Summary

Background, scope and research questions

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. Protecting the privacy of individuals is an important precondition for governmental organizations for opening their data responsibly. In open data settings, where the shared data are observable for everybody, including potential adversaries, data protection boils down to removing personal data from the shared data, i.e., data anonymization in a technical sense, while maintaining the utility of the data as much as possible.

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. SDC technologies for protecting microdata sets are studied in (Bargh, Meijer and Vink, 2018). In this study, we investigate SDC technologies for protecting tabular data sets.

Tabular data are constructed from microdata. A tabular data set is a table consisting of some rows and columns that correspond to a number of *grouping attributes*, which are a subset of the attributes of the microdata. Any combination of the values of the grouping attributes defines a so-called *cell* in the tabular data set. Often a table contains also *marginals* or *margin* cells that hold the sums of the values of the cells in the corresponding rows or columns. There are two types of tabular data, namely: frequency tables and magnitude tables. In a *frequency table*, the quantitative values in the cells are the counts (or the fractions) of the records in the microdata set that match the grouping attribute values of the corresponding cells. In a *magnitude table*, the cells represent the sums of the quantitative values of the corresponding records in the microdata set.

The objective of this study is to investigate *post-tabular* statistical disclosures of personal data and the SDC technologies for protecting personal data *in tabular data sets*, especially in the context of *non-interactively* opening privacy sensitive data sets (as in the case of, e.g., justice domain data sets). In *post-tabular disclosure control* the SDC technologies are applied to the already aggregated data, i.e., to the cells of tabular data sets, and not to the underlying microdata set. A non-interactive release of a data set means that the data set is defined and shared by a data controller in a single release. In contrast, an interactive release means that the data consumers carry out multiple queries on the original data sequentially.

The main research questions addressed in this study are:

- Q_1: What are the ways for disclosing personal data when tabular data sets are published?
- Q_2: What are the methods for protecting tabular data sets?
- Q_3: What are the main functionalities of available SDC tools for protecting personal data in tabular data sets and preserving data utility therein?

Methodology

To answer the research questions, we carried out desk research on SDC topics for tabular data sets. Additionally, we have analyzed several use-cases within the Dutch Ministry of Justice and Security. These use-cases helped us to identify relevant personal data disclosure threats as well as the typical methods used for preventing them in practice. Further, we have examined the SDC tools for protecting tabular data sets and studied their documentations for gaining insight into the capabilities and limitation of these tools.

Main results

In the following, we briefly describe the main results of the study, which can be mapped to the research questions mentioned above.

On the ways for disclosing personal data in tabular data sets

The types of personal data disclosures in a table can be categorized in three groups:
- Reidentification: Reidentifying an individual contributing to a cell of the table.
- Individual attribution: Learning something (new) about an individual from the table.
- Group attribution: Learning something (new) about a group of individuals.

These disclosures can occur in varying degrees of certainty.

Various factors in the data environment influence the risks associated with personal data disclosures. The background knowledge that is available to intruders constitutes a main factor in the data environment for deriving personal data from a released table. The background information available to intruders can be from the other releases with a similar purpose to that of the released table (for example, when organizations release data about victims via multiple, sequential, continuous and collaborative releases) or from data totally other than the released table (for example, when victims share information about themselves via social networks).

In literature, different attacker types are proposed to model disclosure attacks and some aspects of the data environment relevant for these attacks. Three well-known attacker types are: prosecutor, journalist and marketer attackers. For example, an intruder that wants to learn about a specific individual falls under the prosecutor archetype, whereas an intruder that seeks information about any individual or a specific group can belong to either the journalist archetype or the marketer archetype.

Personal data disclosures in a cell of a table can concern either the number of the contributors to the cell or the distribution of the contributions of the contributors to the cell. The former is important for both frequency and magnitude tables, whereas the latter is only relevant for magnitude tables when certain contributions to the cell have large magnitudes, compared to the other contributions.

Based on our literature study, we have categorized personal data disclosure scenarios for tabular data sets in the following types:
1 Few contributors to a cell, whereby it may become possible to reidentify the contributor(s) to such a cell based on (a) the values of the grouping attributes

that the cell represent and (b) the identity associated with these grouping attribute values that can be found in other databases.
2 Differencing among the values of the cells in a table, due to which one can derive small cell values as specified in the first type above.
3 Differencing among the overlapping sub-populations that appear in a table, due to which one can derive small cell values as specified in the first type above.
4 Linking among the values of the cells in different tables, due to which one can derive small cell values as specified in the first type above.
5 Skewed cell values where the distribution of the values of the cells in a row or column of a table is concentrated in a few cells. Hereby, group attribution may occur for those individuals represented by the cells of the row or column.
6 Dominating contributor to a cell in a magnitude table, where the intruder can infer the contribution of the dominating contributor to the cell. Here we assume that, as background knowledge, the intruder knows about the dominance of that contributor.

While disclosure scenarios 1-5 may occur for both frequency and magnitude tables, the last scenario is exclusive to magnitude tables.

On disclosure risk measures

There are several measures proposed in literature that data controllers can use for specifying disclosure risks. These measures define a disclosure risk slightly differently. Firstly, the risk of disclosure can be measured per cell by indicating whether every cell in a table is safe or not by using sensitivity rules. After determining unsafe cells based on a sensitivity rule, the disclosure risk of an entire table is simply determined by the proportion of the cells in the table that are considered as unsafe according to the sensitivity rules. Examples of sensitivity rules are:

- The minimum frequency rule, which deems a cell as unsafe when there are fewer contributors in the cell than a predetermined threshold value (like 3).
- The dominance rule that assesses a cell in a magnitude table as unsafe when a few contributors to the cell contribute more than a specific percentage of the total cell value.
- The p% rule that considers a cell as unsafe whenever a contributor to the cell (normally the second largest contributor) is able to guess the contribution of another contributor with more than p% accuracy.

Secondly, another measure of the disclosure risk of a table is Subtraction Attribution Probability (SAP), which assumes that an intruder has background information about a random sample of the contributors, and uses this background information to estimate the contributions of other contributors in the table.

Thirdly, conditional entropy measures can be used to estimate disclosure risks based on some properties of a table as a whole. These measures aim at capturing how uniform the distributions of contributors are in the table (e.g., the more cells with zero or small values are, the more unsafe the table is).

On data utility measures

There are also several measures proposed in literature that data controllers can use to specify the change in the utility of (or information loss in) tabular data sets caused by applying disclosure control techniques. Knowing the data usage, which includes understanding the type of the data set and the existing associations (or

correlations) within the data set, can help in determining an appropriate data utility measure.

A simple measure of data utility is the distance between the original tabular data set and the transformed tabular data set. Two example distance metrics are: Hellinger's distance and absolute average distance. With such a distance measure, the data controller can measure how close the transformed tabular data set is to the original tabular data set. SDC technologies affect a table as whole, causing various groups to have different totals than their original values. Distance measures can be used to limit the difference between the original and the transformed totals.

Distance measures do not capture well the changes in variance that are caused by the data transformation. Several measures can be used to indicate such changes in variance. For instance, the Analysis of Variance (ANOVA) measure indicates the difference in variance between a set of grouping attributes and a target attribute.

Although the data usage may not be known beforehand, in some cases a data controller could have an idea of the data usage expected in the data release. If it is known that data consumers are interested in certain data associations, for example a particular correlation, then the data controller can use several association measures as utility measures. Examples of such measures are: Spearman rank correlation, Cramer's V, Pearson correlation and Wilcoxon signed rank-test.

On methods for protecting tabular data sets

Several SDC techniques can be used for protecting tabular data sets. Each technique transforms the data in a different way and provides different properties related to data utility. The data protection methods found can be categorized into two generic categories: 1) non-perturbative methods and 2) perturbative methods.

Non-perturbative methods maintain the truthfulness of attribute values intact. These methods include:

- Suppression, which is achieved by replacing the value of a cell with an empty value or with a symbol to indicate the suppression.
- Conventional rounding, which is achieved by rounding every cell value to its nearest base value.
- Small Cell Adjustment (SCA), which is achieved by adjusting (normally via suppression or conventional rounding) the cells with small values to contain information loss.
- Table redesign and sampling, which is achieved by reconfiguring the values of the rows and columns in a table via, for example, merging the smaller intervals of cell values into larger intervals.

Perturbative methods add an element of noise (i.e., a random value) to the data and do not maintain the truthfulness of attribute values. Examples of perturbative methods are:

- Random rounding, which is achieved by rounding a cell value probabilistically to either the upper base value or the lower base value.
- Controlled rounding, which constrains the rounding of cell values so that the rounded values of internal cells sum up to their respective marginal.
- Controlled Tabular Adjustment (CTA), which aims at adjusting all vulnerable cells (not just small value cells) in a way that they remain at a minimum distance

from their original values. Not only rounding but also other methods are used to change cell values here.
- Cyclic perturbation, which is achieved by adding random noise to the cell values. In every cycle, it transforms cell values pairwise by increasing and decreasing the cell values by 1. Hereby, the technique retains the additivity property in the transformed tables.
- Synthetic data generation, which is achieved by generating a completely new table with similar statistical properties to those of the original table.
- Cell-key, which is achieved by consistently adding noise across tables. To this end, random keys are assigned to data records in the original microdata set, which are in turn used to derive cell keys and determine the amount of noise added to every cell in the table.

Empirical studies are needed to better understand how these data protection methods perform in practice.

Applying SDC technologies essentially requires preserving data utility as much as possible and mitigating data disclosure risk as much as possible. When selecting an appropriate method, a data controller should first choose which data utility properties are important, thereafter, the data controller should look at which appropriate methods provides the best trade-off between reducing disclosure risk and retaining utility.

On SDC tools for protecting tabular data sets

Some organizations involved in collection and processing of personal data (like statistical agencies and universities) have developed SDC tools that make it easier to apply aforementioned SDC methods and measures. In this study, we surveyed the following tools: τ-ARGUS, sdcTable, and CellKey packages, all of which are open source and freely available.

Of the tools surveyed, τ-ARGUS provides the largest number of post-tabular techniques for protecting tabular data sets (like suppression, controlled tabular adjustment and controlled rounding), which are accessible through a GUI. Additionally, τ-ARGUS comes with an extensive manual which includes the theory behind the techniques, some recommended parameter settings, and a practical example of how to use their interface to protect a dataset. Although the manual is very extensive, it still misses certain critical explanations, making it difficult for users to use the tool.

The other packages studied are smaller than τ-ARGUS and provide limited features and no GUI. Their documentation is not as elaborate as that of τ-ARGUS and, therefore, using the other packages requires more preliminary knowledge from the user than using τ-ARGUS does. The CellKey packages do provide a method that is not yet implemented in τ-ARGUS and the sdcTable provides access to τ-ARGUS methods in R, which might be preferred by more advanced users. Furthermore, these tools are in active development, which may improve them in the future.

Discussion and follow-up research

We have identified the scenarios of personal data disclosures for tabular data sets that are currently known in the literature. As these scenarios capture the current state of the art, continuous research will be required to remain aware of the newest

disclosure scenarios. The scope of this research should not be limited to only tabular data, but should also include microdata. Particularly, investigating how the risk of personal data disclosures in microdata relates to that for tabular data can be instrumental to harness the knowledge in one field in the other one.

We have also provided an overview on SDC methods that could be used to protect tabular data against personal data disclosures. The list of possible methods is long and varied. A data controller has to understand the properties of the SDC methods in order to select the correct SDC methods in a given context. We include a list of common SDC properties as well as a table of which SDC methods satisfy which properties to help data controllers in their choices of SDC methods.

Similar to the SDC methods, the usage of the SDC tools is complex and users could need additional guidance. In our follow-up reports, we aim at facilitating the use of SDC technologies in practice. To this end, we have to take into account the data type and the data environment in order to appropriately select and configure SDC methods. This research will be done based on the findings in this report together with expert interviews, case studies and our own previous experience. This work will result in some guidelines for applying SDC technologies in practice.

Another 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. Unstructured data are vastly produced and shared within the justice domain (for example, the textual data produced in court proceedings, verdicts and police reports). Text anonymization is a very difficult and time-consuming task. Research into recognizing subjects in grammatical sentences, such as named entity recognition, can make it possible to identify explicit identifiers automatically. This would help pseudonymization efforts for textual data. Furthermore, there is a practical need for the research on state-of-the-art for identifying objects in grammatical sentences as they may not be identifying on their own, but become identifying when they are combined with other data items (i.e., act as quasi identifier). This future research direction would benefit the anonymization efforts for textual data.