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.