δ is, the higher is the degree of data anonymisation. Domain experts in charge of SDC-based data protection determine the abovementioned values in a given data environment, based on some guidelines that are out our scope.

4.3.2 Advanced measures

Data disclosure risks depend not only on the anonymised microdata set, but also on the contextual conditions and constraints surrounding the data sharing process. As mentioned in Subsection 3.2.2, Mackey and Elliot (2013) use the term *data environment* to refer to these contextual factors, which is formalised in four components: data, agency, governance process, and infrastructure. In order to measure data disclosure risks, one needs to consider these data environment factors. In the following, we describe some aspects of these factors that are relevant within the scope of this report for measuring data disclosure risks.

To measure data disclosure risks, one should make some assumptions about the intruder's method of attack (El Emam, et al., 2013). Consequently, the corresponding measurements are valid within the scope defined by these assumptions. An intruder's attack method can be captured in a so-called *data disclosure scenario*[43] that relates (some of) the components of the data environment to a given data sharing operation and context. A frequently used data disclosure scenario is that in which the intruder tries to reidentify a single record in a transformed microdata set by using some background knowledge found in public registries or gained through searches. An aspect that can also be captured in disclosure scenarios is the intruder's degree of motivation. Intruders' motivations can vary depending on their objectives and expected incentives in (or costs for) reidentifying uncertain records in transformed microdata sets (El Emam, et al., 2013). As a result, intruders can be content with a random match or be willing to take extra efforts to reidentify uncertain records. For example, journalists and marketers may take extra efforts to find a rewarding story or to increase their product sells (El Emam, et al., 2013).

Example of an attack scenario from (El Emam et al., 2013), adopted with adaptation: A scenario of an unmotivated adversary who is content with a random match is described in a number of steps as follows.

- The intruder selects a record with unique values from a transformed microdata set.
- The intruder checks for potential matches in public registries having full identity information by matching the QIDs corresponding to the chosen record.
- If there exists one match in the public registers, then reidentification takes place. In this case, a reidentification occurs with certainty.
- If there exist more than one match in the public registers, the intruder can pick up one of the options randomly (assuming that the intruder does not look for or cannot find additional information). In this case, a reidentification occurs with uncertainty (e.g., with 20% certainty if there are 5 matches in the public registers).

In addition to the degree of the intruders' motivations, the success of the efforts to verify uncertainties may be quantified at various levels. For example, in verifying a match, the journalist may be unable to trace and contact a potential victim or may

[43] Iso called 'attack scenario' (Elliot & Dale, 1999).

receive a misleading reply. 'Therefore, the ability to verify a match is probabilistic' (El Emam, et al., 2013). The term probability here is used in its general sense semanically. It can refer to the result of either a random process or a subjective process for associating a probability value to an event/outcome/match. 'In the literature, although a random match (with probabilistic outcome) is not generally considered a reidentification, a low probability of success is still desired (0.2 is a common requirement …)' (El Emam et al., 2013).

The disclosure scenario can be characterised by *attacker type*, which specifies the agency component of a data environment. In literature (for example see Prasser, Kohlmayer & Kuhn, 2016c; El Emam, et al., 2013) and the references therein, three types of attackers are recognised, namely: prosecutor, journalist, and marketer.

With respect to the *prosecutor type***,** it is assumed that the intruder already knows that the transformed microdata set contains the record of an individual (i.e., the victim). The intruder looks for the specific record of the victim in the transformed data set by using (and due to knowing) the EC of the victim (i.e., due to the intrud-r's background knowledge).

Example of a prosecutor attacker: Assume the police publish a microdata set about the types of crimes occurred in people's homes, together with the location (given by the region) and time of the crime (given per week). Every record in the microdata set corresponds to a crime-house combination. The QIDs are the region and week. The SAT is the type of the crime. Someone, i.e., a nosy neigh-bor, who knows that a crime has occurred a week ago in his neighbor's home can search the released microdata set by using the EC of that particular region and week. In this way, the neighbor can find out the type of crime that occurred if there is one match in the transformed/released microdata set (i.e., to achieve reidentification or attribution with certainty). It is possible that there are multi-ple matches in that EC, then the neighbor can infer the type of crime with uncer-tainty.

For the *journalist type*, it is assumed that the intruder has no prior knowledge about the membership of the victim in the published microdata set. Given a potentially high-risk EC, which also exists in the published microdata set, the intruder carries out some searches in auxiliary information sources (possibly together with executing some extra steps like calling individuals with the same EC in the population and ask-ing them some questions) in order to identify a victim from the EC. Subsequently, the intruder randomly links the identified victim to one of the records from the corresponding EC in the published microdata set.

Example of a journalist attacker: Consider the microdata set published by the police as mentioned in the previous example. A data-journalist aims at reidentify-ing / attributing at least one piece of information in the published / transformed data. The journalist learns from a public microdata set that there are only three homes in a given region and sees the names of the homeowners in the public microdata set. Subsequently, the journalist looks at the published microdata set and in the corresponding EC (i.e., of that region and time interval) sees two domestic violence crimes reported. Consequently, the intruder can infer with some likelihood the occurrence of domestic violence in each of the households in that region.

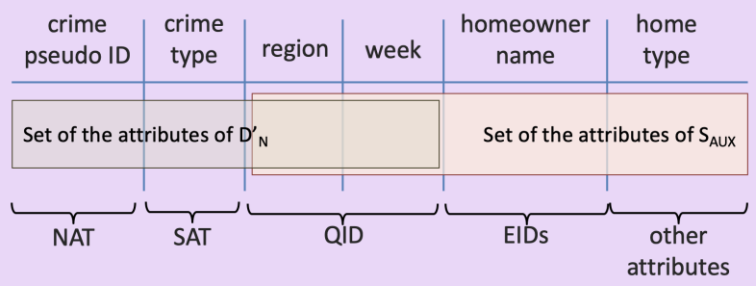
For the *marketer type*, like in the journalist type, it is assumed that the intruder has no prior knowledge about the membership of the victims in the published/ transformed microdata set. The intruder, however, intends to reidentify/attribute a larger number of victims than the journalist for, for example, marketing purposes.

Example of a marketer attacker: Consider the microdata set published by the police as mentioned in the previous example. A marketer of a burglary protection product considers all ECs that are low populated (known from a public register) and have five or more crimes of the burglary type (known from the published microdata set). Subsequently, the marketer learns from the public register the names (and addresses) of all homeowners in the corresponding regions with such a high rate of burglary and puts their contact information in the CRM (Customer Relationship Management) system of the company as potential buyers of the product without asking their consent.

One important step in determining the data disclosure risks is the degree of *uniqueness* of individuals in the published/transformed microdata set and in the population. To illustrate these degrees of uniqueness, let's assume that transformed microdata set D''_N in Figure 9, which comprises N records of M attributes in set $\{A_1, A_2, \ldots, A_M\}$, is released. Every record in the released microdata set D''_N can potentially be identified based on the QIDs that link the sensitive information in the transformed microdata set D''_N to the identification information in an auxiliary microdata set[44] S_{AUX} with attributes $\{S_1, S_2, \ldots, S_{M'}\}$ being available to intruders (i.e., being in the data environment). As illustrated also in the following example and Figure 10, QIDs $= \{A_1, \ldots, A_M\} \cap \{S_1, \ldots, S_{M'}\}$.

Example of data linkage based on QIDs: The linkage between the records of the released data set with an auxiliary data set based on QIDs is illustrated in Figure 10 for the example of the police crime data set.

Figure 10 An illustration of data linkage for the released police crime microdata set



We start with determining the degree of uniqueness of individuals in the published microdata set D''_N. Note that determining the uniqueness of individuals corresponds to determining the uniqueness of their records in microdata set D''_N because, as we assumed, there is a one-to-one relationship between a record and an individual.[45] The degree of uniqueness is defined with respect to those attributes that are known also in auxiliary information sources and serve as background knowledge for microdata set D''_N. In other words, the degree of uniqueness is determined according to

[44] Or, so-called, the 'identification database' in El Emam (2008).

[45] Later we shall relax this condition by assuming multiple data records per individual (e.g., in case of the records of a household).

the QIDs of the records in the published microdata set D''_N. For an EC S in D''_N (i.e., a value pattern of QIDs in microdata set D''_N), let |S| be the number of the data records of D''_N in that EC S. The value of |S| determines the uniqueness of the records/individuals in that EC S. If |S| = 1, then the corresponding record/individual is unique in the published microdata set D''_N. A larger value of |S| makes the corresponding records less unique.

Example of uniqueness in a microdata set: Consider the EC S in the released police crime microdata set, where the pattern value of the QIDs is (region = "Rotterdam-northwest", week= "05-2018").

If |S| = 1, i.e., one crime occurred in week 5 of 2018 in Rotterdam northwest, then the corresponding record/individual is unique in the published microdata set.

Now we define the concepts of *sample uniqueness* and *population uniqueness*. With respect to the set of QIDs, let us assume that microdata set D_N is a sample of a larger population microdata set P_L (i.e., $N \leq L$). Alternatively said, all data records in sample microdata set D_N are also in population microdata set P_L, where these microdata sets D_N and P_L have QIDs in common.

Example of a population microdata set: Assume that the released police crime microdata set is concerned with crimes that occurred in Rotterdam or the whole of the Netherlands. Then, the population data sets are the register data sets of all homeowners in Rotterdam or the Netherlands, respectively, in which the location of houses and periods of ownership are registered and can be used as QIDs.

For two other examples of population and sample microdata sets, see Section 3.4.2.

Let us further assume that the QIDs in microdata sets D_N and P_L are generalised in the same way, resulting in microdata sets D''_N and P''_L with the same ECs (i.e., the same patterns of values for the generalised QIDs). Uniqueness of an individual/record that in both microdata sets D_N and P_L can be defined as follows. Assume that a data record/individual belongs to an EC, which has |S| and |P| records in microdata sets D_N and P_L, respectively. The sample uniqueness and population uniqueness of the record (i.e., the corresponding individual) is defined by values |S| and |P|, respectively. We note that

- population uniqueness results in sample uniqueness (if |P| = 1, then |S| =1); and
- sample uniqueness does not necessarily result in population uniqueness (if |S| = 1, then $|P| \geq 1$).

One should also note that while a data controller can easily validate sample uniqueness by investigating the released microdata set, (s)he cannot easily validate population uniqueness because population microdata sets are generally accessible to intruders and not to data controllers. Nevertheless, in some cases, population uniqueness is more relevant than sample uniqueness to determine data disclosure risks. If the intruder knows that an individual's record is in the sample microdata set, as in the case of prosecutor attacker (e.g., the background knowledge that a nosy neighbour has), then it is important to investigate sample uniqueness. If the intruder is uncertain whether an individual's record is in the sample microdata set, as in the cases of journalist and marketer attackers, then it is important to investigate population uniqueness. The rationale behind the latter is that it is not risky if a

record, which appears alone in an EC in the sample microdata set, shares the EC with multiple records in the population microdata set, as illustrated in the following example.

Example of sample uniqueness: Assume that attributes 'age' and 'marital status' are QIDs. Further, assume that we know Alice being a widow of 15 years old (the example is adopted with modifications from (Elliot et al., 2016)). Based on the registration database of Rotterdam municipality (i.e., the sample micro-data set), there is one record in EC of ('age' = 15, 'marital status' = widow), while there are ten records in that EC in the national registration database of the Netherlands (i.e., the population microdata set). Further assume that we have access to the sample microdata set (i.e., the registration database of Rotterdam municipality).

Case of prosecutor attacker: if we know that Alice lives in Rotterdam, then we can learn more about Alice from the sample microdata set.

Case of journalist attacker: if we don't know that Alice lives in Rotterdam, then there is a 10% chance that Alice appears in the sample microdata set. Thus, we can only attribute the information in the sample microdata set to Alice with 10% certainty.

To illustrate this dependency of uniqueness to attacker types, see the following box for a simplified calculation of record reidentification risks. Note that we do not elaborate here on complex attack scenarios where, for example, the intruder takes extra measures by going to a region and doing field research by asking verification questions from local residents. (For an introduction to such motivated attackers see El Emam et al., 2013; Dankar et al., 2012.)

Example of a probability model from Prasser, Kohlmayer and Kuhn (2016c): Assume that |S| and |P| denote the sizes (i.e., the number of the records) of the EC, to which an individual belongs in the sample and population microdata sets, respectively.

For a prosecutor attacker, the probability of correctly linking the individual with a record from the sample microdata set is

$$\Pr(\text{correct linkage} \mid \text{being sure about the membership in S}) = 1 / |S|.$$

In this case sample uniqueness, captured by |S| above, is important.

For a journalist attacker with moderate motivation (i.e., the one who stops after looking at the population microdata set without posing further questions or doing further field investigation), the probability of correctly linking the individual with a record from the sample microdata set is

$$\Pr(\text{correct linkage} \mid \text{being sure about the membership in S}) \times \Pr(\text{being sure about the membership in S}) = (1 / |S|) \times (|S| / |P|) = 1 / |P|.$$

In this case the population uniqueness, captured by |P| above, is important.

The marketer attacker wants to reidentify a large number of the records in the sample microdata set. Thus, the marketer's probability of successfully reidentifying victims in the sample microdata set can be expressed as an average of the reidentification probabilities of all records, each of which are calculated based on the journalist attacker type above. As such, also in this case population uniqueness is important.

Generally,[46] protecting microdata sets against prosecutor attackers makes them protected also against journalist attacks and, similarly, protecting microdata sets against journalist attackers makes them also protected against marketer attackers (Prasser, Kohlmayer & Kuhn, 2016; El Emam, 2010). This can symbolically be represented as:

> Protection against prosecutor 〉 Protection against journalist 〉 Protection against marketer.

If one should protect a microdata set in an environment with two or more of these attacker types, then the microdata set should be protected against the most severe attacker type in that environment. For example, when both prosecutor and journalist attackers are important for an open data initiative, the protection should be tuned to the prosecutor type (assuming that $|P| \geq |S|$).

As mentioned before, using sample uniqueness is too conservative for assessing data disclose risks in cases that are not concerned with the prosecutor attacker type. Anonymizing against sample uniqueness requires more severe generalisation of QIDs and suppression of records in the released microdata set than those anonymisation measures needed against population uniqueness – remember that one can afford having unique records in the sample microdata set if there are many of such records in the population microdata set. These extra adaptations inflict more adverse impacts on the utility of the released microdata set. For example, Prasser et al.(2016b) reports a 10-25% improvement of data utility when data anonymisation is applied to the population microdata sets than when it is applied to the sample microdata sets. Therefore, in such cases some authors have proposed to define data disclosure risks based on population uniqueness.

To define the data disclosure risks associated with population uniqueness, the data controller should either

a have access to a copy of the population data set; or
b estimate the uniqueness in the population data set by using only the disclosed data set (Dankar et al., 2012).

Option (a) is usually seen to be resource intensive and costly, as it requires updating population microdata sets regularly (Dankar et al., 2012). We argue, nevertheless, that it seems more feasible nowadays to acquire such a population microdata set as the number of open (and big) data initiatives increases. For option (a) where the data controller has access to the population microdata set, one can use the $\hat{k}$-map method (El Emam & Dankar, 2008). As explained in Subsection 3.4.1, in the $\hat{k}$-map method one applies the $\hat{k}$-anonymity method to the population microdata set P_L directly and subsequently maps the structure of $\hat{k}$-anonymity from the population microdata set to the sample microdata set.

[46] As long as there is a one-to-one relationship between individuals and the data records in the sample and population microdata sets.

For option (b), where there are no data about the population microdata set, the uniqueness in the population microdata set could be estimated with statistical models. To this end *super-population models* are used to estimate the characteristics of the overall population with appropriate probability distributions that are parameterised with the characteristics of the sample microdata set. In Dankar et al. (2012), an overview of these estimation models is provided from literature and the accuracy of four promising estimators are evaluated using a Monte Carlo simulation on six clinical health microdata sets. These four promising candidate estimators are: the Zayatz estimator (Zayatz, 1991), slide negative binomial estimator (Chen & Keller-McNulty, 1998), Pitman's estimator (Pitman, 1996; Hoshino, 2001), and μ-ARGUS estimator (Benedetti & Franconi, 1998). Dankar et al. (2012) find out that the Pitman and Zayata models are the most accurate for low and high sampling ratios, respectively. As a rule of thumb, the Pitman model is advised for sampling ratios below 10% (ARX manual, 2018). This decision rule is implemented in, for example, ARX to estimate the marketer reidentification risks.

For case (b), in addition to using super-population model to calculate disclosure risks associated with population uniqueness, one can calculate individual disclosure risks per record from the sample microdata set, and then simply add them up, weighted with the corresponding sampling weights (Templ et al., 2017, p. 7).

4.4 Measures of data utility

Assessing the degree of data utility is an indicator for deciding on whether the outcome of the applied SDC method and model is useful, given the purpose of data usage. In this section, we describe two categories of data utility measures, so-called general-purpose and special-purpose measures (Fung et al., 2010), in Subsections 4.4.1 and 4.4.2, respectively. Per category we describe a few representative data utility measures, without intending to be exhaustive.

4.4.1 General-purpose measures

The aim of SDC-based data anonymisation is to deliver a transformed microdata set that is good enough according to some quantifiable costs (or, alternatively, according to some data utility measures). To this end, the data transformation component in Figure 8 which applies a combination of various SDC methods, introduces some information/utility loss in the output microdata set $D''_{N'}$ (adopting the notation introduced in Figure 9).

General-purpose metrics do not consider the purpose for which the data are going to be used. As such, they are useful for, for example, open data initiatives where the data publisher does not know how data recipients will use and analyse the published data. Kohlmayer et al. (2015) define two types of data utility measures, which are applicable as general-purpose metrics, namely:

- *Class-based* measures, which are based on the sizes of the ECs. Example measures are discernibility and AECS.
- *Attribute-based* measures, which are computed for each attribute individually and are subsequently compiled into a global measure. Example measures are Precision and those based on information theory entropy rate.

The discernibility measure or Discernibility Metric (DM; see Bayardo & Agrawal, 2005) aims at capturing the level of maintaining the discernibility between the

records in original microdata set D_N and those in transformed microdata set $D''_{N'}$. The metric assigns a penalty to every record in microdata set D_N based on the amount of its difference with the corresponding record in the transformed microdata set $D''_{N'}$, i.e., how much indistinguishable they are according to Bayardo and Agrawal (2005). The value of the penalty applied to a record in D_N is

- k′ for generalised records. Value k′ is the number of the records of the EC to which the generalised record belongs in $D''_{N'}$ (note that $k' \geq k$). This penalty is because a record in an EC cannot be distinguished from the other k′-1 records in the same EC (i.e., considering the values of the QIDs).
- $|D_N|$ for suppressed records. Value $|D_N|$ is the number of the records of the original microdata set D_N because a suppressed record cannot be distinguished from any other records in the original microdata set D_N.

The discernibility measure can mathematically be stated as (Bayardo & Agrawal, 2005):

$$C_{DM} = \Sigma_{\{\forall \text{ i where } |EC(i)| \geq k\}} |EC(i)|^2 + \Sigma_{\{\forall \text{ i where } |EC(i)| < k\}} |EC(i)|.|D_N|.$$

In using this loss function for data anonymisation, one can make the penalty of record suppression infinite in order to push the data anonymisation towards a generalisation only approach. Further, note that record suppression is a rather drastic operation, as can be seen in the amount of its penalty (i.e., $|D_N|$) considered in the discernibility measure. Often a hard limit is imposed on the number of suppressions allowed, like maximum 10% of the records of original microdata set D_N can be suppressed (see Samarati, 2001; Bayardo & Agrawal, 2005).

The Average Equivalence Class Size (AECS) measure is proposed by Lefevre et al. (2006) as an alternative to the discernibility measure. It is given by:

$$AECS = |D''_{N'}| / (k \times \text{number of ECs of } D''_{N'}).$$

The higher the value of AECS, the higher is the amount of information loss. If the size of all ECs of microdata set $D''_{N'}$ is k, then the AECS measure value is one, i.e., its minimum and best value. The optimisation objective of data anonymisation here is to reduce the AECS value to 1 (i.e., to find a partitioning that approaches the best case). Apparently the AECS does not consider the impact of record suppressions as severely as the discernibility measure does.

The *precision* measure (Sweeney, 2002b) aims at measuring the amount of distortion in a generalised microdata set $D''_{N'}$. Every cell in the anonymised microdata set may be generalised to a level h out of maximum level H. For example, a cell denoting the birthdate, which is represented by data format dd/mm/yyyy, can be generalised to levels mm/yyyy, while yyyy is the maximum generalisation level. In this case, the h and H values of the cell are 1 and 2, respectively. The ratio of h/H is defined as the cell's *distortion* (or information loss), which is ½ for the example mentioned. The precision measure is defined as:

$$1 - (\text{the sum of all cell distortions}) / \text{the total number of cells}.$$

To define it mathematically, let $d''_{n, m}$ denote the n^{th} and m^{th} cell of microdata set $D''_{N'}$. The corresponding generalisation height and maximum generalisation height are denoted by $h_{n, m}$ and $H_{n, m}$ respectively. Then the precision measure is (see also definition 5 in Sweeney, 2002b):

$$1 - ((\ \Sigma_{m=1,\ \ldots,\ M;\ n=1,\ \ldots,\ N'}\ (h_{n,\ m} / H_{n,\ m})) / N' \times M).$$

Here M is the number of attributes in $D''_{N'}$. The precision measure adds up cell distortions and therefore does not consider the relative importance among different types of cells(for an example see Bui et al., 2015, p. 169).

Gionis and Tassa (2009) define three measures of information loss, based on information theory entropy rate. They call these measures as the *entropy* measure, the *monotone entropy* measure, and the *non-uniform entropy* measure. These measures are calculated based on the distribution of values in the original microdata set D_N, given the distribution of values in the transformed microdata set $D''_{N'}$. For example, let attribute A be a QID, which takes values from set $\{a_1, a_2, \ldots, a_I\}$ in the original microdata set D_N. In the transformed microdata set the values of attribute A are generalised to values $\{a_{1,2}, a_{3,4}, \ldots, a_{I-1,I}\}$, i.e., values a_1 and a_2 in D_N are generalised to value $a_{1,2}$ in microdata set $D''_{N'}$ and so on. Let A" denote the attribute in $D''_{N'}$ that corresponds to attribute A in D_N, noting that the values of A" are from $\{a_{1,2}, a_{3,4}, \ldots, a_{I-1,I}\}$. Then, for example, given that $A = a_1$ and $A'' = a_1$-a_2, the information loss due to generalisation for this outcome is proportional to $- \log_2 ((\#\text{ of } a_1) / (\#\text{ of } a_1 + \#\text{ of } a_2))$.

Let A_m, where m:1, …, M, denote a QID in microdata set D_N and A''_m denote the QID in microdata set $D''_{N'}$ that corresponds to A_m in D_N. We use $\{a_m\}$ and $\{a''_m\}$ to denote the sets of values of QID A_m and the corresponding QID A''_m, respectively. Further, we denote information theory entropy rate function and conditional entropy rate function by H(.) and H(.|.) in the following, respectively.

The *entropy* measure for microdata sets D_N and $D''_{N'}$ is defined in Relation (4) in Gionis and Tassa (2009) as (with slight adaptation, using our notation defined above):

$$\Sigma_{m=1,\ \ldots,\ M}\ H(A_m \mid A''_m).$$

The *monotone entropy* measure is defined in Relation (5) in Gionis and Tassa (2009) as (with slight adaptation, using our notation defined above):

$$\Sigma_{m=1,\ \ldots,\ M}\ \Sigma_{a''\_m\ \in\ \{a''\_m\}}\ \Pr\ (A''_m = a''_m)\ .\ H(A_m \mid A''_m = a''_m).$$

The *non-uniform entropy* measure is defined in Relation (6) in Gionis and Tassa (2009) as (with slight adaptation, using our notation defined above):

$$-\ \Sigma_{m=1,\ \ldots,\ M}\ \Sigma_{a\_m\ \in\ \{a\_m\}\ \text{and}\ a''\_m\ \in\ \{a''\_m\}}\ \log_2\ \Pr\ (A_m = a_m \mid A''_m = a''_m).$$

Unlike 'entropy measure', both 'monotone entropy measure' and 'non-uniform entropy measure' are monotonic with respect to generalisation. Monotonicity of these measures means that they increase monotonically with increasing degrees of generalisation. In other words, if the value a_1 is generalised to value $a_{1,2}$ at level 1 and then to value $a_{1,4}$ in level 2, then the corresponding values of the measure increase when moving from level 1 to level 2.

In Kohlmayer, Prasser and Kuhn (2015) the authors show that the non-uniform entropy measure is not good for evaluating the quality of locally recoded data records, i.e., when the generalisation for an attribute does not occur similarly for all records (like when the age attribute is rounded to intervals of five years for some records

and to intervals of ten years for the other ones). The authors propose a generic model of the non-uniform entropy measure to address its shortcoming for locally recoded records.

Two attribute-based measures are Hamming distance and generalisation height measures, as described in the following.

The *expected Hamming distance* measure is defined in Wang et al. (2014) as the expected value of the Hamming distance between input microdata set D_N and output microdata set $D''_{N'}$ of a data anonymisation process. The measure can be denoted by avg(d(D_N, $D''_{N'}$)), where d(D_N, $D''_{N'}$) represents the number of rows microdata sets D_N and $D''_{N'}$ differ on. The expected Hamming distance averages the individual Hamming distances d(D_N, $D''_{N'}$) based on their joint probability Pr(d(D_N, $D''_{N'}$)). The Hamming distance characterises the neighbouring relation between D_N and $D''_{N'}$ (therefore, it is useful when considering ε differential privacy). We suspect, this Hamming distance is not useful as normally all rows of a microdata set D_N are modified in a typical data anonymisation process.

The *generalisation height* measure is defined in LeFevre et al. (2005) and Samarati (2001) as the height of a transformed microdata set in the generalisation lattice, i.e., the number of generalisation steps performed for the set of the attributes generalised. For example, assume the generalisation lattice consists o 'gender' and 'postal code'. If each of these attributes are generalized one level higher in their taxonomy trees, then the height of the corresponding multi-attribute generalisation method is 2.

As seen above, there are many measures to quantify the utility of transformed data sets. One can choose a measure that is suitable for the application at hand or integrate a subset of those measures. For example, the software tool ARX (from version 2.3 on) supports user-defined *aggregate functions for many measures*. Here one can choose the maximum, sum, arithmetic mean (the recommended one), geometric mean and rank arithmetic mean of multiple measures (ARX manual, 2018).

Sometimes it is desirable to reduce information loss for some QIDs more than that for the other ones. Therefore, one can assign *weights to QIDs* so that those attributes with higher weights become subject to lower information losses (ARX manual, 2018).

4.4.2 Special-purpose measures

There are also *special-purpose metrics* that can be used for those cases in which the purpose and usage of the data are known at the time of data publication (Fung et al., 2010). For example, when the transformed microdata set $D''_{N'}$ is going to be used for the classification of a target attribute T_{ATT} in $D''_{N'}$, then the data transformation should preserve the values of those attributes that are essential for discriminating the class labels in target attribute T_{ATT}. Some may wonder why not to publish the result of data mining if the purpose is known beforehand. Fung et al. (2010) answer 'that publishing a data mining result is a commitment at the algorithmic level, which is neither practical for the non-expert data publisher nor desirable for the data recipient. In practice, there are many ways to mine the data even for a given purpose, and typically it is unknown which one is the best until … different ways are tried.' In the following, we describe a few special-purpose measures of data utility.

The *classification* measure (Iyengar, 2002) is applied to those (transformed) microdata sets with records that are assigned to a categorical class label (i.e., that have also a target attribute). After generalisation and record suppression, the measure assigns penalties to those records in the transformed microdata set $D''_{N'}$ that do not contribute to (i.e., do not provide information on) discriminating the class labels of the target attribute. Specifically, let record d''_n be the n^{th} record in the transformed microdata set $D''_{N'}$. The value of the penalty for record d''_n, denoted by $Pen(d''_n)$ is

- 1 if record d''_n is suppressed in $D''_{N'}$;
- 1 if record d''_n is generalised in $D''_{N'}$ and its class label is not the same as the majority class for the records in $EC(d''_n)$, where $EC(d''_n)$ denotes the EC of record d''_n in $D''_{N'}$;
- 0 otherwise.

The Classification Measure (CM) is the normalised sum of all penalties, i.e.,

$$CM = (\Sigma_{n=1, \ldots, N} Pen(d''_n)) / N.$$

The CM penalises those records that are either suppressed or have different class labels in their ECs. In Prasser et al. (2017) the abovementioned model is extended by introducing more fine-grained penalties to introduce more discrimination power. Specifically, the value of the penalty for record d''_n is

- P_r if record d''_n is suppressed in $D''_{N'}$;
- P_m if record d''_n is generalised in $D''_{N'}$ and its class label is not the same as the majority class for the records in $EC(d''_n)$, where $EC(d''_n)$ denotes the EC of record d''_n in data set $D''_{N'}$;
- P_h if record d''_n is generalised in data set $D''_{N'}$ and there is no unique most frequent value of the class label in $E(d''_n)$;
- 0 otherwise.

Classification performance measures (Prasser et al., 2017) aim at assessing the performance of classifiers built from transformed microdata sets $D''_{N'}$. Such measures can also be used for evaluating how well the CM is applied to a microdata set. For these measures a number of classifiers (e.g., C4.5 decision trees and logistic regression models) are trained with the transformed microdata set. Subsequently, the classifiers are evaluated with some appropriate records from the original microdata set. The results obtained with these classifiers can be compared to the results obtained for those classifiers that are trained on the original microdata set. For these evaluations Prasser et al. (2017) mention the following four measures to assess prediction accuracies:

- baseline accuracy: it is determined by using a method that always returns the most frequent class of the training microdata set, trained with and evaluated on the original microdata set;
- original accuracy: it is determined according to the performance of a non-trivial classifier, trained with and evaluated on the original microdata set;
- accuracy: it is determined for a non-trivial classifier, built from the transformed microdata set (see above);
- relative prediction accuracy: it is determined for the classifier trained with transformed microdata set by normalizing its accuracy, using the baseline accuracy and original accuracy.

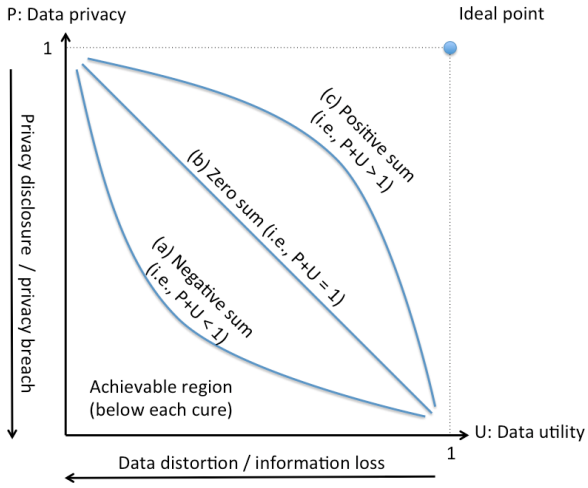
4.5 Data privacy-utility evaluation

The functionality of the data privacy-utility evaluation component in Figure 8 is to determine whether a transformed microdata set is good enough (or, ideally, is the best). The data anonymisation process aims at producing a transformed microdata set that satisfies the related personal data protection requirements and retains as much data utility as possible (Fung et al., 2010).

Sometimes the purpose for which the data are going to be used is known beforehand. For example, the data will be used for creating a classification model to classify a target attribute, based on the other attributes in the microdata set. In this example, it is important to maintain those attributes with class discriminatory properties as much intact as possible. In these purpose specific cases, one should go for maximum level of data anonymisation, given the desired purpose (i.e., so that the purpose can be attained). Of course, the maximum level of data anonymisation must be acceptable from the viewpoint of data disclosure risks. Sometimes, like in case of open data, the data should be usable for as many purposes as possible. In these cases, one should define a minimum acceptable level of data disclosure risks and deliver the data with as much utility as possible. Eventually, the data privacy and data utility should be optimised, i.e., trade-offs should be made, given the data environment (see also Subsection 3.2.2) in which the data are going to be shared/
opened.

For well-structured data anonymisation schemes, i.e., those for which data disclosure risk and data utility measures can be formalised and measured, one can deploy privacy-utility curves to represent the trade-offs between data privacy and data utility, as schematically shown in Figure 11. For example, Wang et al. (2014) define a privacy-utility curve in terms of D (distortion) and ε (privacy leak). For other examples see Sankar, Rajagopalan and Poor (2013), Du Pin Calmon and Fawaz (2012), Salamatian et al (2013); Makhdoumi et al., 2014,). In Figure 11 the data utility measure (which relates inversely to the information-loss measure) and data privacy measure (which relates inversely to data disclosure risk measure) are quantified as positive values in the range of 0 and 1, on a two-dimensional Privacy-Utility (PU) plane. The ideal point on the plane is the perfect data privacy and data utility point with coordinates (1, 1), and the undesired point is the no privacy and no utility point with coordinates (0, 0). The performance of every data anonymisation solution can be modelled by a curve on PU plane when the parameters of the solution are varied. (For example, in a solution based on k-anonymity and maximum 10% record suppression limit, the curve can be obtained when the value of parameters k varies. Note that in this case we do not get a continuous curve as shown in Figure 11.)

Figure 11 Illustrative curves of privacy-utility trade-offs



As schematically shown in Figure 11, the data anonymisation schemes can have three characteristics: Negative sum, zero sum and positive sum. A scheme of zero-sum type yields the same amount of gain in one aspect as it inflicts loss on the other aspect (i.e., P = 1 - U). A positive sum scheme is preferred as it delivers more gain in one aspect than it causes loss in the other aspect (i.e., P > 1 - U).

Note that the problem of finding a solution that optimises the privacy and utility criteria is complex (Mivule & Turner; 2013). For example, Rastogi et al. (2007) show that optimizing privacy-utility in the presence of excessive background knowledge (i.e., when considering the extrinsic factors) would be impossible to achieve. Further, it is shown that achieving k-anonymity with minimal loss of data, being measured by any measures, is NP-hard (see Gionis & Tassa, 2009; Meyerson & Williams, 2004). In practice, nevertheless, there are heuristics proposed to circumvent these complexities and provide near optimal solutions. For example, Prasser et al. (2016a) provide such a heuristic for automatically finding those data transformations that adequately balance data privacy and utility. This heuristic is realised in tool ARX, which partially provides data privacy-utility evaluation functionality. Other tools, like µ-ARGUS and sdcMicro, rely fully on human intelligence to realise data privacy-utility evaluation.

4.6 Summary

SDC tools play a key role in the data anonymisation process for protecting microdata sets against statistical disclosure risks. In this chapter we described a generic functional model of SDC tools. The model comprises four components, namely: data transformation, data disclosure risk measurement, data utility measurement, and trade-off evaluation. The data transformation component uses the SDC methods and models to transform the original microdata set to an anonymised microdata set (ideally).

The data disclosure risk measurement component aims at quantifying the data disclosure (or privacy) risks of the transformed microdata set. We described two categories of data disclosure risk measures, called elementary measures (like the values of k and l in k-anonymity and l-diversity) and advanced measures (which,

in turn, rely on defining data disclosure scenarios such as prosecutor, journalist, and marketer attackers).

The data utility measurement component aims at quantifying the quality of the transformed microdata set. To this end, we presented general-purpose measures (discernibility measure and average EC size measure) and special-purpose measures (classification measure and classification performance measures). Unlike special-purpose measures, general-purpose metrics do not consider the purpose for which the data are going to be used.

The data privacy-utility evaluation component aims at making the trade-offs between the disclosure risks and utility of the transformed microdata set based on the corresponding measurements. The challenge of this evaluation is to achieve a positive sum trade-off where both aspects of data disclosure risk and data utility become acceptable, given the personal, legal, ethical, etc. constraints.

The high-level functional model will be used as a benchmark in the following chapter to describe the functionalities of the SDC tools studied. This benchmark shall serve as a framework to get insight into the functionalities of these tools. Consequently, the chapter presented a theoretical foundation for answering research questions Q_2 and Q_3.

5 On functionalities of SDC tools

In this chapter, we provide an overview of the functionalities offered by three SDC tools that are non-commercial and open source. Further, we shall investigate some non-functional features of these tools, mainly from the perspective of those experts in data science and analytics who want to learn about SDC tools and methods. As such, the chapter contributes to answering research question Q_2 (investigating the main functionalities of available SDC tools for protecting personal data and preserving data utility) and partly answering research question Q_4 (giving insights into promising SDC functionalities or methods).

In this chapter, specifically, we start with describing the criteria and motivations for selecting the SDC tools studied in Section 5.1. Subsequently, we present an overview of the main functionalities of each SDC tool in Sections 5.2, 5.3 and 5.4. We provide a framework for specifying the non-functional aspects of these tools in Section 5.5, and propose some experiments for testing the scalability aspects of these tools in Section 5.6. Finally, we summarise the main topics discussed in Section 5.7.

5.1 Selection of the tools

In this study we have considered three SDC software tools based on Internet searches carried out by the authors and a number of RUAS[47] students, and literature study. The criteria for choosing these tools are:

- providing protection against statistical disclosures;
- being available for the project without extra costs to easily explore their functionalities;
- having enough documentations to explain their functionalities and how to use them; and
- being deployable within our ICT infrastructure without (expected) privacy harms/threats to the data sets studied.

Considering these criteria, we excluded commercial tools and selected the following three open source publicly available tools developed by academia and public organisations. These tools are:

- μ-ARGUS tool (from CBS/Statistics Netherlands);
- sdcMicro (from Eurostat, R-based);
- ARX tool (from Technical University of Munich).

There are other freely accessible data anonymisation tools like Cornell Anonymisation Toolkit (CAT)[48] and UT Dallas Anonymisation Toolbox,[49] which we did not consider in this study due to their limited documentation relatively, as well as our limited resources. Further, note that there are also commercial data anonymisation tools, such as Privacy Analytics Eclipse from Privacy Analytics,[50] which we did not

[47] Two of the authors gave a course on data protection technologies at Rotterdam University of Applied Sciences (RAUS) as part of minor 'Data Science'. This course has been given twice (in 2016 and 2017), each time for about 60 students.

[48] See https://sourceforge.net/projects/anony-toolkit/, retrieved on 12 June 2018.

[49] See http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=home, retrieved on 12 June 2018.

[50] See https://privacy-analytics.com/software/privacy-analytics-eclipse/, retrieved on 12 June 2018.

consider in this study due to the costs associated and due to limited access to documentation and functional specifications of commercial products.

The benefits of investigating these specific software tools are manifold. Primarily, the investigation gives an insight into those data anonymisation methods and models that are important in being developed and deployed in existing tools within the SDC community. The open (source) nature of these tools, on the other hand, offers us an opportunity to easily experiment with these tools, obtain hands-on experience with the tools, and also learn from similar experiences of the research community and academia who basically incline towards investigating those software tools that are open to learn, use, and extend. Knowing the range of feasible functionalities across different software tools, moreover, can be instrumental for developing a vision for joining forces of these tools and developing and/or adopting relevant SDC tools in the future.

In the following sections, we analyse the functional capabilities of the three aforementioned tools.

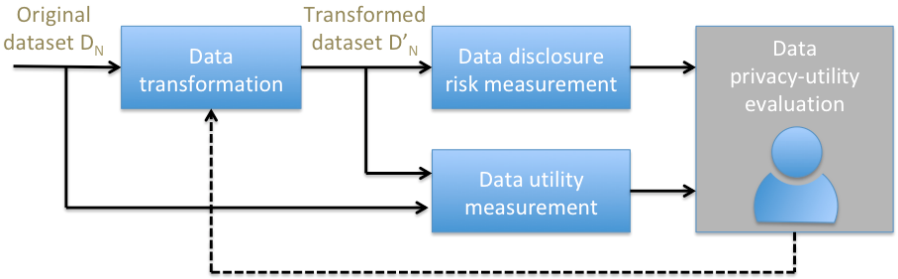
5.2 Main functionalities of µ-ARGUS

For producing safe microdata sets, Statistics Netherlands, in cooperation with its European partners, has created the µ-ARGUS software package.[51] The package is developed under Windows 7, JAVA 7 and SPSS 22. The development of µ-ARGUS has been done with the SDC project (supported by the EU's 4th framework), the CASC (Computational Aspects of Statistical Confidentiality) project (supported by the EU's 5th framework), the CENEX-SDC (2006) project, and the ESSNet-SDC project (both of the latter are supported financially by Eurostat). µ-ARGUS is made Open Source and currently its libraries are being integrated with those of the sdcMircro tool (see Section 5.4) with the support of Eurostat.

Hundepool et al. (2014) provide the general background, principles and usage manual of the µ-ARGUS software package. For an in-depth theoretical background, one can look at µ-ARGUS handbook (Hundepool et al., 2012) and articles (Hundepool & Wolf, 2011; Hundepool et al., 2014). In Figure 12 the functional components of µ-ARGUS are shown, which are similar to those shown in Figure 8 except the data privacy-utility evaluation in µ-ARGUS is done by the end-user (i.e., a domain expert).

[51] See http://neon.vb.cbs.nl/casc/mu.htm, retrieved on 12 June 2018.

Figure 12 The functional components of μ-ARGUS (and sdcMicro)



In the following, we first describe a number of the functionalities of μ-ARGUS that are specific to this tool (based on its last release: Version 5.1.1, released on 30 April 2015) and literature. Subsequently, we summarise the main functionalities of μ-ARGUS in a table.

5.2.1 Data transformation

In order to perform data transformations in μ-ARGUS, one must start with defining the metadata of the original microdata set D_N. In this process one should define the following parameters for every attribute:

- the type, in being numerical or categorical, or being related to a household or not;
- the identification level, in being level 0, which means no identifier (i.e., not being an EID), or a positive natural number from {1, 2, …}, which indicates the ordinal identification significance. Value 1 is the highest identification level;
- the structure, in being hierarchical (like postal codes in the Netherlands);
- the categorical code, in having a file to map the attribute's numerical values to categorical values.

Subsequently, μ-ARGUS offers the possibility of defining QIDs[52] in a flexible way. Unlike the other two tools where one defines n attributes as QID, in μ-ARGUS it is possible to define any m combinations of those n attributes as a *QID pattern* (note that we use the term *QID value pattern* to refer to the EC values that every QID pattern can assume). Alternatively, μ-ARGUS offers an automatic way of combining attributes that are defined with an identification level of 1 or higher (i.e., 1, 2, …). The QID patterns in this automatic case are formed by choosing all combinations in which one attribute is chosen from level 1, one attribute is chosen from level 2, and so on. In our opinion, this flexibility in automatically defining QID patterns is an advantage of μ-ARGUS (despite the fact that it does not allow for defining patterns with two or more attributes from the same level automatically). It this way, we suspect, μ-ARGUS may adapt to the background knowledge with a high granularity, which results in improved data utility.

[52] In μ-ARGUS terminology this is called 'key variables'. For being consistent with the rest of the report, we use QID to refer to these key variables also for μ-ARGUS.

Example of automatic QID patterns in μ-ARGUS: Assume attributes 'age' and 'gender' are assigned with identification level 1, and attributes 'job' and 'postal code' are assigned identification level 2. Then we have the following QID patterns when using the automatic option: (age, job), (age, postal code), (gender, job), (gender, postal code). In the automatic way, it is not possible to define a pattern like (age, gender) or (job, postal code). The latter patterns should be defined manually.

Example of QID pattern values: the QID pattern (gender, job), where attribute 'gender' can be 'male' or 'female' and attribute 'job' can be 'professional' or 'artist', can have the following QID value patterns: (male, professional), (male, artist), (female, professional) and (female, artist).

For those attributes that appear in the QID patterns, μ-ARGUS offers the following data protection methods (Hundepool & Wolf, 2011):

- *Generalisation* in a global recoding way, where selected attributes can be generalised individually based on their structure (if they are defined as hierarchical, like the postal code in the Netherlands) or some user defined recoding taxonomy. For the latter, the user should define a *one-level* taxonomy-tree for every attribute that (s)he wants to generalise accordingly.
- *Top/bottom coding*, where any high/low values of QIDs (being ordinal-categorical or continuous-numerical) that are not generalised as mentioned above are recoded to their maximum/minimum values. This method can also be applied to any distinguishing SATs (like high salaries).

Example of generalisation in μ-ARGUS: Attribute 'age' can be any value from 0 to 99. In μ-ARGUS one can group every ten years together and assign values, for example, 1 to 10 to these groups. Alternatively, one can group every five years together and assign values 1 to 20 to these groups. Note that it not possible to have these two assignments at the same time in a hierarchical way.

Example of top/bottom coding in μ-ARGUS: Assume that attribute salary is a SAT (or a QID), which is not generalised by the aforementioned generalisation in a global recoding way. As very high salaries can be identifying in certain situations, one can recode all salaries above a maximum value to a certain value like 'max' or '100,000 €'.

When the abovementioned generalisations are applied to QID patterns, μ-ARGUS calculates the numbers of the records that appear in the QID value patterns. If the number of the records in a QID value pattern is fewer than a specified number, then these records are called risky records (similar to having fewer than k records in the ECs in the k-anonymity model). For risky records the *local suppression* method is applied, whereby the value of an attribute in the QID value pattern that occurs insufficiently in the microdata set is set to 'unknown'. For example, the QID value pattern (mayor, Amsterdam) is set to (mayor, 'unknown'). Choosing which QID in a QID pattern to be suppressed is done in two ways: (a) the QID with a high entropy (i.e., having a high number of categories/outcomes) or (b) the importance value assigned by users to the QID.

In addition, μ-ARGUS offers other SDC methods for manipulation of QIDs or SATs in order to prevent statistical inference. These methods are:

- PRAM (Post randomisation Method) for global misclassification of categorical attributes based on a probability model;

- micro aggregation for numerical attributes;
- multivariate fixed size micro-aggregation for numerical attributes;
- rank swapping for ordinal/numerical attributes (i.e., swapping within the rank rage of P%);
- synthetic data creation for numerical attributes; and
- adding noise to the weight attribute that specifies the weights of records based on the sampling design (used for disclosure risk estimation).

5.2.2 Offered measures

μ-ARGUS offers three data disclosure risk measures: individual, global and household (Hundepool et al., 2014, Section 3.3).

The *individual risk measure*, implemented from version 4.0 onwards, is a measure of disclosure risk per-record. It is an estimation of population uniqueness for those records that are unique/rare in the sample microdata set (according to their QID value patterns). This estimation of population uniqueness from sample uniqueness is done based on the framework proposed in Benedetti and Franconi (1998) by using the sampling weights, which statistical institutes use to capture how a population microdata set is reduced to a sample microdata set. Local suppression is used to protect risky records. For these risky records the values of QID patterns are suppressed without changing the values of sensitive attributes.

The *global risk measure* provides a measure of risk at a microdata set level (i.e., for all of its individual records). It is expressed in terms of the expected number of reidentified records in the microdata set, where the likelihood of a record reidentification is determined by the record's individual risk measure. For microdata sets with hierarchical relations among their records (e.g., a microdata set consisting of the records of households, where a group of records belong to one group/house), μ-ARGUS defines *household risk measure* by adding up the individual risk measures of the records per household/group. A household, for example, represents the members of a family living together. Starting from a threshold value for the global/household risk measure, one can estimate a threshold value for the individual risk measure. The threshold on the individual risk measure makes it possible to identify and suppress risky records, as mentioned above.

μ-ARGUS uses two simple measures for data utility loss (called information loss in μ-ARGUS), where the loss is induced due to two data protection methods of local suppression and generalisation in a global recoding way (Hundepool et al., 2014, Section 3.5). Specifically,

- Data utility measure for local suppressions: the number of local suppressions. The higher the number of suppressions is, the higher the amount of information loss.
- Data utility measure for global recoding: an information loss measure based on user assigned importance and predefined coding of an identifying attribute (note that how this measure is realised is unclear in the reference cited).

5.2.3 *Overview*

Table 2 summarises the methods, data disclosure risk measures, and data utility measures of the μ-ARGUS software package.

Table 2 A summary of the main μ-ARGUS features

	Specifics	Explanation
Data protection methods		
Generalisation	Full-domain, global recoding (i.e., for all records)	For QIDs (mostly)
Top/bottom coding	For outliers (QIDs and SATs)	For ordinal-categorical or continuous attributes
Suppression (called 'local suppression')	≈ cell (value) suppression, i.e., applied to risky QID value patterns, setting some values to, e.g., 'unknown'	For attributes with high entrop user assigned high importance values
Perturbation	Data-swapping via PRAM	For categorical attributes
	Data-swapping, rank based	For ordinal/numerical attributes
	Synthetic data	For numerical attributes
	Adding noise	For weight attribute
Permutation	Micro-aggregation	For numerical attributes
	Multivariate fixed size micro-aggregation	For numerical attributes
Data protection models		
$\hat{k}$-anonymity		Added to its last release, Version 5.1.3 (22-03-2018)
Data discloser risk measures		
Individual risk	An estimation of population uniqueness based on sample uniqueness	For individual records
Global risk		For all records in a data set
Household risk		For hierarchically structured records
Data utility measures		
Number of local suppressions		
Generalisation loss	Based on attribute importance and amount of generalisation	For every generalised QID

5.2.4 *Data utility and privacy evaluation*

In μ-ARGUS the end-user has to evaluate the privacy and utility of every data transformation carried out, based on the corresponding measures provided by the tool. Should the transformation result in an unsatisfactory trade-off, the end-user needs to repeat the whole data anonymisation operation with a new set of parameters.

5.3 Main functionalities of ARX

ARX[53] is an open source data anonymisation tool for protecting personal microdata. This tool was the result of a cooperation among the chairs of medical informatics, information security, and database systems at the Technical University of Munich between 2011 and 2013.[54] Since 2013, the software has been maintained and fur-

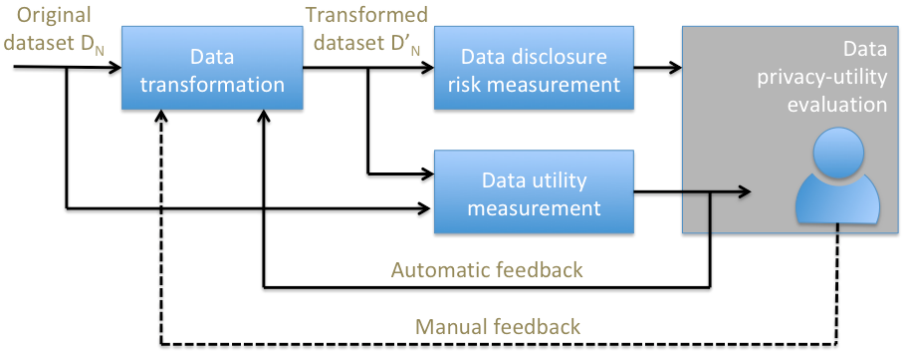
[53] Link: http://arx.deidentifier.org

[54] Link: www.imse.med.tum.de/en/institute-medical-informatics-statistics-and-epidemiology

ther developed by a number of developers[55] under the leadership of the institute of Medical Informatics, Statistics and Epidemiology.[56] In the following, we first describe a number of the functionalities that are specific to this tool, based on its latest release (i.e., Version 3.6.0) and literature. Subsequently, we summarise main functionalities of ARX in a table.

To use ARX, the user chooses a data anonymisation model, like the k-anonymity model or the k-anonymity combined with l-diversity model with the desired values of k or/and l. The chosen model can be perceived as an elementary measure of data disclosure risk, see Subsection 4.3.1. Subsequently, ARX automatically searches for a data transformation that delivers the user-defined data anonymisation model. This automatic search, shown by the 'automatic feedback' link in Figure 13, is unique for ARX (i.e., the other two tools studied do not offer this automatic mechanism). The automatic search algorithm aims at optimizing data utility by using a number of rules – like the monotonicity of data anonymisation models and data utility measures with respect to generalisation only (see Kohlmayer, Prasser & Kuhn, 2015) – and some heuristics, as the problem of finding a transformation that maximises data utility is NP-hard (Prasser et al., 2016c). In ARX, the user can specify the data utility measure in various ways, being measured in the 'data utility measurement' component of the ARX tool shown in Figure 13.

Figure 13 An illustration of the functionalities of ARX



5.3.1 Data transformation

To attain the desired data protection model, ARX uses two methods in the 'data transformation' component, namely:

- full domain generalisation: applied to QIDs according to their predefined taxonomy-trees throughout the whole microdata set (i.e., in a global recoding way);
- record suppression: suppressing all outlier records (i.e., the records of the ECs that do not satisfy the aimed data protection model, e.g., the k-anonymity model or the k-anonymity with l-diversity model).

The tool allows the user to define the maximum percentage of the records to be suppressed. The recommendation is to set the suppression percentage to 100, as the tool's search algorithm does not allow suppressing records extravagantly by default. This self-adjustment stems from the fact that the data utility value goes down rapidly if too many records are suppressed.

[55] Link: http://arx.deidentifier.org/development/contributions/

[56] Link: www.imse.med.tum.de/en/institute-medical-informatics-statistics-and-epidemiology

The search algorithm looks for an appropriate transformation, i.e., specific levels of generalisation for QIDs in their taxonomy-trees and an appropriate number of suppressions for risky/outlier records in the solution space. The search of this solution space aims at maximizing the data utility measure chosen. In short, analysing a transformation from the solution space encompasses the following steps, which are adopted with adaptation from Kohlmayer, Prasser and Kuhn (2015):

1 Initialise the QID generalisation levels to their leaves in the taxonomy-trees.
2 Generalise the microdata set to appropriate levels in the QID taxonomy-trees.
3 Suppress all outlier records, i.e., those that do not satisfy the privacy model in the ECs obtained in step 2.
4 Is the number of suppressed records lower than the given threshold?
   a Yes: The transformation is a candidate solution.
   b No: Try another transformation (go to step 2).
5 Compute the utility of the candidate solution.
6 Put this candidate solution in the list of candidate solutions, where the list is ordered according to the data utility values of the candidate solutions found.
7 Is it expected to find a better solution (i.e., with a higher data utility value)?
   a Yes: Try another transformation (go to step 2).
   b No: Stop.

At the end, ARX provides a list of solutions in decreasing order of the values of the data utility measure chosen. Note that all of the solution found satisfy the data anonymisation model specified by the user in the beginning.

5.3.2 Offered measures

ARX provides a wide range of data disclosure risk measures and data utility measures. These are listed in the table in the following subsection. The foundations of the data disclosure risk measures and data utility measures offered by ARX have already been introduced in Chapters 3 and 4.

Note that there are other data utility measures mentioned in ARX's manual. Examples are: Ambiguity, normalised non-uniform entropy (Prasser, Bild & Kuhn, 2016b), KL divergence, Loss, publisher payout, and entropy-based record-level information loss. For brevity purposes, we omit descriptions of these features and functionalities in this report.

5.3.3 Overview

The main methods and models of ARX for data anonymisation and its measures for data disclosure risks and for data utility are summarised in Table 3. Note that we do not attempt to include all measures and features of ARX exhaustively in Table 3 for preserving the expressiveness of the table content.

Table 3 A summary of the main ARX functionalities

	Specifics	Explanation
Data protection methods		
Generalisation	Full domain, global recoding	
	Local recoding[a]	Applying different generalisation schemes to different parts of a data set
Permutation	Micro aggregation	Combined with either record suppression or different generalisation applied to different parts
Suppression	Record suppression	For outlier records in the ECs that do not satisfy the data protection model
Data protection models (and their possible combinations)		
k-anonymity		
l-diversity	Distinct	
	Entropy diversity	
	Recursive (c, l)-diversity	
t-closeness	Based on generalisation hierarchies	
	δ-disclosure privacy	
δ-presence		
Data disclosure risk measures		
Primitive measures	The values of k, l and δ of the privacy models above	Used within the automatic feedback loop (in Figure 13)
Strict (average) risk	Based on the (average) size of ECs in the sample data set	
Global risk	Using the sampling weights	Based on sampling weights of record the population data set
	Using statistical super population models	
Data utility measures		
Average Equivalence Class Size (AECS)	Based on the sizes of equivalence classes,	Generic purpose measures
Discernibility		
Height	Based on the importance (weight) of attributes, calculated independently for each attribute, then compiled to a total value	
Loss		
Precision		
Non-Uniform Entropy		
Others	Ambiguity & KL-Divergence	
Classification performance measures	Specific purpose, how well the data are for classification, see Prasser et al. (2017)	Specific purpose measures

[a] See https://arx.deidentifier.org/anonymisation-tool/analysis/.

5.3.4 Data utility and privacy evaluation

As mentioned above, ARX provides an internal feedback loop to automatically search for an appropriate transformation, i.e., specific levels of generalisation for QIDs in their taxonomy-trees and appropriate numbers of suppressions for risky/outlier records in the solution space. The search of this solution space aims at maximizing the data utility measure chosen. At the end of this automatic search, ARX provides a list of solutions in decreasing order of the values of the data utility measure chosen.

Subsequently, the end-user can examine the best solution(s) found (one-by-one). To this end, in addition to the data utility values already calculated, ARX allows the user to (visually) inspect and investigate the values of the data disclosure risk measures for different attack scenarios (like prosecutor, journalist and marketer attackers). As such, ARX provides a mechanism to take over some of the responsibilities of the end-user who, otherwise, would have to adjust the parameters of data anonymisation methods manually (e.g., the data generalisation heights and data suppression percentages). Should none of the found transformations fit the privacy and utility trade-off requirements of the end-user, the end-user can initiate a new round of data anonymisation operation with a new data anonymisation model.

5.4 Main functionalities of sdcMicro

sdcMicro is based on the programming language R (Templ, Kowarik & Meindl, 2015). Being an open source, high-level, and extendible statistical computing environment, R has widely been adopted by many researchers and practitioners for statistical data analyses. Therefore, the sdcMicro package can be seen as an important enhancement of R with SDC functionalities. In recent years sdcMicro software has been maintained and further developed by a number of developers.[57] The first version of sdcMicro (i.e., version 1.0.0), released in 2007, was just able to protect small microdata sets (Templ et al., 2015). Almost all methods of sdcMicro in recent releases of sdcMicro (at least since version 4.6.0) are realised in an object-oriented way in C++, resulting in efficient high-performance computations (Templ et al., 2015).

The functional components of sdcMirco are very similar to that of µ-ARGUS, therefore the high-level functional model in Figure 12 applies also to sdcMicro. In the following, we first describe a number of sdcMicro functionalities that are specific to this tool based on its last release (i.e., Version 5.1.0, on 23 March 2018) and literature (Templ, Kowarik & Meindl, 2015; Templ, Meindl & Kowarik, 2017). Subsequently, we summarise main sdcMicro functionalities in a table.

5.4.1 Data transformation

Suppression in sdcMicro is done on the values of a QID[58] for all records that have individual risks above a user-defined threshold (i.e., localSupp()). The tool also offers a method, called localSuppression(), to automatically suppress a minimum number of values in QIDs to achieve k-anonymity.

For micro-aggregation, which is a permutation method typically applied to continuous attributes, sdcMicro offers a number of ways to group the records, as follows:
- *mdav* method, which is based on classical (Euclidean) distance measures;
- *rmd* method, which is based on robust multivariate (Mahalanobis) distance measures;
- *pca* method, which is based on principal component analysis;
- *clustpppca* method, which is based on clustering and (robust) principal component analysis per cluster; and
- *inuence* method, which is based on clustering and aggregation in clusters.

[57] See link https://cran.r-project.org/web/packages/sdcMicro/index.html.

[58] Called "key variable' in sdcMicro.

5.4.2 Offered measures

As a global data disclosure risk measure, sdcMicro provides a *benchmark risk measure*, which is the number of highly risky data records. Risky records can be referred to, for example, those with individual risk values above 0.1 and more than twice of the median of all individual risks. For measuring disclosure risks, sdcMicro also offers a method called *SUDA* (Special Uniqueness Detection Algorithm) that estimates disclosure risks for value patterns of subsets of QIDs (note that the previously mentioned methods estimate disclosure risks for values patterns of all QIDs). The computationally improved version of SUDA is called SUDA2 (Manning et al., 2008). SUDA and SUDA2 are implemented in sdcMicro (but not in its old versions and not in its GUI version called sdcMicroGUI).

In addition to data disclosure risk measures for categorical QIDs (like individual, household and global risk measures), sdcMicro offers measures for calculating risks for continuous attributes. Two of these measures are distance-based record linkage and interval disclosure (Mateo-Sanz et al., 2004).

- In the *distance-based record linkage* approach, one looks for the nearest neighbour of every 'value in the transformed microdata set', where the neighbour is sought in 'the original microdata set'. If the nearest neighbour found is the same as the original data value that corresponds to the transformed data value, then a risky situation arises. The distance-based record linkage measure refers to the total number of such risky situations.
- In the *interval disclosure* approach, one checks whether the original value corresponding to a transformed value falls within a certain interval centred at the transformed value.

As outliers pose more identification risks, sdcMicro offers a mechanism called RMDID2 that is *sensitive for outliers* `(for details see Templ and Meindl, 2008a).

For measuring data utility (i.e., information loss), sdcMicro provides a number of generic and specific measures.

Generic measures are applied to continuous attributes in both original and transformed microdata sets and their results are compared subsequently. These measures can therefore be characterised as precision measures. These generic measures include:

- *Classical measures*: examples include statistical means and co-variance.
- *IL1s measure* (Mateo-Sanz et al., 2004): a scaled distance between original and transformed values of continuous QIDs (sum of those for individual QIDs).
- *eig measure*: the relative absolute differences between the eigenvalues of the co-variances for standardised continuous OIDs of the original and transformed microdata sets.
- *lm measure*: based on regression models (not offered in the sdcMicroGUI).

Specific measures of data utility are basically user-defined and purpose-specific indicators. As such these measures can be perceived as workflow procedures/guidelines. There are two approaches mentioned for defining specific measures:

- The *benchmarking indicators* approach involves selecting a set of (benchmarking) indicators, choosing some criteria for comparison, calculating the indicators, comparing the results, and assessing the results for the transformed microdata set.
- The *model-based* approach involves defining a model that is fitted on the original microdata set, using the model for predicting sensitive attributes with the original and transformed microdata sets, comparing the statistical properties of the model

results for both microdata sets, and assessing the suitability of the transformed microdata set (for the model considered). The lm measure mentioned above can be seen a specific form of this approach.

5.4.3 Overview

Table 4 summarises the main methods and models of sdcMicro for data anonymisation and its measures for data disclosure risks and for data utility. Note that we do not attempt to include all measures and functionalities of sdcMicro exhaustively in Table 4 for preserving the expressiveness of the table content.

Table 4 A summary of the main sdcMicro features

	Specifics	Explanation
Data protection methods		
Generalisation	Full-domain, global recoding	For QIDs, mostly
Top/bottom coding		
Suppression (called 'local suppression')	Local ≈ cell value suppression, i.e., applied to risky QID value patterns, to achieve k-anonymity	To the values of a QID, having high risks
		To the values of some QIDs, in order to minimize # of suppressions
Perturbation	Data swapping using PRAM	Applied to categorical attributes, QIDs or SATs
	Adding noise (to continuous attributes)	Uncorrelated noise
		Correlated noise
		Only to outlying observations
Permutation	Micro-aggregation: Typically applied to continuous QIDs or SATs, by grouping of records, per group aggregate (e.g., average) the values of each attribute independently of other attributes	mdav: grouping based on Euclidean distance
		rmd: grouping based on multivariate distance
		pca: grouping based on principal component analysis
		clustpppca: grouping based on clustering and PCA
		Influence: grouping based on clustering
	Shuffling of continuous QIDs, rank-based shuffling (reverse mapping) within groups	Grouping is conditioned on independent, non-confidential attributes
Data protection models		
k-anonymity		
l-diversity	Distinct	NB: In Templ et al. (2015) the multi-recursive type is also mentioned
	Entropy	
	Recursive	

	Specifics	Explanation
Data discloser risk measures		
Primitive measures	The values of k and l of the privacy models above	In the sample microdata set
Individual risk measures (*: estimations in the population microdata set based on value pattern frequencies in the sample microdata set)	Summing of sample weights for risky QID value patterns	For all QIDs
	Based on supper population models (negative Binomial distribution)	For all QIDs
	Special Uniques Detection Algorithm (SUDA) & SUDA2	For subsets of QIDs
Household risk measure (see * above)	Risk of at least a member of household being reidentified	For hierarchically structured records (called cluster risks)
Global risk measures of an entire microdata set (see * above)	Benchmark approach	Number of individuals/ records with high risk values
	Expected number of reidentification individual risks	Sum of individual (/record) risk values
	Model based estimations	Based on log-linear model
Risk measures for continuous QIDs	Distance based record linkage measure	Number of cases in which the nearest neighbour of a transformed value = the corresponding original value
	Interval disclosure measure	The original value falls within an interval of the transformed value
	Outlier sensitive measures	RMDID2
Data utility measures		
Precision Generic measures, for continuous attributes by comparing the original and perturbed/transformed data	Classical measures	Like means and co-variances
	'IL1s' measure	Scaled distances between original and perturbed values
	'eig' measure	Absolute differences between eigenvalues of co-variances
	'lm' measure	Based on regression models (not present in sdcMicroGUI)
Specific measures	Benchmarking indicators	Choosing, calculating, comparison of indicators based on users' needs
	Modelling based, evaluation of a model constructed from the original microdata set	Use original & transformed microdata sets to predict sensitive attributes, compare results

5.4.4 Data utility and privacy evaluation

Similarly to µ-ARGUS, in sdcMicro the end-user has to evaluate the privacy and utility of every data transformation carried out, based on the corresponding measures provided by the tool. Should the transformation result in an unsatisfactory trade-off, the whole data anonymisation operation has to be repeated with new settings.

5.5 On investigating non-functional aspects

In Subsections 5.2, 5.3 and 5.4, the functionalities of the tools were investigated. However, for evaluating a technology or tool, one can consider three interdependent

aspects. Besides (a) the product itself, these are: (b) interaction between the user and the product i.e., its usability, and (c) experience of using the product, i.e., user experience (Tan, Ronkko & Gencel, 2013). Therefore, in this section we will provide a framework to evaluate the non-functional aspects of these tools, based on their usability. We do not consider the user experience aspect because it is related to nonutilitarian aspects of user-product interactions, where the focus shifts to user affection and sensation and therefore it becomes highly subjective (Tan, Ronkko & Gencel, 2013). We do, however, investigate the experience of using the tools by introducing a usability framework (in this subsection) and designing an experiment (in the following subsection).

Typically, *usability* is characterised by various indicators such as: effectiveness, efficiency, learnability, accessibility, productivity, understandability, generalisability, and safety/error-tolerance (Tan, Ronkko & Gencel, 2013). In order to define our framework for evaluating some of these usability aspects, it is necessary (a) to examine the existential foundation of (in other words, the development methodology behind) these tools and (b) to identify the users that are relevant for our usability evaluation, considering the study context and scope.

The development methodology behind the three tools can be classified as, the so-called, Free/Libre/Open Source Software (FLOSS). In the FLOSS development model, some software developers and domain experts collaboratively create software products. In recent years a considerable number of (successful) software systems have been developed in this way, like GNU/Linux and Android operating systems, Apache tools and technologies (e.g., HTTP server), Mozilla Firefox web server, MySQL and Libre Office (Despalatović, 2013). Despite being developed based on a number of appealing principles, FLOSS products do not generally attract a wide range of users. For example, Linux has not been adopted as much as Windows has been. This can be attributed to a number of issues related to, for example, FLOSS usability, marketing, licence clarity, and user interface (Despalatović, 2013). Therefore, it appears meaningful here to pay attention to the usability aspects of the FLOSS SDC tools.

A useful exercise in usability studies is to investigate the characteristics of usability for each user group (Quesenbery, 2001). Therefore, we first need to identify the user group(s) relevant for our study. To this end, let's look at the four phases of a typical FLOSS project lifecycle (Wynn, 2004) as follows.

- Introductory: the founding developer or a group of developers creates the initial version of the software to demonstrate the vision behind it.
- Growth: users, many of whom are developers, recognise (the need for) the software and start providing feedback about its features, bugs, support requests, etc.
- Maturity: the project reaches a critical mass, i.e., with a maximum number of users/developers. In order to sustain the project, the administration group delegates more tasks to individuals, asking them for self-management and task specific specialism.
- Decline/revive: the project members start losing their interest in the project due to self-management, task delegation and task specialism. Therefore, a small group of project admins, perhaps without the founding developer(s), focus on just the support and maintenance of the existing software. Sometimes the project may revive due to a new innovation by the project members, changing market conditions, a positive response to a new release, etc.

The lifecycles of the three FLOSS SDC tools have passed their introductory phase. Assuming that the SDC tools are in their growth or maturity phase, we categorise their typical users into two groups: (a) active SDC tool developers who actively contribute to or improve the functionality, usability, reliability, performance, etc., of the software and (b) data analysts who want to learn about these tools (and/or, like us, want to understand SDC technologies through experimenting with these tools) and to (possibly) apply them in practice. For our study, the second user group appears to be more relevant than the first group. As such, these SDC tools are meant for experts and specialists – or so-called advanced/power users (Despalatović, 2013), unlike most FLOSS products that are meant for consumers – or so-called simple users (Despalatović, 2013).

Considering the users groups identified above and the nature of FLOSS projects, we define a framework (i.e., a number of indicators) to assess the usability aspects of SDC tools that are relevant for our study. The proposed indicators are devised based on our literature study, the feedback we received from using these tools in an educational setting (by about fifty students), and our own experience (i.e., by two members of the project team). Note that in this study, we do not aim at devising a comprehensive standard to benchmark the non-functional properties of such tools because these tools have heterogeneous features, these tools rely on users (and their expertise) differently, and, most importantly, such a benchmarking falls out of the scope of this study.

Our proposed framework comprises the following criteria:

1. Ease of access or availability: for this indicator, we recognise the following sub-criteria:
   a. being open source;
   b. being free of charge;
   c. being platform independent;
   d. …
2. Ease of use: for determining ease of use we recognise the following sub-criteria:
   a. ease of data import;
   b. ease of data processing;
   c. ease of data export;
   d. having user-interface/GUI;
   e. …
3. Ease of learning: for this indicator, we recognise the following sub-criteria:
   a. availability of documentation, particularly about the data anonymisation methods and models included;
   b. sufficiency/quality of the documentation;
   c. community support;
   d. intuitiveness of the tool, i.e., learning the functionalities and usage of the tool intuitively through interaction with the tool;
   e. …
4. Ease of extension: this indicator looks at how actively the tool is being (or can be) developed and improved by its (dedicated) development team or community. For this indicator, we recognise the following sub-criteria:
   a. integration capability with other software through, e.g., providing APIs;
   b. number of active contributors/developers;
   c. recent activities related to its maintenance, issue list, …;
   d. availability via software development platforms (like GitHub);
   e. developer support through, e.g., having bug/issue list;
   f. …

After defining the usability criteria and sub-criteria, one needs to define the possible values per every sub-criterion. While some sub-criteria can practically assume binary outcomes (like those in category 1: ease of access), the other criteria can be characterised by a multi-level ordinal scale. For example, 'ease of import' can be specified based on a 5 levels scale. Exact specification of these scales and scale values is out of our scope.

5.6 On investigating scalability aspects

In this section, we describe a number of experiments to test some performance aspects of the SDC tools. To this end, we have considered the differences in the functionalities provided by the three SDC tools in order to set up a uniform basis for evaluating their scalability aspects. In other words, our experiments aim at (a) being practically feasible and (b) delivering as much similar tests as possible for these tools.

We start with describing the data preparation activity and then propose an experiment that can be carried out by these tools as similarly as possible.

5.6.1 Microdata set preparation

For our experiments we propose to use a microdata set with a large number of records. We propose to use a real and publicly available microdata set, like the ADULT[59] microdata set in order to create the microdata set needed for the experiments. The ADULT microdata set is an excerpt of 32,561 records from the 1994 US census database. It is widely used by SDC researchers, especially those involved in developing and evaluating the tools studied in this report (see for example Manta, 2013; Dankar, 2012; Prasser et al., 2014, 2016). In order to extend the number of the records of this microdata set, we randomly replicate the records of the ADULT microdata set to get the so-called Extended-ADULT microdata sets. This can be done according to the following scheme, described in a pseudo code:

```
Choose m
For n = 1 to m do
  Choose record-X from data set ADULT randomly;
  Add record-X to data set Extended-ADULT;
  Increase n by one;
  End(do);
End.
```

5.6.2 Experimental design

In this subsection, we describe a number of experiments to tests the SDC tools' performance when the sizes of microdata sets increase or the parameter values of SDC methods vary. We shall describe the tests and their settings (in terms of their purposes and configurations).

One of the SDC models supported by all three tools studied is k-anonymity. Therefore, it appears sensible to investigate their performance in realising k-anonymity.

[59] Download link: https://archive.ics.uci.edu/ml/datasets/adult.

Realising k-anonymity requires taking two steps mainly:

a determining the appropriate generalisation heights for QIDs; and
b generalising QIDs to the determined heights as well as suppressing the outlier records.

Both μ-ARGUS and sdcMicro require the involvement of end-users for step (a), while this step is done automatically in ARX, as illustrated by an inner feedback loop in Figure 13 compared to Figure 12. Based on their experience, end-users define appropriate generalisation heights for QIDs in μ-ARGUS and sdcMicro directly. Subsequently, the tools carry out the generalisation (and suppression) operations determined by end-users. Finally, end-users evaluate whether the privacy-utility trade-offs achieved are satisfactory. If the trade-offs are not satisfactory, then end-users may initiate another round of data generalisation (and suppression). This much user involvement is not required when using ARX, as the inner feedback loop in Figure 13 seeks for the best settings for generalising the heights, given a data utility measure, value of k, and allowed suppression percentage. Based on all generalisation settings found automatically, the end-user in ARX can choose the most suitable one manually. Realising step (b) of k-anonymity is done automatically in all three tools.

As we look for a fair setting to examine/measure the execution times of these tools in realising k-anonymity, we opt for measuring the execution times of step (b) instead of those of steps (a) and (b) together (i.e., the whole time needed to realise k-anonymity). In this way we exclude the influence of human factors in the experiments substantially. In order to measure step (b) of realising k-anonymity, we must have a generalisation setting that can be used for all three tools. Therefore, we have designed our experiments in the following way:

- use ARX to find a number of generalisation settings, ordered according to their data utility measures as calculated by ARX;
- pick up the first generalisation setting from the list above;
- run ARX for the chosen generalisation setting, measure its execution time;
- run μ-ARGUS and sdcMicro for the chosen generalisation setting, measure their execution times.

Note that although the generalisation setting for the three tools become the same according to the scheme described above, (some of) the other configuration parameters still remain tool specific.

5.7 Summary

The SDC tools studied are realised in different application domains (like for protecting medical data or for protecting census data) and by various parties (like universities or national bureaus of statistics). Therefore, these tools do not provide the same set of functionalities, nor do they aim at delivering similar performance measures. Therefore, we tried to provide some insights into their main functionalities and non-functional characteristics. These insights have been instrumental for identifying the main SDC methods and models. Moreover, investigating these tools provided us with valuable hands-on experiences to better understand SDC mechanisms.

For our investigations, we introduced three levels: functional level, non-functional level (i.e., usability), and performance level (scalability). At the functional level we used the high-level model of SDC tools introduced in Chapter 4 as a benchmarking model to give an overview of the main functionalities of those SDC tools. For inves-

tigating the usability aspects of these tools, we introduced a framework to specify the usability aspects relevant for our study. Due to heterogeneity of these tools, particularly in terms of their varying reliance on human operators, we devised a limited number of experiments to examine their scalability with respect to microdata set sizes.

6 Discussion

In this chapter, we discuss the results and insights gained during this study. We reflect on the SDC tools studied and recommend a number of (new) functionalities and features to be included in (future) SDC tools. Further, we elaborate on the (potential) role of SDC tools within the privacy protection processes and procedures of data intensive organisations, particularly, those in the justice domain. This chapter, therefore, aims at giving partial answer to the research question Q_4 (giving insights into promising SDC functionalities or methods). In this chapter we also discuss a number of directions for future research.

We start in Section 6.1 with reflecting on the SDC tools studied. Subsequently, we give a list of the functionalities that are desired to be included in future SDC tools in Section 6.2. In Section 6.3, we argue that SDC tools should be used according to a risk-based approach. In Section 6.4, we remind that using SDC tools is necessary for a responsible way of sharing or opening personal data. Finally, in Section 6.5 we elaborate on future research topics in the legal domain, inspired by the insights gained from our study of SDC technologies.

6.1 Reflection on the studied tools

In order to investigate relevant SDC methods and models, we investigated the SDC methods and models realised in three SDC tools, without intending to be exhaustive (i.e., to describe all functionalities of these tools). The overview aimed at understanding the theoretical foundations (and practical potentials) of SDC technologies. Based on a generic functional model of SDC tools, we categorised the SDC functionalities realised in practice (i.e., those offered in typical SDC tools). This categorisation was based on the set of features reported in the manuals of these tools, in the articles published by the developers of these tools, or the features available in the GUI of these tools. Due to the large number of features available and the much larger number of possible configurations of these features, however, it was not sensible, at least within this study, to investigate how effective the SDC methods and models are implemented in the tools. For such a detailed evaluation of these tools, one may need to investigate the source codes of the methods and models implemented as well as carry out an extensive number of experiments with numerous data sets.

The studied tools rely on human intelligence on a varying degree. ARX provides an internal loop to look for high utility transformations, given a target SDC model. In this approach it is important to use an appropriate data utility measure that assesses the desired utility with a high fidelity. In this case, one can find some of the configuration parameters (e.g., the generalisation heights) of the SDC model effectively (i.e., with a high data utility). Such a tool, therefore, reduces the burden on system users (i.e., data controllers), which makes the tool usable for moderately savvy experts. The other two tools (i.e., µ-ARGUS and sdcMicro), on the other hand, do not have such an automation loop and therefore rely on human expertise and efforts more heavily than ARX does. Therefore, µ-ARGUS and sdcMicro are suitable for highly savvy experts with a lot of experience. They need this to configure SDC tools readily and rapidly, based on their extensive practical experience. This varying

level of reliance on human intelligence, on the other hand, made it difficult for us to devise various experiments for extensive and automatic evaluation of these tools.

The SDC tools studied are realised in different application domains, for instance, for protecting medical data or for protecting census data, by various parties (like universities or national bureaus of statistics), often in an ad-hoc way. Compared to the other two tools, ARX has a substantial number of publications describing its (new) functions, performance, and theoretical foundations. This stems from, we suspect, the fact that ARX is developed in close collaboration with university researchers who, by definition, try to publish their works proactively. The organisations behind the development of μ-ARGUS, and to a lesser degree sdcMicro, are less proactive in sharing the details of their tools. Thus, μ-ARGUS and sdcMicro appear less accessible for newcomers and adopters.

In terms of the number of features, μ-ARGUS offers fewer features than ARX and sdcMicro do. On the other hand, debugging and using μ-ARGUS' features were more difficult than debugging and using the features of ARX' and sdcMicro. It seems that μ-ARGUS is suitable for a small community experienced with the tool, although we have heard news about joining forces between μ-ARGUS and sdcMicro to support each other's functionalities. According to this intention, a user experienced with the front-end of μ-ARGUS, for example, is going to be able to use sdcMiocro functions and libraries seamlessly, and vice versa.

6.2 On desired SDC functionalities

In this section we list a number of SDC functionalities that we find useful to be included in (future) SDC tools, especially for those tools (to be) used for protecting justice domain microdata sets.

6.2.1 Risk assessment with population microdata sets

One important step in determining data disclosure risks is to determine the degree of uniqueness of individuals in the transformed microdata set and in the population microdata set. While a data controller can easily validate sample uniqueness by investigating the transformed microdata set, (s)he cannot readily validate population uniqueness because, traditionally, those population microdata sets have been accessible to intruders and not to data controllers. Not having access to population microdata sets for data controllers is a pity because in some cases, like those of journalist and marketer attackers, population uniqueness can be more relevant than sample uniqueness to determine data disclosure risks.

To define the data disclosure risks associated with population uniqueness, the data controller can use statistical models to estimate the uniqueness in the population microdata set, based on the sample microdata set. To this end, super-population models can be used to estimate the characteristics of the overall population with appropriate probability distributions that are parameterised with the characteristics of the sample microdata set, see Dankar et al. (2012) for an overview. As discussed in Subsection 4.3.2 and Sections 5.2, 5.3 and 5.4, tools μ-ARGUS, ARX and sdcMicro use such estimation models. In those cases where the sample microdata set is randomly sampled down from a population microdata set (for example, some data published by statistics bureaus), one can have the sampling weight per every record in the sample microdata set. Using these sampling weights, one can calculate

the individual disclosure risk of every record in the sample microdata set from the same sample microdata set, and then simply add them up for those records from the same ECs, weighted with the corresponding sampling weights. This weighted summation, as implemented in ARX and sdcMicro, is not accurate and overestimate the disclosure risks.

When random down-sampling is not done, one can (try to) collect the population microdata set, i.e., the microdata set with EIDs and the QIDs (or QID values) for those attributes (or attribute values) that serve as background knowledge for intruders. This option is (or used to be) resource intensive and costly, as it requires updating population microdata sets regularly (Dankar et al., 2012). We argue, nevertheless, that it becomes more feasible nowadays to acquire such a population microdata set as more open data and big data initiatives become available.

For the microdata sets of the justice domain we find it important to have a realistic view of the population microdata set and to determine the data disclosure risks in regard to the population microdata set as accurately and realistically as possible. Therefore, we suggest a new functionality to add to SDC tools for (a) uploading population microdata sets, (b) deriving realistic sampling weights, and (c) applying those weights for deriving disclosure risks as accurately as possible. In order to obtain population microdata sets, future research can rely on big data analytics to pinpoint unique attribute values (like 'mayor of Amsterdam') and to guide defining the QIDs appropriately and adequately. Further research might be needed to calculate disclosure risks, based on sample weights derived from big data analytics.

6.2.2 Automatic data transformation with user involvement

For a data anonymisation model (e.g., the k-anonymity model or the k-anonymity combined with l-diversity model, given the desired values of k or/and l), ARX automatically searches for a data transformation (i.e., appropriate data generalisation heights and/or data suppression values) that delivers the user-defined data anonymisation model while optimizing data utility. The user can choose a desired data utility measure that matches the data usage in mind. The tool solves this optimisation problem by using a number of rules and heuristics. At the end, the user is provided with a number of candidate transformations, prioritised based on their data utility values. Eventually, the user can choose one of the candidate transformations based on, for example, her/his domain expertise.

This automatic data transformation while maintaining user involvement, as we have experienced, is a desired functionality for SDC tools. In this way, the capabilities of users are enhanced while their domain knowledge is paramount in choosing an appropriate data transformation.

6.2.3 Dealing with characteristics of justice domain data

Two key characteristics of justice domain data are: being released continuously and being location dependent.

Example: Consider the police crime microdata set that includes robbery incidents per household (thus being location dependent), where the robbery statistics are released regularly (thus being time dependent).

This case is a typical case of 'continuous data publishing' (Fung et al., 2010). In continuous data publishing, the data controller has previously published microdata sets D_1, …, D_{t-1} and now wants to publish an update microdata set D_t at time instance t. In this scenario, all data releases have the same data scheme (i.e., the same sets of attributes and for every attribute the same range, i.e., the set of values), but every release is different from the previous ones due to insertion and/or deletion of some records.

Example: In case of the police crime microdata set, a household can be removed from a new release if the household is not robbed again in the new time interval, and a new household can be added if it was not robbed in the previous time interval. The set of attributes does not change when the time interval or robbery location changes.

In such a data publishing scenario, the intruder knows the timestamp and QIDs of the victim generally. If no care is taken, it is foreseeable to have record linkage or attribute linkage (Fung et.al, 2010) even if every data release is individually well protected. For example, if only one new record is added to an existing EC, then the intruder can carry out the record linkage attack (assuming that the intruder knows the victim's EC and the interval in which the victim's record is added). Similarly, the same record linkage attack occurs if only one record is removed from an existing EC.

Example: Assume at time t-1 the following table is released about crimes occurred in the last 4 intervals (i.e., those crimes occurred between time instances t-5 and t-1):

Equivalent class	Region	Crime type
1	Downtown	Robbery – without violence
1	Downtown	Attempt for robbery
1	Downtown	Attempt for robbery
2	Northwest	…
2	Northwest	…

Where attribute region is the QID (and each value of this QID represents an EC). Now at time t, the following table is released (i.e., those crimes occurred between time instances t-4 and t), where only the fourth row is added as a new record, compared to the previous table.

Equivalent class	Region	Crime type
1	Downtown	Robbery – without violence
1	Downtown	Attempt for robbery
1	Downtown	Attempt for robbery
1	*Downtown*	*Robbery with violence*
2	Northwest	…
2	Northwest	…

Assume the intruder knows that a crime has occurred for a household in downtown in the time interval between t-1 and t. As only the fourth row is added in the time interval mentioned, then the intruder knows that the type of crime for that household is robbery with violence.

An SDC tool for such a justice domain microdata set should provide a functionality that assists the data controller in preparing these continuous data releases.

6.3 Need for a risk-based approach

Data protection technologies, in general, and SDC tools for data anonymisation, in particular, cannot give 100% guarantees. This lack of guaranteed privacy can particularly be attributed to extrinsic factors in the data environment. For example, background knowledge available to intruders grows enormously with the current explosion of (big) data and advancement of data analytic technologies. As mentioned already, optimizing privacy and utility in the presence of such background knowledge is complex (Mivule & Turner, 2013). This implies that finding an optimal solution that satisfies the myriad requirements of privacy and utility becomes steadily difficult. Therefore, one should be realistic about the potentials of SDC tools and applying these tools should not give a false sense of security nor privacy.

As there is no single solution to deliver guaranteed privacy, many practitioners advocate adopting a risk-based data protection approach instead of a strictly guaranteed data protection one (Cavoukian & Castro, 2014; El Emam, 2010).[60] This requires considering data anonymisation not as a process with a fixed/static binary outcome (i.e., resulting in being anonymous or not being anonymous forever) but as a risk-management process. The GDPR also considers the state of being anonymous dependent on (dealing with) 'all the means reasonably likely to be used … to identify the natural person directly or indirectly' (see Recital 26 of GDPR). The term 'all the means reasonably likely', we suspect, implies considering the risks reasonably and adapting a risk-based approach.

In such a risk-based approach, various mitigation measures might be applied to contain the risk of data disclosure at an acceptable level. The level of risk acceptability depends on many data intrinsic and extrinsic factors and may be subject to changes when new background knowledge becomes available, new technologies are being adopted, new laws are adopted, or the ethical criteria and personal preferences are being adapted.

Therefore, one may not perceive SDC as a silver bullet that can protect data against disclosure risks alone and forever. It should, instead, be perceived as a defence layer in the broader toolset of data controllers. Other technological, procedural, and contractual solutions should also be employed in order to mitigate data disclosure risks adequately.

6.4 SDC tools for data sharing and opening

In light of GDPR, we foresee that SDC tools can be used, or even be necessary, for realising some core principles of GDPR. SDC tools can provide useful insights into data utility aspects and data privacy issues; and are useful for making appropriate trade-offs between data utility aspects and data privacy issues. These insights, in turn, are useful for realising the data minimisation, purpose limitation, and propor-

[60] Note that there are some concerns against the risk based approach, as elaborated upon in Narayanan and Felten (2014) and the discussions at https://freedom-to-tinker.com/2014/07/09/no-silver-bullet-de-identification-still-doesnt-work/

tionality principles of GDPR. To this end, SDC tools provide justifications for and realisation means of making the trade-off choices between data utility and data privacy. As such, SDC tools and methods can be perceived as enablers of data sharing initiatives in general and data opening initiatives in particular.

6.5 On legal aspects

Our study of SDC technologies has resulted in identifying a number of research questions in the legal domain. While all these legal research questions are for future, we discuss them in this section briefly.

6.5.1 Open data and maintaining the original data

According to open data guidelines and our research (see Chapter 2), anonymity plays an important role in opening privacy sensitive data, in particular in the justice domain.[61] According to GDPR a key requirement for data being considered as anonymous is that the data are anonymous for everybody, even for the data controller. Otherwise, the data have to be seen as pseudonymised. We suspect GDPR has adopted this approach because the data controller is theoretically able to link (some) records in the transformed microdata set to the corresponding identities by using the original microdata set. To this end, those sensitive and insensitive attributes that are the same in both the transformed microdata set and the original microdata set, can serve as QIDs for the data controller to link the identity values in the original microdata set to the records in the transformed microdata set.

One option to announce a transformed microdata set as anonymous is to erase the original microdata at the data controller. Due to some reasons (e.g., for archiving purposes and the use of the original microdata for their primary purposes), the data controller must maintain a copy of the original microdata set. In these cases, the transformed microdata is not anonymous and falls within the scope of GDPR.

It is, therefore, for future research to investigate how a transformed microdata set, that satisfies all conditions of GDPR anonymity except the original microdata set being held by the data controller, can be opened in practice. One solution direction would be to define the set of QIDs so that the transformed microdata set (e.g., after generalisation of QIDs) becomes anonymous also for the data controller and thus may be eligible for being made open. Along this solution direction, the research questions are (a) what is the set of appropriate QIDs for opening microdata sets, and (b) what is the impact of defining such QIDs on data utility?

6.5.2 Making trade-offs between privacy and transparency

Wob seems to forbid opening of sensitive personal data in an absolute prohibitive way. According to this, one cannot consider overweighing the benefits of sharing the data, i.e., no trade-offs may be made between the contending values of the rights of individuals on personal data protection versus the public rights on access to information.[62] Meanwhile, Wob seems to provide an exception to the rule of not opening sensitive personal data. According to this exception, sensitive data may not

[61] Whether (or how much) anonymity is a necessary condition for opening a data set is for future research.

[62] Memorie van toelichting Wet openbaarheid van bestuur; Tweede Kamer, vergaderjaar 1986-987 , 19 859, nr. 3.

be opened unless this evidently does not lead to a breach of personal privacy (Wob, Article 10/1/d).[63]

A research question that arises here is how the Wob requirement of 'evidently not leading to a breach' be interpreted according to GDPR. Does this mean that the sensitive personal data should be anonymous in the GDPR sense before being opened? This research question is for future studies, where one should investigate and formalise the legal (and ethical) grounds for establishing a balance between the privacy rights of individuals and the information access rights of the society, particularly in light of GDPR and in relation to already existing laws (like Wob).

6.5.3 Other legal aspects

Based on the reidentification and attribution concepts used in the technological domain, we defined statistical data disclosures in Subsection 3.2.4. Further we noted that in legal terms, reidentification and highly certain attribution are considered as a breach of privacy rights in the UK. It is necessary to align these concepts from the technological domain with their legal counterparts formally (i.e., to determine when a privacy breach occurs legally). Interpreting these technological concepts in light of GDPR seems appropriate and necessary in order to apply SDC technologies within DPIA processes practically.

Data governance should be in place in order to enforce privacy and other fundamental human rights, as required in laws such as GDPR and ECHR. It is necessary to investigate the (needed) legal grounds for data governance. Particularly, processing anonymous data may still have adverse impact on individuals, leading to privacy loss (see Article 8 of the ECHR and Article 7 of the EU Charter of Fundamental Rights for protecting the sphere of an individual's private life). The legal ground for protecting anonymous data in open data settings is a subject of future research.

[63] '(…) tenzij de verstrekking kennelijk geen inbreuk op de persoonlijke levenssfeer maakt" Wob, Article 10, 1-d.

7 Conclusion

In this report we presented an overview of SDC technologies (like SDC methods and models) and the main functionalities of three open source SDC tools. The scope of this study was limited to those SDC methods, models and tools that have been developed for protecting microdata sets against statistical data disclosures. These disclosures, which may arise even when data are accessed legitimately, can be accomplished through, for example, statistical inference. Within this work we have particularly looked at protecting justice domain microdata sets in open data settings.

In this chapter, we provide the conclusions of our study. The chapter is organised per research question, i.e., we present our conclusions about the legal constraints relevant for SDC-based data protection in Section 7.1, about SDC technologies and their main functionalities in Section 7.2, about dealing with background information in Section 7.3, and about promising SDC functionalities in Section 7.4. Finally, we sketch a number of directions for future research in Section 7.5.

7.1 Legal constraints

Concerning research question Q_1 (i.e., the legal constraints relevant for SDC-based data protection, particularly for opening justice domain data) we draw the following conclusions.

Pseudonymisation and anonymisation are two important terms within the domain of SDC technologies. These terms are not defined uniformly and are used differently in legal and technological domains. We elaborated on these terms as used in legal and technological domains in order *to raise awareness about their differences*. Pseudonymisation in the technological domain means replacing explicit identifiers with pseudo-identifiers, while according to GDPR pseudonymisation denotes a personal data protection process where a party can somehow re-identify individuals from the transformed data. Data anonymisation in the technological domain means applying SDC technologies to protect personal data, while according to GDPR the term anonymous data denotes the status of a transformed data set, i.e., that no party can relate to individuals anymore, given *all the means likely reasonably to be used* for identification. In summary, most of the technological data anonymisation mechanisms can be regarded as data pseudonymisation mechanisms in terms of GDPR.

Opening justice domain data has to be compliant with generic and specific privacy laws and regulations. Justice domain data generally contain sensitive personal data (particularly, criminal justice and law enforcement data). Being anonymous data in the GDPR sense plays an important role – if not to say a necessary role – for opening sensitive justice domain microdata sets. Therefore, we also investigated *when a data set can be considered as anonymous according to GDPR*. We concluded that it is plausible to mark data as anonymous after applying appropriate safeguards (e.g., SDC technologies and non-technological procedures) in a way that disclosure risks are contained within an acceptably negligible level (i.e., the risks are below a threshold). This *threshold level is context (and time) dependent* and depends on, for example, available technologies, other data sources, the motivations of attackers and costs of reidentifications. Therefore, data disclosure risks may increase over time, and the currently anonymous data (i.e., anonymous in the GDPR sense) may

become non-anonymous personal data in the future. In this case, one can imagine that the anonymity threshold level rises. We noted that sometimes the threshold level may subside, for instance, in case the background knowledge would no longer be available.

According to GDPR, a necessary condition for the transformed data to be considered as anonymous is that *the data are anonymous for everybody including the data controller*. In case of creating anonymous data sets for open data purposes, it is for future research to investigate the necessity and/or consequences of anonymity at the data controller.

Once transformed data are (considered as) anonymous in the GDPR sense, the GDPR and data protection principles do not apply anymore to the transformed data. Due to increasing background knowledge or new technological developments, the identification of individuals might become possible in the future because one cannot foresee all advancements at the time of data publishing. Consequently, currently anonymous data (i.e., anonymous in the GDPR sense) may adversely be affected by these advancements. This means that the transformed data may become non-anonymous and will fall back within the scope of GDPR. This dynamicity, we argue, may be perceived as the *Achilles heel of the GDPR data protection in open data settings.* Once the transformed data are opened and published on the Internet, the data can no longer be removed (or only with great difficulty). Therefore, it becomes unrealistic to expect that GDPR can be enforced to the transformed data in all regions (as the transformed data may have reached some regions outside of the GDPR jurisdiction).

Processing anonymous data may still have adverse impact on individuals, leading to privacy loss. Here Article 8 of the ECHR and Article 7 of the EU Charter of Fundamental Rights protect the sphere of an individual's private life. Risks of such data processing may be severe, especially when using sensitive justice domain data (such as those about criminal convictions and offences committed by individuals). It is, therefore, important *to devise personal data protection measures that govern the whole lifecycle of anonymous data* (i.e., anonymous in the GDPR sense), especially when/if they turn to become non-anonymous personal data again.

7.2 SDC tools and functionalities

Concerning research question Q_2 (i.e., the main functionalities of available SDC tools for protecting personal data and preserving data utility) we draw the following conclusions.

SDC-based personal data protection technologies rely on a number of basic SDC methods (like removal, suppression, pseudonymisation, generalisation, permutation, perturbation and anatomisation). A subset of these methods is used to realise a specific SDC model (like k-anonymity, l-diversity, t-closeness, k-map or δ-presence). Each SDC tool realises a number of these SDC models to protect microdata sets against statistical disclosures. We studied three non-commercial open source software SDC tools in detail. SDC-based microdata protection tools can be *specified by a generic functional model* that comprises four components, namely: data transformation (transforming original microdata sets to anonymised microdata sets), data disclosure risk measurement (quantifying the data disclosure risks of the transformed microdata set), data utility measurement (quantifying the quality

of the transformed microdata set), and trade-off evaluation (making trade-offs between the disclosure risks and utility of the transformed microdata set). We used this model as a benchmark to categorise the functions of the SDC tools studied.

In order to protect microdata sets against statistical disclosures with SDC technologies, we need to understand the data environment and *specify the relevant data disclosure scenarios*, capturing for example attack types (like record linkage, attribute linkage, table linkage, and probabilistic linkage) and attacker types (like prosecutor, journalist and marketer). One important step in determining the data disclosure risks based on SDC tools is *determining the degree of uniqueness of individuals* in the transformed microdata set (i.e., sample uniqueness) and in the population microdata set (i.e., population uniqueness). Population uniqueness results in sample uniqueness, while sample uniqueness does not necessarily result in population uniqueness.

Compared to protection against population uniqueness, protection against sample uniqueness requires a more severe degradation of data utility. Using sample uniqueness, on the other hand, is too costly for data utility (as requiring sample uniqueness is too pessimistic for assessing risks in those data disclosure scenarios where intruders are uncertain about an individual's record being in the sample microdata set. Remember that one can afford having unique records in the sample microdata set if there are many of such records in the population microdata set). Therefore, in such cases we should *try to adopt the SDC models that deal with population uniqueness* (like $\hat{k}$-map) to have less data utility degradation. Note that, on the other hand, validating population uniqueness is not as easy as validating sample uniqueness. The latter can be done by investigating the released microdata set. For the former, the data controller should either have access to a copy of the population microdata set or estimate the uniqueness in the population microdata set (by using, e.g., the disclosed data set).

To measure data utility in open data settings, general-purpose metrics can be used because they do not consider a specific data usage purpose. Special purpose metrics can be used for those cases where the purpose and usage of the data are well-known at the time of data sharing.

Eventually trade-offs should be made between the data privacy and data utility, given the purpose of data sharing and the data environment in which the data are going to be shared. Sometimes the purpose for which the data are going to be used is known beforehand. In these purpose specific cases, one could aim at attaining a maximum level of personal data protection so that the purpose can be attained. Of course, this target level must be acceptable considering the estimated personal data disclosure risks. Sometimes, like in case of open data, the data should be usable for as many purposes as possible. In these cases, one could define a minimum acceptable level of data disclosure risks and deliver the data with as much utility as possible. In addition to defining the boundary conditions (i.e., of acceptable data privacy and data utility), one should choose appropriate data transformations to result in positive sum outcomes for data privacy and utility. Finding a solution that optimises privacy and utility criteria is complex. *Therefore, studying the strategies for data utility privacy trade-offs based on case studies are recommended for future research*.

In practice, there are heuristics proposed to circumvent part of these complexities and provide near optimal partial solutions. Such heuristics are realised in ARX

through partially-automatic data privacy-utility evaluation functionality. Specifically, ARX provides an internal feedback loop to automatically search for an appropriate data transformation, i.e., specific levels of generalisation for QIDs in their taxonomy-trees and appropriate number of suppressions for risky/outlier records in the solution space. Other tools, like µ-ARGUS and sdcMicro, rely fully on human intelligence to realise the data privacy utility evaluation functionality. In other words, ARX provides a mechanism to take over some of the responsibilities of the end-user who, otherwise, were supposed to manually adjust the parameters of data anonymisation methods (e.g., the data generalisation heights and data suppression percentages).

In addition to considering the functionalities of the SDC tools, we provided a *framework* to examine the non-functional aspects of these tools based on their usability. For this usability we considered only the potential user group of data analysts who want to learn about SDC technologies via studying the literature as well as hands-on experimenting with these tools (so-called advanced/power users). Furthermore, we described a number of *experiments* to test some performance aspects of the SDC tools in realising the k-anonymity model. To this end, we considered the differences in the functionalities provided by the three software tools in order to set up a uniform basis for observing the scalability aspects of the SDC tools. We applied the usability framework and carried out the proposed experiments in a limited scale. It is for future studies to scale up the application of the proposed framework and tests.

Our investigation of the functional (and, to a limited scale, the usability and performance) aspects of the SDC tools show that ARX appears to be more accessible for newcomers and adopters comparatively. In other words, µ-ARGUS and sdcMicro are suitable for more experienced experts relatively.

SDC tools can be used for data sharing as well as for data opening. These tools are important instruments that can be included in the DPIA process to identify and deal with data disclosure risks via data minimisation for a given purpose (thus, applying the data utility privacy trade-offs). The role of SDC tools is to support (thus not to replace) domain experts in identifying data disclosure risks in (large) data sets and to deal with those threats appropriately before opening or sharing them. Applying SDC tools, therefore, can be an important measure of conducting the due diligence principle, indicating that sufficient efforts are in place to protect personal data.

7.3 Background knowledge

Concerning research question Q_3 (i.e., dealing with background knowledge in SDC-based personal data protection) we draw the following conclusions.

Personal data protection requires also considering the extrinsic risk factors, i.e., how data disclosures might occur given, among others, intruders' motivations and means, impacts of disclosures on victims, and the availability/use of auxiliary information sources (i.e., the so-called *background knowledge*) to/by intruders. In order to protect microdata sets with SDC tools, the attributes of microdata sets are divided into four disjoint sets called: explicit identifiers (EIDs), quasi identifiers (QIDs), sensitive attributes (SATs), and non-sensitive attributes (NATs). EIDs can identify individuals directly in microdata sets. Therefore, explicit identifiers are generally removed, suppressed or pseudonymised in order to prevent intrinsic data disclosures. QIDs are those attributes that intruders may use to link the identities available in auxiliary information sources to the corresponding records in the pub-

lished microdata set. In protecting microdata sets via SDC tools, therefore, the *background knowledge available to intruders is captured by appropriately identifying* the QIDs. SATs may have high values for data analytics but, on the other hand, they have also high impacts on the privacy of individuals. Normally they are not transformed for maintaining data usability. SATs are generally specified in legal frameworks (for example, in GDPR, UK's DPA) and they include those attributes capturing the racial or ethnic origin, political opinions, religious beliefs, trade union membership, physical or mental health or condition, sexual life, and some aspects of criminal proceedings of individuals.

There is *no universal and fixed way of attribute mapping*, i.e., defining EIDs, QIDs, SATs and NATs. Attribute mapping depends on the courtesy and skills of data controllers to realistically and carefully identify/estimate QIDs. On the one hand, some QIDs might be overseen during data transformation. Such attributes seem innocent at first sight, for example, the weather condition, but they might enable data linkage when they also appear in auxiliary information sources together with EIDs. On the other hand, background knowledge increases steadily, as we witness nowadays. Therefore, the set of QIDs might grow accordingly in time. As a result, currently anonymous microdata sets may become non-anonymous in the future due to their unprotected attributes that turn to become unprotected QIDs as time passes by. It is for future research to see how data controllers can remain vigilant by monitoring and predicting the background knowledge available to intruders, considering all possible disclosure scenarios. Data controllers, for example, can use big data analytics to estimate the landscape of such background knowledge continuously.

7.4 Promising functionalities

Concerning research question Q_4 (i.e., promising SDC functionalities or methods proposed in literature) we draw the following conclusions.

Investigating the range of feasible functionalities across existing SDC tools, enabled us to develop a vision for joining forces of these tools and/or for extending SDC tools in the future. There are a number of SDC functionalities that we found useful to be included in (future) SDC tools, especially for those tools (to be) used for justice domain data. These useful functions include:

- *Risk assessment based on actual population data set*: For the data of the justice domain we found it important to have a realistic view of the population microdata set and to determine the data disclosure risks in regard to the population microdata set as accurately and realistically as possible. Obtaining population microdata sets can be based on analysing all data available including open data and big data. Future research can investigate whether/how big data analytics can be used to pinpoint unique attribute values (like 'mayor of Amsterdam') and to guide defining the QIDs appropriately and adequately.
- To calculate disclosure risks based on realistic population microdata sets, we suggest adding a new functionality to SDC tools whereby *population microdata sets can be uploaded* to derive realistic sampling weights and, in turn, to calculate disclosure risks as accurately as possible.
- *Automatic data transformation with user involvement*: The automatic data transformation (like generalisation and record suppression) while maintaining user involvement, as we have experienced with ARX, is a desired functionality for SDC tools. In this way, the capabilities of users are enhanced while their domain knowledge is determinant for choosing an appropriate data transformation.

- *Dealing with specificities of justice domain data*: One of the characteristics of justice domain data is that they are released continuously and are location dependent. An SDC tool for justice domain data should provide functionality that assists the data controller in preparing continuous data releases.

Data protection technologies in general and SDC tools for data anonymisation in particular cannot give 100% guarantees. This lack of guarantee can particularly be attributed to extrinsic factors in the data environment. Therefore, one should be realistic about the potentials of SDC tools and applying these tools should not give a false sense of security and privacy. As there is no single solution to deliver guaranteed privacy, many practitioners advocate adopting a risk-based data protection approach instead of a strictly guaranteed data protection one. This requires perceiving personal data protection as a continuous risk management process, not as a onetime operation with a binary outcome (i.e., resulting in being anonymous or not being anonymous forever).

Therefore, one may not perceive SDC as a silver bullet that can protect data against disclosure risks alone and forever. It should, instead, be perceived as a defence layer in the toolset of data controllers. This asks for data controller to be vigilant by monitoring and predicting the background knowledge available to intruders, considering all possible disclosure scenarios.

In light of GDPR, we foresee that SDC tools become necessary for realising some core principles of GDPR. SDC tools can provide useful insights in data utility aspects, data privacy issues, and making appropriate trade-offs between data utility aspects and data privacy issues. These insights, in turn, are useful for realising data minimisation, purpose limitation, and proportionality principles of GDPR. To this end, SDC tools provide justifications for and realisation means of the trade-off choices between data utility and data privacy. As such SDC tools and methods can be perceived as an enabler of data sharing initiatives in general and data opening initiatives in particular.

7.5 Future work

A number of directions for future research have already been mentioned in the previous sections. In the following we mention a number of other research directions.

SDC tools provide a wide range of functionalities, features, and configuration options for data controllers. In practice, it is not trivial to use and configure these tools when there are so many options to choose from. Use and configuration of these tools become even more cumbersome and complex when one considers also the variety of the data to be protected and the diversity of the data environment in/for which the data protection takes place. Further, one needs to be able to interpret and finetune the parameters of SDC tools and methods in order to appropriately support the decision-making process of data anonymisation. Therefore, we recommend conducting further research on how to apply SDC tools to justice domain data, particularly to conduct a number of case studies with real data.

According to GDPR, a necessary condition for the transformed data to be considered as anonymous is that the data are anonymous for everybody including the data controller. Therefore, when a data controller maintains the original (identifying) data, the transformed data (for example after removal or masking of identifiable

data) are not anonymous in a GDPR sense. Because the controller may identify individuals from the transformed data with the help of the original data. In case of achieving anonymity for open data purposes, it is for future research to investigate the necessity and/or consequences of anonymity at the data controller. Further, one can also investigate how the set of QIDs must be defined so that the transformed microdata set (e.g., after generalisation of QIDs) becomes anonymous also for the data controller. To this end, the research questions are (a) what are the appropriate QIDs for making microdata sets anonymous for all parties, and (b) what is the impact of defining such QIDs on data utility?

In the future we need to devise a suitable workflow for using SDC tools in practice, given the data type and environment. Two main challenges of using SDC tools are their configuration and the interpretation of their results. In other words, how can SDC tools and methods be exploited by domain experts to protect data appropriately by making data privacy and utility trade-offs? It is, therefore, for future research to investigate how to choose the methods, models, and parameters of SDC tools in a given context, and how to interpret and finetune the parameters of SDC tools and methods in order to appropriately support the decision-making process of data anonymisation.

Applying SDC tools in practice requires collaboration among legal/ethical experts (regarding data privacy issues mainly), domain experts (regarding data utility aspects mainly), data scientists (regarding to both privacy issues and utility aspects), and data subjects (regarding perceived aspects of privacy). Achieving a collaboration in an effective way is not trivial and it is for future research to develop a methodology for effective collaboration among various stakeholders.

We focused on SDC technologies for protecting microdata sets. In future research, a similar study can be done for SDC technologies that aim at protecting frequency tables, quantitative tables, and (semi-)structured documents. Some case studies are necessary for characterising the requirements of the data sets in the judicial domain for both data sharing and data opening purposes, as far as the data privacy and utility aspects are concerned. For example, whether data are considered anonymous in practice can be investigated per case (or class of cases). We also advise to investigate which measures can be taken in the post-release stage of data opening (like new laws and regulations that prohibit reidentification of anonymous data sets) so that the impact of already opened but currently non-anonymous data can be contained.

Samenvatting

Over Statistical Disclosure Control technieken
Ter bescherming van persoonsgegevens in een open data context

Achtergrond, reikwijdte en onderzoeksvragen
De ontwikkelingen op het gebied van data - in termen van bijvoorbeeld hun volume, variëteit en snelheid - verhogen de risico's op onthulling van persoonsgegevens, ofwel dataonthulling. Enerzijds maakt de groei (in omvang) van een dataset het moeilijk om risico's bij het vrijgeven van data die verborgen zijn in de dataset (dat wil zeggen de *intrinsieke risicofactoren*) te detecteren en onder controle te krijgen. Anderzijds maakt de groei (in omvang of aantal) van andere datasets – ofwel de toename van achtergrondkennis die beschikbaar is voor andere partijen – het moeilijk om de risico's voor het vrijgeven van data te bepalen en onder controle te krijgen: hier betreft het risico's die zich kunnen voordoen bij het combineren van de data met andere datasets (dat wil zeggen de *extrinsieke risicofactoren*). Bijgevolg wordt het voor gegevensbeheerders moeilijker om hun data te openen, dat wil zeggen: hun data te delen met specifieke groepen, individuen of het publiek.
Het vrijgeven van gevoelige informatie over personen kan gebeuren wanneer persoonsgegevens worden overgedragen, opgeslagen of geanalyseerd. Mechanismen voor informatiebeveiliging, zoals dataencryptie en toegangscontrole, kunnen worden gebruikt om data tijdens transport, opslag of analyse te beschermen. Wanneer er al toegang is verkregen tot de data (zij het legitiem of onrechtmatig), is het nog steeds mogelijk om gevoelige informatie over personen onrechtmatig te onthullen (ongeoorloofd gegevensgebruik). Zelfs als direct-identificerende informatie (zoals namen) uit de data wordt verwijderd, kan iemand die (al dan niet op een legitieme of onrechtmatige wijze) toegang heeft verkregen tot die data, statistische onthullingsmethoden gebruiken om sommige data-items alsnog te identificeren, met name door andere informatiebronnen te gebruiken. De term 'burgemeester van Amsterdam' in een dataset kan bijvoorbeeld de identiteit van een persoon onthullen als men al weet wie die burgemeester is of als men dit kan achterhalen met een Google-zoekopdracht. Gegevensbeheerders kunnen op hun beurt de *Statistical Disclosure Control* (SDC-) tools gebruiken om de intrinsieke en extrinsieke risico's voor het vrijgeven van data te verkleinen.
SDC-tools zijn gericht op het elimineren van zowel direct als indirect identificeerbare informatie in een dataset, terwijl de datakwaliteit (dat wil zeggen de bruikbaarheid van de data) zo veel mogelijk wordt gehandhaafd. Direct identificerende informatie (zoals namen en burgerservicenummers) en indirect identificerende informatie (zoals de combinatie van geboortedatum, postcode en geslacht) in een dataset dragen respectievelijk bij aan de intrinsieke en extrinsieke risicofactoren. SDC-tools kunnen worden toegepast op zowel microdatasets als geaggregeerde datasets.
De reikwijdte van deze studie beperkt zich tot de SDC-tools die gericht zijn op het beschermen van microdatasets. Dit zijn datasets die informatie over individuen en individuele eenheden zoals huishoudens bevatten. Binnen deze studie houden we ons met name bezig met het beschermen van datasets uit het justitie domein voor open data doeleinden. Deze focus is gekozen omdat het Nederlandse ministerie van Justitie en Veiligheid van plan is haar open data initiatieven te intensiveren teneinde de transparantie en verantwoording te verbeteren. In deze context is het doel van de studie om SDC-tools te onderzoeken die gericht zijn op het beschermen van

microdatasets. Daartoe definiëren en behandelen we de volgende onderzoeksvragen:

1 Wat zijn de wettelijke beperkingen die relevant zijn voor op SDC-gebaseerde gegevensbescherming, in het bijzonder voor het openen van data uit het justitiedomein?
2 Wat zijn de belangrijkste functionaliteiten van beschikbare SDC-tools voor het beschermen van persoonsgegevens en het behoud van de bruikbaarheid van data?
3 Hoe kan achtergrondkennis worden verdisconteerd in de op SDC-gebaseerde bescherming van persoonsgegevens?
4 Wat zijn (andere) veelbelovende SDC-functionaliteiten of -methoden (voorgesteld in de literatuur)?

Methodologie en resultaten

Om de onderzoeksvragen te beantwoorden, hebben we een uitgebreide literatuurstudie uitgevoerd over de relevante onderwerpen, zoals privacy bevorderende technologieën, SDC-methoden, procedures voor gegevensbeschermingseffectbeoordeling, (nieuwe) wet- en regelgeving en open data initiatieven. Verder hebben we onze tussentijdse resultaten gepresenteerd aan verschillende (expert)groepen (zoals data-analisten, privacy-experts, trainees en hogeschoolstudenten) om de grenzen van de reikwijdte te verfijnen, relevante onderwerpen te selecteren en de resultaten en aanpak te controleren.

Voor het beantwoorden van de eerste onderzoeksvraag hebben we bovendien semigestructureerde interviews afgenomen met drie experts op het gebied van gegevensbescherming die ervaring hebben met privacywetten en -voorschriften. Verder hebben we, om de tweede onderzoeksvraag te beantwoorden, een aantal experimenten uitgevoerd om een voorlopige indicatie te krijgen van de bruikbaarheid en schaalbaarheid van de SDC-tools.

Hieronder beschrijven we in het kort de belangrijkste resultaten van het onderzoek per onderzoeksvraag.

Over wettelijke beperkingen

In het licht van de Algemene Verordening Gegevensbeveiliging (AVG, 2016), kunnen SDC-tools worden gebruikt om de gegevensminimalisatie, doelbinding en proportionaliteitsprincipes te realiseren. SDC-technologieën kunnen met name inzicht verschaffen in en mechanismen bieden voor (a) het transformeren van onbewerkte data, (b) het beoordelen van het nut van de onbewerkte en getransformeerde data, (c) het schatten van de onthullingsrisico's van de onbewerkte en getransformeerde data, en (d) het maken van afwegingen tussen de bruikbaarheid van de data en risico’s verbonden aan het vrijgeven van data. We concluderen dat deze op SDC-gebaseerde inzichten en SDC-mechanismen, noodzakelijk zijn voor gegevensbeheerders om aan de AVG te voldoen bij het delen en openen van hun data.

Pseudonimisering en anonimisering zijn twee belangrijke termen binnen het domein van SDC-tools. Deze termen zijn niet uniform gedefinieerd en worden op verschillende manieren gebruikt in het juridische en technologische domein. We stellen vast dat bijvoorbeeld de meeste data-anonimiseringsmechanismen in de technologische zin kunnen worden beschouwd als data-pseudonimiseringsmechanismen in de AVG-zin. Als onderdeel van de context van onze studie, gaan we in op deze terminologische verschillen.

Data uit het justitiedomein betreffen voornamelijk gevoelige persoonsgegevens (bijvoorbeeld data over strafrechtspleging en wetshandhaving). Het niet opnemen van persoonlijke (identificerende) informatie speelt een belangrijke rol – zo niet een noodzakelijke rol – bij het openen van data uit het justitie domein. Daarom onder-

zoeken we ook wanneer een dataset kan worden beschouwd als zijnde zonder persoonlijke informatie (of anoniem) conform de AVG. Hiertoe stellen we het idee van een drempel voor om de grens van data-anonimiteit te markeren. Deze drempel is in principe afhankelijk van de context (en tijd). Dat wil zeggen dat deze drempel afhankelijk is van bijvoorbeeld beschikbare technologieën en hun vooruitgang, andere beschikbare gegevensbronnen en de motivatie voor en kosten van heridentificering. Daarom kunnen risico's voor het vrijgeven van gegevens in de toekomst toenemen – gegevens die op dit moment anoniem zijn, kunnen niet-anonieme persoonsgegevens worden, omdat de drempelwaarde voor anonimiteit met de tijd toeneemt. Soms kan het drempelwaarde echter afnemen, bijvoorbeeld als de achtergrondkennis verdwijnt.

Over de belangrijkste functionaliteiten van SDC-tools
In deze studie hebben we drie niet-commerciële *open source software* SDC-tools onderzocht, namelijk: μ-ARGUS, ARX en sdcMicro. Enerzijds heeft het onderzoek van de tools ons in staat gesteld om (a) een inzicht te krijgen in de belangrijkste SDC-functionaliteiten, (b) hands-on ervaring op te doen met SDC-tools (door te experimenteren met deze tools), en (c) te leren van de ervaringen van de onderzoekgemeenschap en academische wereld. Anderzijds leidde het onderzoek van de SDC-tools (samen met onze literatuurstudie) ertoe dat we de SDC-tools konden karakteriseren op basis van een generiek functioneel model dat uit vier componenten bestaat:

- *datatransformatie* waarin een originele microdataset getransformeerd wordt naar een microdataset met behulp van SDC-methoden en -modellen;
- *dataonthullingsrisicometing* waarmee de onthullingsrisico's in de getransformeerde microdataset gekwantificeerd kunnen worden door middel van het in overweging te nemen van verschillende onthullingsscenario’s en mogelijke koppelingen;
- *bruikbaarheidsmeting* waarin de gegevenskwaliteit van de getransformeerde microdataset in termen van bruikbaarheid gekwantificeerd wordt; en
- *privacy-utility-evaluatie* waarmee afwegingen gemaakt kunnen worden tussen de onthullingsrisico's en bruikbaarheid van de getransformeerde microdataset.

Met behulp van het functionele model bieden we inzicht in de belangrijkste functionaliteiten van de SDC-tools, per component van het functionele model. De datatransformatie component omvat SDC-methoden (zoals verwijdering, onderdrukking, pseudonimisering, generalisatie, permutatie, perturbatie en anatomisatie) en SDC-modellen (zoals k-anonimity, l-diversity, t-closeness, k-map en δ-presence). Over het algemeen wordt een combinatie van SDC-methoden gebruikt om een SDC-model te realiseren en een combinatie van SDC-modellen wordt binnen een SDC-tool gerealiseerd. De data-onthullingsrisicometing neemt de onthullingsscenario's en aspecten van de mate van uniekheid van data-items in beschouwing. Deze data-onthullingsrisicometing omvat twee categorieën risicometingen: elementaire metingen (zoals de waarden van k en l in k-anonimity en l-diversity) en geavanceerde metingen (die op hun beurt weer steunen op het definiëren van data onthullingsscenario's, zoals het scenario van de openbare aanklager, journalist en marketeer aanvaller). De bruikbaarheidsmeting omvat algemene metingen (zoals de onderscheidingsmaatstaf) en maatregelen voor speciale doeleinden (zoals classificatiemaatstaven en maatregelen voor classificatieprestaties). De privacy-utility-evaluatiecomponent vertrouwt voornamelijk op menselijke expertise om een afweging te maken tussen de onthullingsrisico's en de bruikbaarheid van de getransformeerde microdataset op basis van de hierboven genoemde metingen.

Daarnaast stellen we een raamwerk voor om de niet-functionele aspecten van SDC-tools te onderzoeken, op basis van een bruikbaarheidsperspectief dat relevant is voor onze studie (d.w.z. voor datamanagers die meer willen weten over SDC-tools).

Dit kader omvat de volgende criteria:
1 toegankelijkheid of eenvoudige beschikbaarheid, bijvoorbeeld *open source*, gratis en platformonafhankelijk;
2 gebruiksgemak, bijvoorbeeld eenvoudige import, verwerking van gegevens, en export van gegevens, en een heldere gebruikersinterface;
3 leergemak, bijvoorbeeld beschikbaarheid en kwaliteit van documentatie, community-ondersteuning en intuïtiviteit van de tool;
4 uitbreidbaarheid, bijvoorbeeld integratiemogelijkheid met andere software, aantal actieve ontwikkelaars, recente onderhoudsactiviteiten en ondersteuning door ontwikkelaars.

Ten slotte beschrijven we een experiment voor het testen van een specifiek aspect van de prestaties - de uitvoeringstijd - van de drie onderzochte SDC- tools. Daartoe hebben we de verschillen in de functionaliteiten van de drie SDC-tools meegenomen om zo een uniforme manier te vinden om deze tools te testen. Het experiment heeft tot doel (a) praktisch uitvoerbaar te zijn en (b) zo veel mogelijk vergelijkbare tests voor deze tools te leveren. Het experiment is als volgt opgezet:
- gebruik ARX om een aantal generalisatie-instellingen te vinden, gerangschikt volgens hun datafunctionaliteit, zoals berekend door ARX;
- neem de eerste generalisatie-instelling op uit de bovenstaande lijst;

Voer ARX, µ-ARGUS en sdcMicro uit voor de gekozen generalisatie-instelling, meet hun uitvoeringstijden.

Ons onderzoek naar de functionele aspecten van de SDC-tools laat zien dat ARX relatief meer toegankelijk lijkt voor nieuwkomers en *early adopters*. Maar µ-ARGUS en sdcMicro zijn daarentegen relatief beter geschikt voor meer ervaren experts.

Over achtergrondkennis

Achtergrondkennis – steeds meer beschikbaar voor indringers – is een belangrijke extrinsieke risicofactor. Achtergrondkennis omvat de informatie in voor het publiek beschikbare databanken of directory's (zoals kiesregisters, telefoongidsen, handelsgidsen, registers van beroepsverenigingen), in persoonlijke en informele contacten (vanwege of via bijvoorbeeld lokale nabijheid), in sociale media; of in organisatiedatabases (beschikbaar voor, bijvoorbeeld, overheidsinstanties en commerciële bedrijven). Tijdens het SDC-proces gericht op het in kaart brengen van de attributen worden sommige kenmerken van microdatasets aangeduid als Quasi-ID's (QID's). QID's zijn attributen die indringers kunnen gebruiken om de identiteit van sommige betrokkenen, beschikbaar in externe informatiebronnen, te koppelen aan de gegevensitems in de getransformeerde microdataset. Bij het beschermen van microdatasets via SDC-tools wordt daarom de achtergrondinformatie die beschikbaar is voor indringers vastgelegd door de QID's op de juiste manier te definiëren. We merken op dat er geen universele manier is om attributen in kaart te brengen, bijvoorbeeld om QID's te definiëren. Daarom moeten gegevensbeheerders deze attribuuttoewijzing zorgvuldig uitvoeren binnen een SDC-proces om de risico's te beperken en de onthullingsniveaus op acceptabele niveaus te houden.

Over veelbelovende SDC-functionaliteiten

Onderzoek naar het bereik van SDC-functionaliteiten, dat gebaseerd is op het bestuderen van de drie SDC-tools en de literatuur, heeft ons in staat gesteld een visie te ontwikkelen voor het bundelen van de krachten van deze tools en voor het uitbreiden van deze tools in de toekomst. We identificeren een aantal SDC-functionali-

teiten die nuttig zijn om te worden opgenomen in (toekomstige) SDC-tools, in het bijzonder voor het beschermen van data uit het justitiedomein:

- risicobeoordeling bepaald op basis van werkelijke populatie data (bijvoorbeeld het aantal inwoners van een bepaalde leeftijdscategorie met een specifieke opleiding);
- semiautomatische datatransformatie, maar samen met de bij het proces betrokken gebruikers;
- data-anonimisering op basis van de kenmerken van data uit het justitie domein (omgaan met, bijvoorbeeld, doorlopende publicatie en locatie-afhankelijkheid).

Discussie en vervolgonderzoek

Gegevensbeschermingstechnologieën, in het algemeen, en SDC- tools in het bijzonder, kunnen geen 100% bescherming bieden tegen data-onthullingsrisico's. Dit kan met name worden toegeschreven aan de extrinsieke risicofactoren in de (data-) omgeving. Daarom moet men realistisch zijn over de mogelijkheden van databeschermingstechnologieën en het toepassen ervan mag geen vals gevoel van privacy geven. Aangezien er over het algemeen geen enkele oplossing is om gegarandeerde privacy te bieden, pleiten veel professionals ervoor om een op risico gebaseerde benadering voor databescherming aan te nemen, in plaats van een strikt gegarandeerde gegevensbeschermingsmethode. Dit vereist dat databescherming wordt beschouwd als een continu risicobeheerproces en niet als een eenmalige bewerking met een binaire uitkomst (resulterend in voor altijd anoniem zijn of voor altijd niet-anoniem zijn). Wij denken dat SDC-tools een essentieel onderdeel zijn van een dergelijk risicobeheerproces. Om ervoor te zorgen dat gegevensbeheerders AVG-compliant worden bij het delen en openen van hun data, dienen SDC-tools te worden opgenomen in het proces voor gegevensbeschermingseffectbeoordeling (DPIA). Zij kunnen daarmee de risico's identificeren en controleren middels dataminimalisatie, terwijl de datakwaliteit voor het doel aanvaardbaar blijft. Daarom pleiten wij er ook voor dat de SDC-tools worden gebruikt om domeinexperts te ondersteunen en dus niet te vervangen. *Samenvattend zien we het toepassen van SDC-tools als een noodzakelijke stap voor het realiseren van het zorgvuldigheidsprincipe dat vraagt om voldoende inspanningen om persoonsgegevens in een bepaalde context te beschermen*.

SDC-tools bieden een breed scala aan functionaliteiten, opties en configuratiemogelijkheden voor gegevensbeheerders. In de praktijk is het echter niet eenvoudig om deze tools te gebruiken en te configureren, juist wanneer er zo veel opties zijn om uit te kiezen. Het gebruik en de configuratie van deze tools worden nog omslachtiger en complexer als ook wordt gekeken naar de verscheidenheid van de gegevens die moeten worden beschermd en de diversiteit van de dataomgeving waarin de gegevensbescherming moet worden uitgevoerd. Verder moet men ook de parameters van SDC-tools en -methoden kunnen interpreteren en afstemmen om het besluitvormingsproces van dataminimalisatie adequaat te ondersteunen. *Daarom adviseren wij verder onderzoek te doen naar de toepassing van SDC-tools op justitiële gegevens, met name door een aantal concrete studies uit te voeren met operationele gegevens uit het justitiedomein*.

Ten slotte zien we op basis van de inzichten die in dit onderzoek de volgende mogelijkheden voor toekomstig onderzoek:

- Onderzoek naar de noodzaak en gevolgen van anonimiteit in de AVG-zin, ook bij oe verantwoordelijke voor de gegevensverwerking en voor open-data-initiatieven;
- een workflow ontwikkelen voor het in de praktijk gebruiken van een SDC-tool;
- een leidraad bieden voor de configuratie en interpretatie van SDC-parameters en -resultaten;

- een methodologie ontwikkelen voor effectieve samenwerking tussen verschillende belanghebbenden in het data-anonimiseringsproces, zodat SDC-tools effectief in de praktijk kunnen worden gebruikt;
- een aantal studies uitvoeren om de SDC-vereisten van datasets voor het justitie domein voor het delen van gegevens (inclusief het openen van gegevens) in kaart te brengen;
- het ontwikkelen van aanvullende (wettelijke) maatregelen die nodig zijn voor, tijdens en na het beschermen van gegevens met SDC-tools.

Glossary of terms

Anonymisation (in technological terms): It is characterised as 'a process of ensuring that the risk of somebody being identified in the data is negligible' (Elliot et al., 2016). Data anonymisation aims at hiding the identity and/or the sensitive data of data subjects, while retaining sensitive data for the purpose of data analysis (Fung et al., 2010).

Anonymous information (in legal terms and according to GDPR, Recital 26): It refers to the 'information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable'. (NB: We used the term 'anonymous data' throughout this report alternatively because we were mostly concerned with microdata.)

Attribute mapping: A process whereby the type of every attribute in a microdata assigned. The type of every attribute can be EID, QID, SAT or NAT.

Attribution (or *attribute disclosure*): A process of associating a piece of information with a population unit like a person or a family. Via attribution we learn something new about a (some) person(s).

Auxiliary information sources: Representing background knowledge, auxiliary information sources encompass some QIDs of types 'indirect identifiers', 'key variables', or both, and the EIDs of the corresponding data subjects. The types of auxiliary information sources include: the original microdata sets at the data controller, open data, public registers, social media, proximity knowledge, and personal knowledge.

Background knowledge: Refers to the information that an intruder has access to via external data resources, using which the intruder can disclose personal information from a transformed microdata set.

Data controller: A 'natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data' (Article 4(7) of GDPR, 2016).

DPIA (Data Protection Impact Assessment) process: It is a process 'designed to describe the processing, assess its necessity and proportionality and help manage the risks to the rights and freedoms of natural persons resulting from the processing of personal data by assessing them and determining the measures to address them' (WP29, 2017b).

*Data subject***:** An identified or identifiable natural person to whom personal data refer to.

Equivalent Class: Denoted by EC, refers to a pattern of the values of QIDs.

Explicit Identifier: Denoted by EID, also called direct identifiers, refer to the set of attributes in the original microdata set D that structurally and on their own could uniquely identify an individual, i.e., a data subject.

Extrinsic characteristics of data (in relation to disclosure risks): Refers to those personal information disclosures that are indirectly resulted from having access to a

released form of the data. Such disclosures arise via linking the released data with some background knowledge.

Global recoding: Refers to the case where the generalisation method is applied to all records in a microdata set (see also local recoding).

De-identification: To protect a microdata set against the intrinsic threats by transforming direct identifiers (like names, social security numbers and digitised unique biometrics) via (a) Replacing them with pseudo identifiers, (b) Masking/suppressing them or (c) Removing them. Note that de-identification in North America means anonymisation in technological sense (as in other places).

Identifiable natural person: 'An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person' (Article 4 of GDPR, 2016).

Inference: When the statistical disclosure is not 100% certain, one talks about inference. In other words, 'the capability of a user of some data to infer at high degrees of confidence (short of complete certainty) that a particular piece of information is associated with a particular population unit' (Elliot et al., 2016).

Intrinsic characteristics of data (in relation to data disclosure risks): Refers to those personal information disclosures that are directly resulted from having access to a released form of the data alone.

Intruders (in SDC setting): A party who has either a legitimate or an illegitimate access to some personal data (i.e., internal intruder or external intruder, respectively), and applies (statistical) data analysis (e.g., data linkage and information inference methods) to derive privacy sensitive information from the accessed data illegitimately.

Justice domain data: The term encompasses all data that pertain to the justice branch of the government, ranging from the data of court proceedings and judgments to the data that are gathered within the administration processes and procedures of the whole justice branch of the government. These justice administration and procedural data are often gathered by a number of independent organisations that are involved in a country's justice domain.

Local recoding: Refers to the case where the generalisation method is applied to a few records in the microdata set (see also global recoding). Cell generalisation is such a type.

Non-sensitive Attribute: Denoted by NAT, refers to those attributes that are not Explicit Identifiers, Quasi-Identifiers or Sensitive Attributes.

Original microdata set (or *microdata set*, in short)**:** Denoted by $D_N(A_1, A_2, \ldots, A_M)$, is a relational table with N rows/records, representing individuals and individual units (like households), and M columns/attributes, representing some attributes about those individuals (like their age, gender and occupation).

Population microdata set: A population microdata set includes the records of transformed microdata set.

Pseudonymisation (in legal terms and according to Article 4 of GDPR, 2016): It refers to 'the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person'.

Pseudonymisation (in technological terms): A method whereby direct identifiers are replaced with fictitious values (i.e., a pseudo identifier) that uniquely specify or refer to individual records. Referencing to individuals can be local unique or global unique in a data set (or in a set of related data sets).

Quasi Identifier: Denoted by QID, refers to the set of attributes in the original microdata set D that could 'potentially' identify individuals, i.e., data subjects. This identification is achieved through using the QIDs to link the records of microdata set D with the other microdata sets and knowledge bases wherein both EIDs and QIDs are present for some individuals.

Reidentification (or identity disclosure): A process of attaching an identity to some data items (e.g., to a data record in case of microdata).

Sample microdata set: A transformed microdata set can be considered as a sample of a larger population microdata set that includes the records of the sample microdata set.

Sensitive Attribute: Denoted by SAT, refers to those attributes that capture privacy-sensitive information about data subjects who (possibly) do not want to disclose them. Examples of sensitive attributes are disease, salary, loan, disability status, and crime type.

Statistical data disclosure (i.e., statistical personal data disclosure or, in short, data disclosure): It refers to a reidentification or attribution that occurs with confidence/certainty or at a high-enough degree of confidence/certainty.

Transformed microdata set: Denoted by D'_N or $D''_{N'}$, refers to the transformed microdata set that is resulted from applying some SDC technologies (i.e., SDC methods and models) to the original microdata set D_N.

References

Almasi, M.M., Siddiqui, T.R., & Mohammed, N., & Hemmati, H (2016). The risk-utility tradeoff for data privacy models. In *Proceedings of the 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS), (*pp. 1-5), 24-26 June, Canary Island, Spain. S.l.: IEEE.

ARX manual (2018). ARX – Data Anonymization Tool (online). Retrieved from (on 2 October 2018): https://arx.deidentifier.org/anonymization-tool/.

Bargh, M.S, Vink, M.E., & Choenni, S. (2018). On using obligations for usage control in joining of datasets. In P. Mori P., S. Furnell & O. Camp (Eds.), *Information Systems Security and Privacy (ICISSP*), *Communications in Computer and Information Science (CCIS),* (vol. 867, pp. 173-196). Cham: Springer.

Bargh, M.S., Choenni ,S. & Meijer, R. (2016a). Meeting open data halfway: On semi-open data paradigm. In *Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance (ICEGOV),* 1-3 March. Montevideo, Uruguay. S.l.: ACM.

Bayardo, R.J., & Agrawal, R. (2005). Data privacy through optimal k-anonymization. In *Proceedings of the 21st International Conference on Data Engineering (ICDE)* (pp. 217-228), Tokyo, Japan. S.l.: IEEE.

Benedetti, R., & Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination. In *Pre-proceedings of New Techniques and Technologies for Statistics* (Vol. 1, pp. 225-232). Sorrento, Italy.

Bui, T.V., Nguyen, T.D., Sonehara, N., & Echizen, I. (2015). Trade-off between the price of distributing a database and its collusion resistance based on concatenated codes. In G. Wang, A. Zomaya, G. Martinez & K. Li, K. (Eds.), *Algorithms and Architectures for Parallel Processing (ICA3PP) (*pp. 163-182*)*. Cham: Springer. Lecture Notes in Computer Science, vol. 9529.

Cavoukian, A., & Castro, D. (2014). Big data and innovation, setting the record straight: De-identification does work. Toronto, Ontario: Information and Privacy Commissioner of Ontario. Retrieved from (on 2 October 2018): www2.itif.org/2014-big-data-deidentification.pdf.

Chen, G., & Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata. *Journal of Official Statistics*, *14*(1), pp. 79-95.

Choenni, S., Bargh, M.S., Roepan C., & Meijer, R. (2015). Privacy and security in smart data collection by citizens. In J.R. Gil-Garcia, T.A. Pardo & T. Nam (Eds.), *Smarter as the New Urban Agenda: A Comprehensive View of the 21st Century City* (pp. 348-366). Cham: Springer LNCS.

Choenni, S., van Dijk, J., & Leeuw, F. (2010). Preserving privacy whilst integrating data: Applied to criminal justice. *Information Polity*, *15*(1, 2), 125-138.

Dankar, F. K., El Emam, K., Neisa, A., & Roffey, T. (2012). Estimating the re-identification risk of clinical data sets. *BioMed Central (BMC) Medical Informatics and Decision Making*, *12*(1), 1. Retrieved from (on 2 October 2018): http://doi.org/10.1186/1472-6947-12-66.

De Haan, G, Choenni, S., Mulder, I., & Kalidien, S. (2011). Bringing the research lab into everyday life: Exploiting sensitive environments to acquire data for social research. In S.N. Hesse-Biber (Ed.), *The Handbook of Emergent Technologies in Social Research* (pp. 522-541). New York: Oxford University Press.

Despalatović, L. (2013). The usability of free/libre/open source projects'. *International Journal of Computer and Information Technology*, *2*(5), 958-963.

Directive EU 2016/680 (2016). *Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with re-*

gard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA. Retrieved from (on 2 October 2018): https://publications.europa.eu/en/publication-detail/-/publication/182703d1-11bd-11e6-ba9a-01aa75ed71a1/language-en.

Du Pin Calmon, F. & Fawaz, N. (2012). Privacy against statistical inference. In *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing (*pp. 1401-1408), 1-5 October, Allerton House Monticello, Monticello, Illinois, USA. S.l.: IEEE.

Dwork, C. (2006). Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming* (pp. 1-12). Berlin/Heidelberg: Springer. Retrieved from (on 2 October 2018): http://doi.org/10.1007/11787006_1.

Elliot, M., & Dale, A. (1999). Scenarios of attack: The data intruder's perspective on statistical disclosure risk. In J. Pannekoek, & L. Willenborg (Eds.), *Netherlands Official Statistics, Special issue Statistical disclosure control* (No. 14, pp. 6-10), Voorburg/Heerlen: Statistics Netherlands.

Elliot, M., Lomax, S., Mackey, E., & Purdam, K. (2010). Data environment analysis and the key variable mapping system. In D. Hutchison & J.C. Mitchell (Eds.), *Privacy in Statistical Databases* (pp. 138-142), Berlin/Heidelberg: Springer.

Elliot, M., Mackey, E., Kieron O'Hara, & Tudor, C. (2016). *The anonymisation decision-making framework. A technical report by UK Anonymisation Network (UKAN).* UK: UKAN, Retrieved from (on 2 October 2018): http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-makin g-Framework.pdf.

El Emam, K., Dankar, F.K., Neisa, A., & Jonker, E. (2013). Evaluating the risk of patient re-identification from adverse drug event reports. *BioMed Central (BMC) Medical Informatics and Decision Making*, *13*. Retrieved from (on 2 October 2018): www.biomedcentral.com/1472-6947/13/114.

El Emam, K., & Malin, B. (2014). *Concepts and methods for de-identifying clinical trial data*. Paper commissioned by the Committee on Strategies for Responsible Sharing of Clinical Trial Data.

El Emam, K. (2010). Risk-based de-identification of health data. *IEEE Security & Privacy*,*8*(3), 64-67. Oakland, CA: IEEE.

El Emam, K., & Dankar, F. K. (2008). Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, *15*(5), 627-637.

Fung, B.C.M., Wang, K., Chen, R., & Yu, P.S. (2010). Privacy-preserving data publishing. *ACM Computing Surveys*, *42*(4), 1-53. Retrieved from (on 2 October 2018): http://doi.org/10.1145/1749603.1749605.

GDPR (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC* (General Data Protection Regulation).

Gionis, A., & Tassa, T. (2009). k-anonymization with minimal loss of information, *IEEE Transactions on Knowledge & Data Engineering*, *21*(2), 206-219. S.l.: IEEE.

Hoshino, N. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment. *Journal of Official Statistics*, 17(4), 499-520.

Hundepool, A., & Wolf, P. de (2011). *Methods series: Statistical disclosure control*. The Hague/Heerlen: CBS. Retrieved from (on 2 October 2018): http://research.cbs.nl/casc/Related/Methods_sdc_CBS.pdf

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Wolf, P.-P. de, & Spicer, K. (2012). *Statistical Disclosure Control*. S.l.: John Wiley.

Hundepool, A., Wolf, P.-P. de, Bakker, J., Reedijk, A., Franconi, L., Polettini, S., … Domingo, J. (2014). *μ ARGUS user's manual, version 5.1*. The Hague: CBS. Retrieved from (on 2 October 2018): http://research.cbs.nl/casc/mu.htm.

Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. In *Proceedings of International Conference on Knowledge Discovery and Data Mining* (KDD), (pp. 279-288), Edmonton, Canada. New York: ACM.

Kohlmayer, F., Prasser, F., & Kuhn, K.A. (2015). The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. *Journal of Biomedical Informatics*, *58*, 37-48. Retrieved from (on 2 October 2018): http://doi.org/10.1016/j.jbi.2015.09.007.

Lefevre, K., Dewitt, D.J., & Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE).* Atlanta, GA, USA. S.l.: IEEE.

LeFevre, K., DeWitt, D.J., & Ramakrishnan, R. (2005, June). Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data,* June (pp. 49-60), Baltimore, MD, USA. New York: ACM.

Li, N., Li, T., & Venkatasubramanian, S. (2007). *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity. In *Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE)* (pp.106-115)*,* 16-20 April, Istanbul, Turkey. S.l.: IEEE.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on* Knowledge Discovery from Data, 1(1). Retrieved from (on 2 October 2018): http://doi.org/10.1145/1217299.1217302.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE)*, 3-7 April, Atlanta, GA, USA. S.l.: IEEE.

Mackey, E., & Elliot, M. (2013). Understanding the Data Environment. XRDS: Crossroads, *The ACM Magazine for Students*, *20*(1), 36-39.

Makhdoumi, A., Salamatian, S., Fawaz, N., & Medard, M. (2014). From the information bottleneck to the privacy funnel. In *Proceedings of IEEE Information Theory Workshop (ITW), (*pp. 501-505), 2-5 November 2014, Hobart, TAS, Australia. S.l.: IEEE.

Manning, A.M., Haglin, D.J., & Keane, J.A. (2008). A recursive search algorithm for statistical disclosure assessment. *Data Mining and Knowledge Discovery*, *16*(2), 165-196.

Manta, A. (2013). *Publishing privacy sensitive open data, using an automated decision support tool*. (Master Thesis at Delft University of Technology.)

Mateo-Sanz, J. M., Sebé, F., & Domingo-Ferrer, J. (2004). Outlier protection in continuous microdata masking. In J. Domingo-Ferrer & V. Torra (Eds.). *Proceedings of International Workshop on Privacy in Statistical Databases (PSD)* (pp. 201-215). Berlin/Heidelberg: Springer.

Memorandum (1986). *Memorie van toelichting Wet openbaarheid van bestuur* [Explanatory memorandum to the open government act]. Tweede Kamer, vergaderjaar 1986-1987, 19 859, nr. 3.

Meyerson, A., & Williams, R. (2004). On the complexity of optimal k-anonymity. In *Proceedings of the ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems,* (pp. 223-228). S.l.: ACM.

Mivule, K., & Turner, C. (2013). A comparative analysis of data privacy and utility parameter adjustment, using machine learning classification as a gauge. *Procedia*

Computer Science, *20*, 414-419. Retrieved from (on 2 October 2018): http://doi.org/10.1016/j.procs.2013.09.295.

Narayanan, A., & Felten, E.W. (2014). *No silver bullet: De-identification still doesn't work*. Retrieved from (on 2 October 2018): http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf%5Cnhttps://freedom-to-tinker.com/blog/randomwalker/no-silver-bullet-de-identification-still-doesnt-work/.

Nergiz, M.E., Atzori, M., & Clifton, C.W. (2007). Hiding the presence of individuals from shared databases. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)* (pp. 665-676), 12-14 June, Beijing, China. New York: ACM.

NODA letter (2015), In Dutch: *Kamerbrief over nationale open data agenda 2016,* 2015-0000710114, 30 November. Retrieved from (on 2 October 2018): www.rijksoverheid.nl/documenten/kamerstukken/2015/11/30/kamerbrief-over-nationale-open-data-agenda-2016-noda.

Pitman, J. (1996). Random discrete distribution invariant under size based permutation. *Advances in Applied Probability 28(*2), 525-539. UK: Applied Probability Trust (APT).

Prasser, F., Bild, R., Eicher, J., Spengler, H., & Kuhn, K.A. (2016a). Lightning: Utility-driven anonymization of high-dimensional data. *Transactions on Data Privacy*, *9*(2), 161-185.

Prasser, F., Bild, R., & Kuhn, K.A. (2016b). A generic method for assessing the quality of de-identified health data. In A. Hoerbst et al. (Eds.), *Studies in Health Technology and Informatics,* (Vol. 228, pp. 312-6). S.l.: European Federation for Medical Informatics (EFMI) and IOS Press. Retrieved from (on 2 October 2018) http://doi.org/10.3233/978-1-61499-678-1-312.

Prasser, F., Kohlmayer, F., & Kuhn, K.A. (2016c). The importance of context: Risk-based de-identification of biomedical data. *Methods of Information in Medicine*, *55*(4), 347-355. Retrieved from (on 2 October 2018): http://doi.org/10.3414/ME16-01-0012.

Prasser, F., Eicher, J., Bild, R., Spengler, H., & Kuhn, K.A. (2017). A tool for optimizing de-identified health data for use in statistical classification. In *Proceedings of the 30th IEEE International Symposium on Computer-Based Medical Systems* (pp. 169-174). S.l.: IEEE.

Quesenbery, W. (2001). What does usability mean: Looking beyond ease of use. In *Proceedings of the 48th Annual Conference, Society for Technical Communication*, Chicago: S.n. Retrieved from (on 2 October 2018): www.wqusability.com/articles/more-than-ease-of-use.html.

Rastogi, V., Suciu, D., & Hong, S. (2007). The boundary between privacy and utility in data publishing. In *Proceedings of the 33rd International Conference on Very large Data Bases (VLDB)* (pp. 531-542). Vienna: ACM.

Samarati, P. (2001). Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, *13*(6), 1010-1027.

Salamatian, S., Zhang, A., Calmon, F.D., Bhamidipati, P., Fawaz, N., Kveton, B., Oliveira, P., & Taft, N. (2013). How to hide the elephant – or the donkey – in the room: Practical privacy against statistical inference for large data. In *Proceedings of IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (pp. 269-272), 3-5 December, Austin, Texas, USA. S.l.: IEEE.

Sankar, L., Rajagopalan, S., & Poor, H. (2013). Utility-privacy trade-off in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, *8*(6), 838-852. Retrieved from (on 2 October 2018): http://doi.org/10.1109/TIFS.2013.2253320.

Sweeney, L. (2002a). k-anonymity: A model for protecting privacy. *International Journal on Uncertainty*, *10*(5), 557-570. Retrieved from (on 2 October 2018): http://doi.org/10.1142/S0218488502001648.
Sweeney, L. (2002b). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 571-588.
Sweeney, L. (2000). *Uniqueness of simple demographics in the U.S. population*. *Laboratory for International Data Privacy*, Pittsburgh, PA: Carnegie Mellon University. Technical Report LIDAP-WP4.
Tan, J., Ronkko, K., & Gencel, C. (2013). A framework for software usability and user experience measurement in mobile industry. In *Proceedings of the Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement (IWSM-MENSURA) (*pp. 156-164), 23-26 October, Ankara, Turkey. Washington, DC: IEEE Computer Society.
Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. *Statistical Software*, 67(4). Retrieved from (on 2 October 2018): http://doi.org/10.18637/jss.v067.i04.
Templ, M., Meindl, B., & Kowarik, A. (2017). Introduction to statistical disclosure control (SDC). *International Household Survey Network,* Vienna, 19 October, 2017. S.l.: S.n. Retrieved from (on 2 October 2018): https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf.
Tishby, N., Pereira, F.C., & Bialek, W. (2000). *The information bottleneck method*. arXiv preprint physics/0004057. S.l.: S.n. Retrieved from (on 2 October 2018): http://arxiv.org/abs/physics/0004057.
Verheul, E., Jacobs, B., Meijer, C., Hildebrandt, M., & de Ruiter, J. (2016). *Polymorphic encryption and pseudonymisation for personalised healthcare*. Technical report. S.l.: S.n. Retrieved from (on 2 October 2018): www.semanticscholar.org/paper/Polymorphic-Encryption-and-Pseudonymisation-for-Verheul-Jacobs/7dfce578644bc101ae4ffcd0184d2227c6d07809.
Vlek, C. (2013). How solid is the Dutch (and the British) national risk assessment? Overview and decision-theoretic evaluation. *Risk Analysis*, *33*(6), 948-971.
Wang, W., Ying, L., & Zhang, J. (2014). On the relation between identifiability, differential privacy and mutual-information privacy. In *Proceedings of the 52nd IEEE Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (pp. 1086-1092). S.l.: IEEE. Retrieved from (on 2 October 2018): http://arxiv.org/abs/1402.3757.
Wbp (2000). *Wet bescherming persoonsgegevens*. Retrieved from (on 2 October 2018): http://wetten.overheid.nl/BWBR0011468/2018-05-01.
Wpg (2007). *Wet politiegegevens*. Retrieved from (on 2 October 2018): http://wetten.overheid.nl/BWBR0022463/2018-05-01.
Willenborg, L., & de Waal, Y. (1996). *Statistical disclosure control in practice*. New York: Springer-Verlag.
Willenborg, L., & de Waal, T. (2001). *Elements of statistical disclosure control*. New York: Springer-Verlag.
Wjsg (2002). *Wet justitiële en strafvorderlijke gegevens*. Retrieved from (on 2 October 2018): http://wetten.overheid.nl/BWBR0014194/2016-01-01.
Wob (1991). *Wet openbaarheid van bestuur*. Retrieved from (on 2 October 2018): http://wetten.overheid.nl/BWBR0005252/2018-07-28.
WP29 (2017a). *Opinion on some key issues of the Law Enforcement Directive (EU 2016/680), WP258*. Retrieved from (on 2 October 2018) http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=610178.

WP29 (2017b). *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679. WP248 (rev.01)*. Retrieved from (on 2 October 2018) http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236.
WP29 (2014). *Opinion 5/2014 on anonymisation techniques, WP216*. Retrieved from (on 2 October 2018): https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
Wynn Jr., D.E. (2004). Organizational structure of open source projects: A lifecycle approach. In *Proceedings of the 7th Annual Conference of the Southern Association for Information Systems (SAIS)* (Vol. 47, pp. 285-290). S.l.: SAIS.
Zayatz, L. (1991). *Estimation of the percent of unique population elements on a microdata file using the sample*. Technical report. Washington: Census Bureau.

Appendix 1 List of the persons involved in the study

Members of the project advisory committee

prof.dr.ir. Marijn Janssen (chairman)	Delft University of Technology
dr.ir. Maurice van Keulen	University of Twente
Henk-Jan van der Molen	CISSP/CISM/CISA, Ministry of Justice and Security
mr.dr. Marc van Opijnen	Ministry of the Interior and Kingdom Relations
mr. Just Stam	Ministry of Justice and Security

Interviewees

mr. Pieter de Groot	Ministry of Justice and Security
mr.dr. Bas van der Leij	DPA Professionals
dr. Ellen Beem	Ministry of Justice and Security

Research advisors

drs. Walter Schirm	the Police
drs. Fanny Wallebroek	Ministry of Justice and Security

Reviewers of the report

dr.ir. Sunil Choenni	Ministry of Justice and Security (WODC)
dr. Susan van den Braak	Ministry of Justice and Security (WODC)