Summary

Background, scope and contributions

Governments seek to improve their transparency, accountability and efficiency through proactively opening their publicly funded data sets to the public. In this way, governments intend to support participatory governance by citizens, to foster innovations and economic growth for public and/or private enterprises, and to facilitate making informed decisions by citizens and organizations. Public organizations also share data with others for various reasons like facilitating their operational activities, gaining statistical insights in the status of their operational activities and strategical objectives, and conducting scientific research relevant to their mission such as the impact of their policies. Protecting personal data is an important precondition for governmental organizations for opening and sharing their data responsibly. Minimizing personal data in shared/opened data to the level that is needed for the data usage in mind is one of the main principles of personal data protection. Particularly in open data settings, where the opened data are observable for everybody including potential adversaries, personal data protection boils down to data minimization mainly.

There are various technologies for protecting personal data in a data set. Statistical Disclosure Control (SDC) technologies refer to a subset of personal data protection mechanisms, developed for minimizing personal data while sharing useful data for a given purpose (i.e., maintaining data utility). SDC technologies can be applied to microdata sets as well as tabular data sets. Microdata sets, which may have (very) large sizes, are structured tables with some rows, representing individuals, and a number of columns, representing the attributes of those individuals (like their age, gender and occupation). Tabular data sets are constructed from microdata. A tabular data set contains one or more tables consisting of some rows and columns that correspond to a number of grouping attributes, which are a subset of the attributes of the corresponding microdata. We studied SDC techniques for protecting microdata sets in (Bargh et al., 2018) and for tabular data sets in (Bargh et al., 2020).

The main objective of our research project on personal data protection and SDC technology is to enhance the level of knowledge within the Dutch government, and more specifically, the Ministry of Justice and Security, about SDC technology, its capabilities and limitations, and its usage. Following the previous publications, i.e., (Bargh et al., 2018; 2020), in this report we take another step towards adopting SDC technology within governmental organizations by developing some initial guidelines for using SDC technology in practice. In this way, we expect that SDC technology becomes more accessible to data stewards, who are responsible for sharing or opening data sets responsibly.

Applying SDC technology is a multidisciplinary task requiring, among others, legal and technical expertise. The initial guidelines presented in this report aim at enhancing the technical SDC knowledge and usage skills of data stewards who are entry-level users of SDC technology. The main contribution of this report is to provide an initial set of guidelines concerned with:

- the process of using SDC technology for protecting microdata and tabular data,

- the main actions to be taken in every step of the process of using SDC technology, and
- tthe configuration of specific steps in practice.

Applying SDC technology into practice (i.e., adopting it within organizations) is not a one-off endeavor due to its complexity, multidisciplinary nature and context dependency. Therefore, we also envision and present a framework according to which SDC technology can incrementally and gradually be introduced to and embedded in an organization. The initial SDC guidelines presented in this report serve as one of the first steppingstones of this framework.

Note that throughout this report, we use the term data anonymization to denote the process of removing (direct) identifiers from a data set and applying SDC techniques to it in order to be able to share or open the data set.

Methodology

We used various methods such as literature study, case studies, expert interviews, experiments, prototyping and simulations to develop the SDC guidelines and framework.

Our literature study resulted in two technical reports (Bargh et al., 2018; 2020) that act as the auxiliary reading material for explaining the theoretical foundation of the initial guidelines and for providing some illustrative examples of the methods used in the guidelines presented here. We conducted four case studies chosen from the judicial domain to learn about the current practices for data anonymization. Further, we carried out four expert interviews to base and/or reinforce some of the choices we made for the initial guidelines. We applied the first version of the initial guidelines to four data sets from the justice domain to evaluate the applicability of the guidelines in practice. Some user-interface-related aspects of the initial guidelines were designed and realized in a mockup-type prototype of a SDC software tool. Finally, we conducted a number of simulations with an open-source SDC software tool to support some design decisions we made for the guidelines.

Exercising all these methods allowed us to make design choices for the initial guidelines, narrowing down their scope to a manageable level. Furthermore, they helped us to learn about the limitations of the work and identify a number of directions for future research.

Main results

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

On (the importance of) data anonymization
We have shown that there are a number of driving forces behind personal data minimization. The necessity of limiting the processing of personal data to the purpose in mind is emphasized in various (recent) privacy laws and regulations like the GDPR.

Personal data minimization often starts with removing directly identifying information (like names and social security numbers) from a data set. When

indirectly identifying information needs to be limited, the process of protecting data becomes more complex. In such cases, SDC technology is a main technology used to adjust the amount of indirectly identifying information about individuals in a data set to a desired, required or allowed level, depending on the data usage purpose. To help in this process, SDC technology can provide insights into and mechanisms for:
a tansforming raw data,
b asessing the utility of the original and the transformed data,
c etimating the data disclosure risks of the original and the transformed data, and
d mking trade-offs between data utility aspects and data disclosure risks.

Like any other (data protection) technology, the capabilities of SDC technology should be considered together with some reservations. One of these reservations is that personal data minimization via applying SDC technology does not deliver guaranteed anonymity in the way that the term anonymous is defined in the GDPR. Therefore, SDC technology should not be considered as a silver bullet solution as it does not provide a fully-fledged data protection solution. Nevertheless, using SDC-based insights and applying SDC technology are *necessary* in order to comply with data protection regulations when sharing or opening personal data. In data sharing situations, SDC technology should be paired with complementary technical measures and/or non-technical data governance mechanisms (like contracts, policies and organizational procedures) in order to mitigate residual risks.

On the envisioned organizational embedding framework
SDC technology is a cutting-edge expertise area, being actively researched and continuously developed. The use of SDC technology requires adopting a holistic approach by considering technological, legal, ethical, public and business administration aspects. Furthermore, applying SDC into practice is dependent on many contextual factors like the availability of background knowledge to intruders and the sensitivity level of the shared data. Therefore, we envision and present a framework for embedding SDC technology within an organizational setting.

The envisioned framework includes a structural model to distribute SDC responsibilities within an organization and an iterative organizational learning process to develop relevant SDC knowledge across the organization, based on the rising needs of the organization.

- The envisioned structural model benefits from the advantages of distributing SDC knowledge and skills across local parties within an organization who control privacy sensitive data sets and a central party within the organization who has expertise in SDC technology. A possible implementation could be that routine SDC tasks are delegated to local parties and complex and advanced tasks are delegated to the central party.
- The iterative organizational learning process within our framework aims at gradually delegating the SDC tasks to the local parties as much as possible. In this way, we seek to create SDC expertise at local parties eventually, without imposing immediate burden and accountability on them to learn and apply complex SDC tasks. According to this learning process, the initial set of SDC guidelines is gradually expanded (and/or modified) via learning from practice.

On the generic guidelines
The initial SDC guidelines presented in this report aim at specifying the process of using SDC technology and providing some recommendations for, among others, the protection models and methods as well as their parameter configurations. We start

with a basic set of SDC guidelines that are carefully developed based on our literature study, case studies, expert interviews, experiments, prototyping and simulations. These initial SDC guidelines can be used for education purposes as well as for conducting routine data anonymization tasks in practice by local parties. According to the envisioned evolutionary process for organizational learning, the initial SDC guidelines should be gradually expanded (and/or modified) via learning from practice.

The initial SDC guidelines are divided into generic and specific ones. The specific ones are for protecting microdata or tabular data, which are described in the following section. The generic guidelines describe (the tasks of) the SDC process that are applicable to both microdata and tabular data protection. These generic tasks are for selecting the data to share, specifying the objective(s) of data sharing, specifying the data environment, transforming data, analyzing data and sharing data.

Based on relevant strategic objectives, policies and considerations, the parties acting as data controllers select the data set for sharing. Data selection can be done reactively in reply to a concrete request of data processors or proactively for creating transparency (e.g., in case of Open Data). Specifying the objective of data sharing, which is typically determined outside the data anonymization process, can be used for, for instance, defining some aspects of the data environment, choosing appropriate measures for assessing data utility and data disclosure risks, and making trade-offs between data utility and data privacy. Specifying the data environment is concerned with modelling the context within which the data are shared and utilized. Such a context modelling is important for determining data disclosure risks. The factors for context modelling include the agency (e.g., the intruder types), the auxiliary data sources used by intruders as background knowledge to disclose personal information from a shared data set, data governance mechanisms used for mitigating residual disclosure risks, and the infrastructures used to protect the shared data set or to derive personal information from the shared data set.

The guidelines related to the tasks of data transformation and data analysis are dependent of data type (i.e., being microdata or tabular data) and, thus, we describe them in the following. Finally, data sharing task includes those actions needed after the data set is anonymized satisfactorily. One of the main data sharing actions is the documentation of the data anonymization process for both internal and external usages.

On the data specific guidelines
The data transformation and analysis tasks of the generic guidelines are functionally similar but technically different per data type. Therefore, we present them separately in the report.

The microdata specific SDC tasks are about transforming data and analyzing data. For *transforming data*, in turn, we define a number of tasks, namely: determining the privacy approach, mapping attributes, choosing privacy models and methods, configuring parameters, applying the chosen models, methods and their configurations.

- Two *privacy approaches*, namely syntactic and noise-based approaches, are distinguished for protecting data. We choose the former one to base the initial SDC guidelines due to being truthful and easily linkable to legal privacy concepts.

In the future editions of the guidelines noise-based approaches can also be considered for inclusion.

- For the syntactic approach, the *attribute mapping* task is used to assign four types to the attributes of the shared data set, namely explicit identifiers, quasi identifiers, sensitive attributes and non-sensitive attributes. The explicit identifiers (like names and ID numbers) are generally removed. The quasi identifiers model the background knowledge of intruders who can use them to re-identify data records.
- In the initial guidelines, we suggest using two *privacy models* of k-anonymity and l-diversity to protect quasi identifiers and sensitive attributes. To this end, however, one should be aware of their capabilities and limitations.
- Subsequently, the parameters of the chosen privacy models and methods should be *configured*. To this end, various factors can be considered such as the sensitivity degree of the shared attributes, the sensitivity degree of the shared attribute values, the objective of data sharing, the existence or lack of complimentary data protection mechanisms after data sharing, the type of attackers expected, the reputation of data processors, and the sampling rate and type (i.e., being a random sample or else) of the shared data set with respect to the corresponding population data set, to name a few.
- Using a software tool, the chosen privacy models, methods and parameters can be *applied* to the data set.

For *analyzing data*, which is concerned with the analysis of the transformed data set, one should analyze the data utility and the disclosure risks of the transformed data set. Subsequently, one decides whether a satisfactory trade-off is achieved between data utility and data privacy (i.e., personal data disclosure risks) or not.

Tabular data are an aggregation of microdata. The SDC tasks for protecting tabular data sets are functionally similar to those for protecting microdata sets. Compared to microdata, tabular data have a smaller dimension (i.e., fewer attributes), but may have many more dependencies. These dependencies, which occur for example when multiple tables are produced from a microdata set, may be used to disclose privacy sensitive information. Therefore, tabular data protection focuses more on identifying and resolving these dependencies. This focus makes the process for protecting tabular data slightly different from the process for protecting microdata.

For tabular data specific SDC tasks, which are, in turn, part of the transforming data and analyzing data tasks, we define designing the table, choosing disclosure risk measures, choosing protection methods, configuring parameters, and analyzing data tasks.

- During the *table design*, one determines the desired table type (i.e., frequency table or magnitude table), selects the grouping attributes and their values, and specifies the structure of the table in terms of the existing relations within table and with the other tables that are (going to be) present in the data environment.
- In *choosing disclosure risk measure(s)* one identifies the cells that are at risk based on (a limited number of) sensitivity rules. For the initial set of the guidelines, the suggested rules are frequency rule, p%-rule, and zero cells and skewed distributions.
- Through *choosing protection method(s)*, one tries to mitigate the threats of the cells at risk. To this end, we propose a generic workflow to choose and apply protection methods based on the desired properties of the protected table.
- Via *configuring parameters*, one fine-tunes the parameters of the chosen disclosure risk measures and the chosen protection methods.

- Finally, through analyzing data, one assesses data utility and data privacy (i.e., personal data disclosure risks) based on some measures. Subsequently, one decides whether a satisfactory trade-off is achieved between data utility and data privacy or not.

For both categories of specific SDC guidelines (i.e., those for protecting microdata and those for protecting tabular data), the data transformation and data analysis tasks are iteratively carried out until a satisfactory trade-off is made between data utility and data privacy.

Discussion and follow-up research

As mentioned above, applying SDC technology into practice (i.e., adopting it within organizations) is not a one-off endeavor due to its complexity, multidisciplinary nature and context dependency. Therefore, according to our envisioned framework, SDC technology can incrementally and gradually be introduced to and embedded in an organization. This asks for establishing an iterative organizational learning process to develop relevant SDC knowledge across the organization, based on the rising needs of the organization.

During the design and development of the initial SDC guidelines, we identified a number of issues. As addressing these issues was beyond the scope of the current study, we consider these issues as future research directions and mention some of the most generic ones in the following.

The *usability* of the SDC guidelines and the state-of-the-art tutorial reports was not evaluated *with the target user group(s)*, i.e., the entry-level users (or data stewards) of SDC technology. To this end, it is necessary to:

- Train data stewards on SDC technology, using our state-of-the-art reports and organizing training workshops about SDC tools.
- Develop a high-fidelity prototype for data protection to enable the target user group to gain hands-on experiences with SDC technology. Such a prototype for microdata protection, for example, can be based on the user interface designed in (Rawat, 2020). Note that the designed user interface should still be coupled to an existing SDC tool.

Hereby, moreover, data stewards become familiar with the relevant SDC concepts so that they can provide insightful feedback for SDC experts to develop the guidelines in the future according to the needs of the data stewards.

Proposing detailed and comprehensive guidelines appears to be an impossible task. However, it is possible to *produce a compendium of worked examples* from practical settings, which show the details of the way that a data set is anonymized in a specific case (e.g., how the SDC models and methods are chosen and applied, how the risk and utility are measured, and how the trade-offs are made). Such an example-based approach with a number of worked examples from the real-world scenarios could be used as a basis for how-to knowledge sharing towards data stewards in the future.

There were a limited number of case studies and experiments carried out with real world data sets. In the future, it is desirable *to conduct more case studies* in real world settings (i.e., in a close collaboration with the data stewards). These case

studies can be presented as worked examples and/or be used to fine-tune the initial SDC guidelines.

A future research topic is to investigate *the relation between the initial guidelines and the legal aspects* of personal data protection. For example, it is necessary to determine the required amount of resources (i.e., time, money, employees, etc.) that should be put into the process of data anonymization in a given situation. To this end, applying the due diligence principle is a key legal requirement. Another important research topic is to investigate the ways that one can *adequately model the data environment* in which a data set is shared, considering many existing uncertainties.

Finally, an emerging research direction is to expand the available work on SDC based microdata and tabular data protection to the domain of *protecting unstructured data*, specifically for protecting textual data written in natural languages.