# On Protecting Microdata in Open Data Settings from a Data Utility Perspective

Afshin Amighi[1], Mortaza S. Bargh[2], Sunil Choenni[3], Alexander Latenko[4], Ronald Meijer[5]
[1,2,3] Rotterdam University of Applied science, Research Center Creating 010, Rotterdam, The Netherlands
[2,3,4,5] Research and Documentation Centre Ministry of Justice and Security, The Hague, The Netherlands
Email: [1] a.amighi@hr.nl [2] m.shoae.bargh@wodc.nl [3] r.choenni@wodc.nl [4] a.latenko@wodc.nl [5] r.f.meijer@wodc.nl

*Abstract*—**In modern societies, opening data is playing a crucial role in innovations and economic growth. Public organizations and private enterprises constantly are collecting data. To support the growth of societies, these organizations and enterprises intend to be more active in data opening. However, disclosure of personal data is one of the main threats for data opening. Data transformation techniques for Statistical Disclosure Control (SDC) aim at removing personal data while maintaining the utility of the data at an acceptable level. Applying SDC methods always faces the struggle of maintaining a balance between data utility and personal disclosure risk. In this research, we investigate different options for a common set of transformations for protecting microdata. We study a set of common scenarios which target (or specify) two types of data environments (i.e., those with and without the original microdata sets) and two approaches for privacy protection (i.e., those based on normative heuristics and a formal approach). Employing ARX, we run a series of experiments to observe the behaviours of various measurement factors. At the end, we discuss the consequences of choosing each of the options that can be used by policymakers for opening privacy-sensitive microdata sets.**

***Keywords*–*Data Protection; Disclosure Scenarios; Microdata; Statistical Disclosure Control.***

## I. Introduction

Often, public organizations and private enterprises collect data about citizens and their clients or employees. These parties collect such personal data directly as the input necessary for provisioning their services (like the contact and demographic information about crime-victims or patients). They may collect personal data also indirectly as the byproduct of their service provisioning (when, for example, a judicial or healthcare process proceeds through a chain of actions and interventions). Consequently, personal data are collected and processed in various forms such as microdata, tabular data, semi-structured data as well as unstructured data.

Governments try to improve their transparency, accountability and efficiency through proactively opening their public funded data sets to the public. Hereby, they intend to support participatory governance by citizens, to foster innovations and economic growth for enterprises, and to enable citizens and organizations to make informed decisions. An important precondition for any data opening is to Open Data responsibly, without violating the fundamental human rights such as privacy, liberty, autonomy and dignity [1]–[3]. Considering the scope of this contribution, we focus on the privacy risks or harms associated with such Open Data initiatives. For example, processing personal data may lead to wrong classification of individuals, adversely impacting their liberty, autonomy and income. Even correctly classifying individuals may be harmful and illegal when, for example, individuals become subject to unjustifiable and/or unjust discrimination [4]. Further, linking such personal data to other data sets can reveal even more privacy-sensitive information about individuals than was initially shared [1][5].

Protecting the privacy of individuals in the open data settings, where the shared data is observable for everybody, including the adversaries, boils down to removing personal data from the shared data while maintaining the utility of the data as much as possible. This operation is called data anonymization in a technical sense, which relates to the data minimization, purpose limitation, and accuracy principles of General Data Protection Regulation (GDPR) i.e., Articles 5-1b/1c/1d. Data anonymization is not an easy task as there have been many supposedly anonymized data sets that were re-identified in practice [6].

A common way for data anonymization is by using SDC methods. These SDC methods are applicable to microdata as well as tabular data (i.e., frequency and magnitude tables). Our scope in this contribution is limited to microdata sets which are structured as a table of records/tuples corresponding to individuals, and attributes corresponding to some (privacy-sensitive) properties of those individuals. Generally, applying SDC methods affects data privacy (or personal data disclosure risks) and data utility inversely, i.e., when one increases the other decreases. In practice, therefore, one should make a trade-off among personal data disclosure risks and the utility of opened data.

In Open Data setting, the purpose of data usage is not predetermined. The data consumers (i.e., the public) should be able to apply any (legitimate) analysis they are interested in. This implies that the objective is to transform microdata such that the risk/threat of personal disclosure becomes negligible (i.e., practically eliminated), while the data utility remains high as much as possible. Inspired by [7], we call such a data publishing as *privacy preserving microdata opening*.

In order to apply SDC methods for protecting personal data in Open Data settings, there is a need for gaining an insight in the utility of data when different SDC protection methods (of varying data protection level) are applied. The objective of this study is to empirically investigate the impact of applying the SDC methods on data utility when opening sensitive microdata sets. Specifically, our research questions can be formalized as: *For a common set of data transformation options, what are data utility and data disclosure implications? What are the implications of these options, which policymakers should consider when opening privacy sensitive microdata sets?*

For this study, we have carried out desk research, expert interviews, and extensive experiments with an SDC software tool called ARX [8]. We identify a number of cases that are relevant for data opening. The results of this study can clarify the difference between the identified cases and demonstrate some of the implications of opting for each of these cases. Legal consultants, legislators and policymakers can use the study results when choosing their strategy for opening privacy-sensitive microdata sets.

The remainder of this paper is organized as follows. In Sections II and III, we provide some related work and the background of and the motivations for data opening. In Section IV, we present two generic approaches for applying SDC methods and in Section V, we define the data quality and data privacy measures used for our experiments. Subsequently, we provide the experiment cases in Section VI. Then, in Section VII, we present the experiments results obtained for a publicly available data set, we reflect on the study results and lay down the limitations of the study. Finally, we draw our conclusions and mention some future research directions in Section VIII.

## II. Related work

One criterion for defining our cases is whether or not the intruder has access to the original microdata set. The issue of whether the original microdata can be used for re-identification of some individuals from a technically anonymized microdata set is elaborated upon in [9] and [10]. Elliot et al. [9] discuss the UK privacy regulations, which recognize two environments: one with the original microdata (as it is often the case for the environment of the data controller) and the other without it. In the first environment, the technically anonymized microdata set is considered as personal data while in the other it is considered as an anonymized data. This is because the entity in possession of the original data (e.g., the data controller) can use the original data set to re-identify some individuals in the technically anonymized microdata set. They also mention that an anonymized data set that is re-identifiable for some party (like data controllers) is personal in the EU jurisdiction. Similarly, El Emam and Malin [10] emphasize that when data controllers are able to re-identify some individuals with the original data set, the data set is not anonymous.

In [11], we build on this argument and further argue that, based on an investigation of the relevant legal regimes, criminal justice data cannot be opened when they are personal. Further, we note that, unlike the claim in [9], the condition of not being personal for everybody (i.e., being anonymous in a GDPR sense) is not unanimously accepted (yet) as a precondition for opening privacy sensitive data sets (like criminal justice system data sets). In this contribution, we extend our previous work [11] by defining a number of possible options for protecting microdata sets against re-identification by parties with and without original microdata sets. Subsequently, we investigate the utility of the resulting microdata sets.

Further, we investigate the utility of an anonymization method that yields a sound (i.e., formally provable) data protection mechanism according to a new definition of privacy ($\epsilon$-differential privacy). The need for formal approaches to define privacy and realize personal data protection rigorously is at the centre of focus in recent studies [12]–[14]. The authors argue that past technologies for protection against personal data disclosure rely on intuitive, heuristic understandings of privacy, and the privacy regulations have often endorsed such heuristic techniques implicitly or explicitly. For example, by making an implicit assumption that re-identification may primarily (or even solely) occur via record linkage, where a record is de-identified by those in a publicly available data set, ”many privacy regulations require protecting personal information that can be linked to an individual in order to safeguard against record linkage” [12]. Such regulations, which capture some aspects of normative privacy, do not satisfy all expectations of privacy protection. Therefore, these studies ask for more understanding of the gaps between technical/formal approaches to privacy and the normative approaches to privacy so that future privacy regulations can be improved. Inspired by these works, we investigate the impact of applying a formal privacy protection method (specifically, $\epsilon$-differential privacy) on data utility and compare it with a heuristic normative approach (i.e., $k$-anonymity) as often applied against record linkage attacks.

Based on the impacts of such solution directions on data utility and on privacy risks, as presented in this contribution, policymakers can make an informed strategy for opening their privacy sensitive microdata sets.

## III. Opening Microdata

### A. Motivation(s)

Governments seek to improve their transparency, accountability and efficiency through proactively opening their publicly funded data sets to the public. Via Open Data, governments intend to support participatory governance by citizens, to foster innovation and economic growth, and to empower citizens and businesses for making informed decisions.

Often, public organizations and privacy enterprises possess personal data about citizens as well as clients, employees or partners in the form of microdata sets. Microdata records may include (privacy-sensitive) properties of individuals (like demographic, behavioral, health and/or business information).

In order to achieve the objectives of Open Data, namely transparency, accountability and efficiency, public organizations strive to open their microdata sets as raw as possible. But, microdata sets pertaining to natural persons (very often) contain (sensitive) personal data (like demographic, behavioral, health and/or business information). Opening such microdata as raw as possible, therefore, can inflict (severe) privacy breaches (i.e., personal data disclosures) with adverse impacts on the fundamental human rights as well as on individuals' dignity, liberty, autonomy and income [1]–[3]. Therefore, protecting the privacy of citizens and individuals is an important precondition for (governmental) organizations in order to open their data responsibly [11].

Further, for validation and reproduction of their results, scientists and scholars are supposed to make their research data available for their peers and the scientific community. These research data are often in the form of microdata. In these cases, the protection of personal data is also one of the preconditions for conducting these researches and, even more importantly, for sharing the research data with the scientific community.

### B. Opening Personal (Sensitive) Data

Personal data refer to any information that relates to an identified or identifiable natural person (so-called a *data subject*). One can distinguish several types of personal data in legal domains. For example, GDPR discerns three personal data types: Directly identifiable data, indirectly identifiable data, and sensitive data. Directly identifiable data relate to a person straightforwardly, for instance, someone's name or address. Indirectly identifiable data do not relate to a person straightforwardly but may influence the way a person is perceived or treated in the society (for instance, the type of someone's house or car), or may contribute to someone's identification when combined with other data sets. Sensitive data are related to the fundamental rights and freedom of individuals. According to GDPR, sensitive personal data are of two types: (a) Special categories of personal data such as someone's racial or ethnical 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; and (b) the personal data related to criminal convictions and offences. If sensitive data are (or can be) related to an identified or identifiable natural person, they may be processed only if the data processing complies with strict data protection measures. Bargh et al. [11] argue that such sensitive data sets can be opened to the public if they are without personal information, i.e., they cannot be related to identified or identifiable natural persons.

A data set can be regarded as without personal information in a given, so-called, *data environment*. When a data controller transforms a microdata set to a protected one, and shares the result with a partner organization, the boundary of the partner organization defines a data environment. Within the scope of this study (i.e., opening data to the public), two types of data environments are interesting to investigate, namely: those with the original microdata set and those without it. Making this distinction is based on the fact that the original microdata set is one of the richest knowledge bases that can be used for linking, via re-identification or attribution [15], the records or attributes in a protected microdata set to natural persons. This richness can be associated with the facts that the protected microdata set is the result of applying SDC methods to the original microdata set and that the original microdata set itself contains one or more identifying attributes (like names and social security numbers). These identifying attributes together with the other attributes in the original microdata set can facilitate linking the records in the protected microdata set to the corresponding identities (thus, to re-identify the records in the protected microdata set). A typical data environment with the original microdata set is that of the data controller.

In [11], it is shown that protecting a microdata set for a data environment without the original microdata set delivers a transformed microdata set that is anonymous in a GDPR sense (i.e., being anonymous for everybody in that data environment), while protecting a microdata set for a data environment with the original microdata set delivers a transformed microdata set that is pseudonymized in a GDPR sense (i.e., being potentially identifiable for a party, for example, the data controller, who is possession of the original microdata set). Note that these two types of data environments exists in Open Data settings in cases where the data controller does not maintain or maintains, respectively, a copy of the original microdata set. In Section III-D, we elaborate further on these two data environments types (i.e., those with the original microdata set and those without it).

### *C. Protecting Microdata in Open Data Setting*

In the context of Open Data, the data spread over and reach all areas, some of which fall out of GDPR jurisdiction. As mentioned above, GDPR requires that sensitive personal data are processed (i.e., shared in case of Open Data) with strict data protection measures. The data protection mechanisms that can be applied to this setting are those that minimize data by stripping off, ideally, all the personal data from the data to be opened. To this end, the data minimization mechanisms can be applied via the following processes:

- Data anonymization: This process ensures "that the risk of somebody being identified in the data is negligible" [9]. Data anonymization aims at hiding the identity and/or the sensitive data of data subjects, while retaining sensitive data for the purpose of data analysis [16]. To achieve this, the so-called SDC methods and tools are used.
- Data de-identification: This process aims at protecting a microdata set against the intrinsic threats by transforming direct identifiers (like names, social security numbers and digitized unique biometrics). This transformation is carried out via replacing direct identifiers with pseudo identifiers, masking/suppressing them or removing them.

Note that the term anonymization above is used in a technological sense (and not in a GDPR sense). Further, note that the term de-identification in North America means anonymization in a technological sense. As part of strict data protection measures, cybersecurity controls such as access control and cryptography are not suitable for protecting data in Open Data settings. In other words, data disclosure threats due to cybersecurity attacks of Information System (IS) hacking and due to decrypting encrypted personal data, while data being in transit, storage and processing, are out of the scope in Open Data settings.

### *D. Attribute Mapping*

The process of applying SDC methods for data anonymization and de-identification starts with the subprocess of dividing the set of the attributes of a microdata set into various categories. This subprocess is called *attribute mapping*. To describe attribute mapping, we start with formalizing the concept of microdata sets. Microdata sets are structured in the form of a table of records/tuples and attributes. Within the context of this study, we assume that every record corresponds to an individual and every attribute corresponds to a (privacy-sensitive) property of the corresponding individual.

More specifically, a microdata set $DS_N$ comprises $N$ rows or records denoted by $x^n$, where $n : 1, \ldots, N$. We assume that every record $x^n$ corresponds to one individual. Further, every record $x^n$ comprises $D$ attributes, denoted by $a_d$, where $d : 1, \ldots, D$. Each attribute $a_d$ assumes a nominal or ordinal value from domain $A_d$ (or, in other words, attribute $a_d$ assume a value that is an element of set $A_d$). Domain $A = A_1 \times A_2 \times \ldots \times A_D$ denotes the super domain, over which all attributes are defined. Every record $x^n$ is defined over $A$, consisting of attribute values $x_d^n \in A_d$, $d : 1, \ldots, D$.

In attribute mapping, the set of attributes $\{a_1, a_2, \ldots, a_D\}$ are normally divided into four disjoint sets called: Explicit identifiers, quasi identifiers, sensitive attributes, and non-sensitive attributes. *Explicit Identifiers* (EIDs) refer to those attributes in $DS_N$ that structurally and on their own could uniquely identify individuals, i.e., data subjects. Examples of EIDs are a data subject's name and social security number. *Quasi Identifiers* (QIDs) refer to the set of attributes in $DS_N$ that could be used to identify (some of) the data subjects in $DS_N$. To this end, the QIDs should also be present in some other data sets or information sources together with the corresponding EIDs. The QIDs in microdata set $DS_N$, therefore, capture the background knowledge that intruders have with respect to data set $DS_N$. *Sensitive ATtributes* (SATs) refer to those attributes that capture privacy-sensitive information, while conveying useful information for a data analysis purpose (e.g., someone's disease type and salary). Unlike QIDs, SATs are known only within $DS_N$ and, therefore, they cannot be characterized as background knowledge for intruders. *Non-sensitive ATtributes* (NATs) refer to all the other attributes that are not EIDs, QIDs or SATs.

Through attribute mapping, attributes $a_1, a_2, \ldots, a_D$ of microdata set $DS_N$ are divided into 4 disjoint subsets EID, QID, SAT and NAT. Defining the EIDs is straightforward and is based on the intrinsic aspects of microdata set $DS_N$. Defining NATs becomes trivial once the other three subsets are determined. Defining QIDs and SATs is not straightforward as it depends on subjective and contextual aspects related to the data environment. QIDs capture the background information already known in the so-called *auxiliary information* sources (i.e., in the other data sets than $DS_N$) about the identities of (some of) the data subjects in $DS_N$. In other words, in the other data sets one can find a combination of attributes QIDs and one or more EIDs for (some of) the data subjects in $DS_N$. For illustration, the example in Figure 1 shows an attribute mapping for original microdata set $DS_N$, assuming the background information available to intruders as shown by the attributes of auxiliary data set $A_{aux}$. The last row in the figure indicates the attributes of the transformed microdata set $DS'_N$ due to applying SDC methods to microdata set $DS_N$.

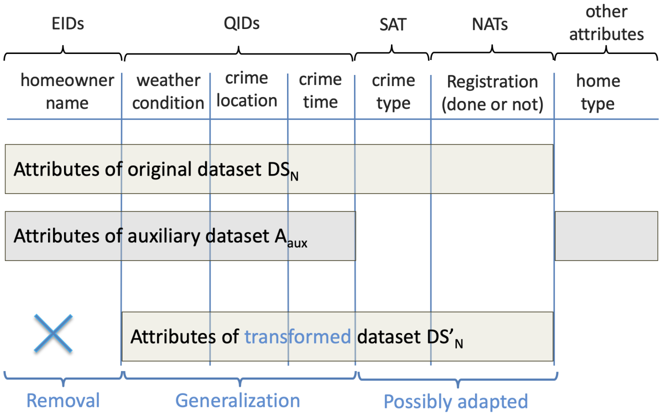


Figure 1. An illustration of attribute mapping for a data environment without the original microdata set.

Some legal frameworks specify specific attributes as SATs. For example, the UK's Data Protection Act (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* [9]. Further, the situational context and personal preferences (of data subjects) influence an attribute in being considered as a SAT. In some situations, the attributes related to one's income, wealth, credit record and financial deals can be considered as SATs. Attribute religion might be considered as a SAT in some countries and NAT in others.

In environments where the original microdata set $DS_N$ acts as an auxiliary information source (like that of the data controller, as also mentioned in Section III-B), all the other attributes in $DS_N$ with exception of the EIDs act as QIDs. In other words, QIDs, SATs and NATs shown in Figure 1 can act as the extended set of QIDs, given the original and the protected microdata sets, as shown in Figure 2. This extended set of QIDs can facilitate linking the records in the protected microdata set $DS'_N$ to the EIDs in the original microdata set $DS_N$.

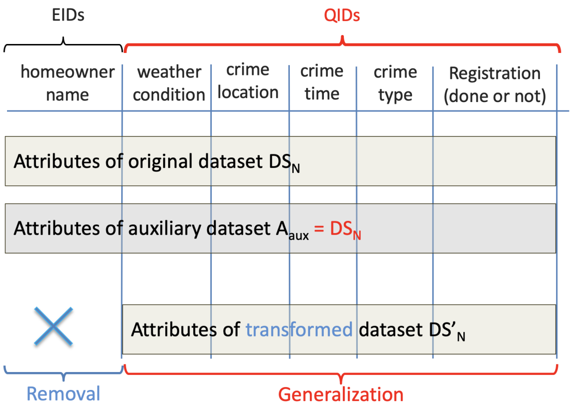

EIDs
QIDs
homeowner name
weather condition
crime location
crime time
crime type
Registration (done or not)
Attributes of original dataset DS_N
Attributes of auxiliary dataset A_aux = DS_N
Attributes of transformed dataset DS'_N
Removal
Generalization

Figure 2. An illustration of attribute mapping for a data environment with the original microdata set.

## IV. Guiding Principles for Personal Data Protection

### A. *Normative Approaches*

Often, legal regimes and definitions of privacy are based on the normative and intuitive assumptions "about how pieces of information interact" [12]. According to the normative notions of privacy with respect to data minimization (i.e., data protection via the de-identification and anonymization processes), a given microdata set $DS_N$ is considered as personal data if it can reveal personal information when it is combined with any other auxiliary data set $A_{aux}$ that is available to (legitimate and illegitimate) data recipients (i.e., the intruders). Data set $A_{aux}$ encompasses the background information available to intruders. We note that such auxiliary data sets are growing rapidly in the current Big Data era.

### B. *Towards Formal Approaches*

Dwork et al. [17] showed that it is impossible to enforce the stringent definition of privacy protection as proposed by the current normative definitions, when the intruder has an arbitrary amount of background knowledge. To protect privacy in those microdata sets that are used for statistical computation (and also for Open Data purposes), one should deal with the shortcomings of the heuristic data protection approaches that partially capture the normative notions of privacy. There is currently a trend to move from the current normative heuristics of privacy to the formal privacy protection approaches. For example, some legal scholars advocate to base legal privacy regimes, which are mostly based on the normative and intuitive assumptions about how pieces of information interact, on formal privacy models [12]. Formal privacy models, which are based on mathematically and rigorously proven techniques such as differential privacy [17], are inherently not subject to interpretation in different contexts, particularly in regard to other data sets. In other words, formal concepts do not rely on intuitive assumptions about how pieces of information interact, but rather on the properties of a data set itself which can be examined by scientific and mathematical principles.

A pioneering work that provides a formal definition of privacy is [17] that introduces the $\epsilon$–differential privacy technique. According to this definition, the presence or absence of the (personal) data of an individual in a data set must not have an observable impact on the output of an analysis/computation over that data set. In other words, it requires "the output distribution of a privacy preserving analysis to remain stable under any possible change to a single individual's information" [12]. The technique of $\epsilon$–differential privacy is already deployed in some Information Systems (ISs) currently by, for example, Google, Apple, Uber, and the U.S. Census Bureau. Apple uses the technique in iOS10 for increasing its security and privacy, Google uses it for protecting urban mobility data to ensure that individual users and journeys cannot be identified, and the U.S. Census Bureau wants to apply it to 2020 US census data for safeguarding the information it gathers from the US citizens [12].

One should note that the $\epsilon$–differential privacy technique guarantees privacy protection in the sense defined in the beginning of the previous paragraph. Whether this definition of privacy is comprehensive and adequate is not established. Although the formal approaches and definitions of privacy and privacy protection have not been introduced to legislation and regulations yet, there is a growing trend to do so in academia due to being independent of environmental conditions that are highly dynamic in the era of big and Open Data. Therefore, we shall examine the impact of such approaches and compare it with those of traditional normative approaches in Section VII for a specific formal technique (i.e., an $\epsilon$–differential method implemented in SDC tool ARX, see Subsection VI-E).

## V. Data Utility and Risk Measures

In our study, we define five data protection cases as possible scenarios applicable to Open Data settings and observe the data utility per each case. In this section, first, we describe the data utility and data disclosure measures used in the study. In Section VI, we present details of these cases.

To illustrate the notation adopted from this point on, Figure 3 summarizes these notations per each data transformation stage when applying SDC methods. From the original microdata set $DS_N$, the EIDs are removed or suppressed to obtain microdata set $DS'_N$. The QIDs of the result are generalized to get microdata set $DS''_N$. Finally, in order to achieve $k$-anonymity, some records of $DS''_N$ are suppressed to yield microdata set $DS''_{N'}$. Note that $N$ and $N'$ denote the number of the records in the corresponding data sets, where $N' \leq N$, i.e., the the number of records may decrease in last step in Figure 3.

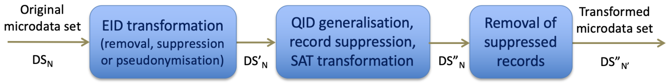

Original microdata set
DS_N
EID transformation (removal, suppression or pseudonymisation)
DS'_N
QID generalisation, record suppression, SAT transformation
DS''_N
Removal of suppressed records
Transformed microdata set
DS''_N'

Figure 3. The notation convention used from this point on.

### A. General-Purpose Data Utility Measures

Data utility measures are indicators for assessing the usefulness of the data transformations that are applied to the microdata by using SDC methods. There are two categories of data utility measures, the so-called special-purpose measures and general-purpose measures, which depend on whether or not the usage of data is already known, respectively [16]. In this paper, we consider the latter one, as the purpose of data usage is not determined beforehand in Open Data settings (see also the research objective part in Section I). In the following, we explain three general purpose data utility measures of Average Equivalence Class Size, Non-Uniform Entropy, and Granularity from literature in a way that they are provided/realized within ARX tool.

*1) Average Equivalence Class Size:* The Average Equivalence Class Size (AECS) measure [18] is given by:

$$\text{AECS} = \frac{|DS''_{N'}|}{k \times N_{EC}} = \frac{N'}{k \times N_{EC}},$$

where $N_{EC}$ is the number of the ECs of $DS''_{N'}$. The AECS $\geq 1$ and it is a measure of information loss. The higher the value of AECS, the higher is the amount of information loss. If the size of all ECs of microdata set $DS''_{N'}$ is $k$, then the AECS measure value is one, i.e., its minimum and best value for a given $k$. The optimization objective of data anonymization 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 suppression.

ARX provides the AECS as an information loss measure, which is a bit differently than the definition given above (namely, normalizing without the value of $k$ and with the value of $N$), as:

$$\text{AECS}_{\text{ARX}} = \frac{1}{N} \times \frac{N'}{N_{EC}}.$$

*2) Non-Uniform Entropy Measure:* Gionis and Tassa [19] 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 $DS_N$, given the distribution of values in the transformed microdata set $DS''_{N'}$. For example, let attribute $a$ be a QID, which takes values from set $\{v_1, v_2, \ldots, v_I\}$ in the original microdata set $DS_N$. In the transformed microdata set $DS''_{N'}$, the values of attribute $a$ are generalized to values $\{v_{1,2}, v_{3,4}, \ldots, v_{I-1,I}\}$, i.e., values $v_1$ and $v_2$ in $DS_N$ are generalized to value $v_{1,2}$ in microdata set $D''_{N'}$, and so on. Let $a''$ denote the attribute in $DS''_{N'}$ that corresponds to attribute $a$ in $DS_N$, noting that the values of $a''$ are from $\{v_{1,2}, v_{3,4}, \ldots, v_{I-1,I}\}$. Then, for example, given that $a = v_1$ and $a'' = v_{1,2}$, the information loss due to generalization for this outcome is proportional to

$$-\log_2 \frac{\#\text{ of } v_1}{\#\text{ of } v_{1,2}} = -\log_2 \frac{\#\text{ of } v_1}{\#\text{ of } v_1 + \#\text{ of } v_2}.$$

Let $a_m$, where $m : 1, \ldots, M$, denote a QID in microdata set $DS_N$ and $a''_m$ denote the corresponding QID in microdata set $DS''_{N'}$. We use $A_m = \{v_m\}$ and $A''_m = \{v''_m\}$ to denote the sets of values of QID $a_m$ and $a''_m$, respectively. The non-uniform entropy measure is defined in Relation (6) in [19] as (with slight adaptation, using our own notation defined above):

$$E = \sum_{m=1}^{M} \sum_{\substack{v_m \in A_m, \\ v''_m \in A''_m}} - \log_2 Pr(a_m = v_m | a''_m = v''_m)$$

This non-uniform entropy measure is monotonic with respect to generalization. Monotonicity of a measure here means that the measure increases monotonically with increasing degrees of generalization. In other words, if the value $v_1$ is generalized to value $v_{1,2}$ at level 1 and then to value $v_{1,4}$ at level 2, then the corresponding values of the measure increase when moving from level 1 to level 2. ARX provides a non-uniform entropy-based utility model, which is simply $1 - E$.

*3) Granularity:* Granularity is an information loss measure defined in [20], which is realized in ARX tool [21][8]. The measure expresses the degree to which the values of an attribute in the transformed microdata set cover the original domain of the attribute.

The information loss for every QID attribute $a_m$, which is generalized (or suppressed) from domain $A_m$ in original microdata set $DS_N$ to the domain $A''_m$ in the transformed microdata set $DS''_{N'}$, is computed as the average loss for every (i.e., per record) value of that attribute, denoted by $x^n_m \in A_m, n : 1, \ldots, N$. The loss for every value of the QID attribute is calculated based on the generalization taxonomy tree T of that attribute.

For a categorical QID $a_m$, let $M_m$ denote the total number of leaf nodes in the taxonomy tree $T_m$ of QID attribute $a_m$. Assume that the value of the QID attribute $a_m$ is generalized to node $P_m$, at which the sub-tree possesses $M_{p,m}$ leaf nodes. The loss of information when the value of the QID $a_m$ for the $n$-th record (i.e., value $x^n_m$) is generalized from a leaf node of tree $T_m$ in the original microdata set to the sub-tree node $P_m$ is calculated by:

$$\text{Information loss for attribute value } x^n_m = \frac{M_{p,m} - 1}{M_m - 1}.$$

When the value of the QID attribute $a_m$ is suppressed, the worst-case information loss occurs, i.e., the generalized node is the root of the taxonomy tree. This worst case leads to information loss $\frac{M_m - 1}{M_m - 1} = 1$ for that suppressed value of the QID attribute.

For numerical QIDs, the information loss can be defined similarly. Consider the value of such a numerical QID attribute $a_m$ for the $n$-th record (i.e., value $x^n_m$) is generalized to an interval $i$ defined by the lower and upper end points $L_{m,i}$ and $U_{m,i}$, respectively. Further, assume that the lower and upper bounds of that QID attribute $a_m$ in original data set are $L_m$ and $U_m$, respectively. Then, the information loss for this value of the QID attribute $a_m$ is given by:

$$\text{Information loss for } x^n_m = (U_{m,i} - L_{m,i})/(U_m - L_m).$$

Indeed, the granularity measure for every value of a QID attribute quantifies the loss when a leaf node value cannot be disambiguated from another leaf node value due to the generalization (i.e., when both belong to the same sub-tree of the generalization node $P$).

The information loss for the QID attribute $a_m$ is computed by averaging the loss for every (i..e., per record) value of that attribute.

$$\text{Information loss for } a_m = \frac{1}{N} \sum_{n=1}^{N} \text{Information loss for } x^n_m.$$

The granularity measure is the total information loss due to generalizations and suppressions for all QID attributes. It is computed by summing up the information loss of each QID as defined above (assuming that the QID attributes are equally important for identification potentially) and normalizing the outcome. Similar to non-uniform entropy, ARX supports an utility model based on granularity calculated as $1 - \frac{1}{M} \sum_1^M l(a_m)$, where $l(a_m) \in [0, 1]$ is the information loss for $a_m$.

### B. Data Disclosure Risk Measures

In the following, we explain three general purpose data disclosure risk measures of Prosecutor Record at Risk, Journalist Average Risk and Marketeer Success Rate. Note that these risk measures capture those risks associated with the external risks factors, as the background knowledge of intruders is modelled by the QIDs. Therefore, these data disclosure risk measures are applicable to Cases III, IV and V (to be mentioned in the following section). The risks associated with EIDs (i.e., the internal risks factors) and SATs (i.e., the risks associated with attribute linkage and table linkage attacks [16] are not captured by the risks measures studied in the following.

Typical measures for quantifying disclosure risks turn around the concepts of sample uniqueness and population uniqueness. With respect to the set of QIDs, let us assume that microdata set $DS_N$ is a sample of a larger population microdata set denoted by $P_L$ (i.e., $N \leq L$). Alternatively said, all data records in sample microdata set $DS_N$ are also in population microdata set $P_L$, where microdata sets $DS_N$ and $P_L$ have QIDs in common. Note that for re-identification of (some of) the records, it is necessary that data set $P_L$ includes the combination of attributes EIDs and QIDs. To this end, the EIDs can actually be present in $P_L$ or can potentially be present in $P_L$ in the sense that the intruder can somehow deduce the corresponding EIDs in the future via, for example, interrogation (e.g., asking neighbours), testing (e.g., testing someone's DNA), searching digital media (via search engines like Google and Bing), and so on. Population microdata set $P_L$ can be seen as background information, which does not contain attributes SATs and NATs of $DS_N$ (as, otherwise, these SATs and NATs should have been considered as QIDs).

Let us further assume that the QIDs in microdata sets $DS_N$ and $P_L$ are generalized in the same way, resulting in microdata sets $DS''_{N'}$ and $P''_{L'}$ with the same ECs (i.e., the same patterns of the values for the generalized QIDs). For a given EC, the EC size in $DS''_{N'}$ is smaller than or equal to the EC size in $P''_{L'}$. The uniqueness of a data record (i.e., an individual) in microdata sets $DS''_{N'}$ and $P''_{L'}$ with respect to QIDs can be defined as follows. Assume that the data record belongs to an EC, which has $|EC_S|$ and $|EC_P|$ records in microdata sets $DS''_{N'}$ and $P''_{L'}$, respectively. The sample uniqueness and population uniqueness of the record are defined by $|EC_S| = 1$ and $|EC_P| = 1$, respectively.

We note that population uniqueness results in sample uniqueness (i.e., if $|EC_P| = 1$, then $|EC_S| = 1$); and sample uniqueness does not necessarily result in population uniqueness (i.e., if $|EC_S| = 1$, then $|EC_P| \geq 1$). One should also note that while a data controller can easily validate sample uniqueness by investigating the (to be) released microdata set, the data controller cannot easily validate population uniqueness because population microdata sets are generally inaccessible to data controllers.

It is important to determine/know which of population uniqueness and sample uniqueness is more relevant for estimating 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 neighbor has [22][23]), 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 [22][23], then it is important to investigate population uniqueness. The rationale here is that the (likelihood of a) risk might be not high if a record, which appears alone in an EC in the sample microdata set, shares the same EC with multiple records in the population microdata set.

To further explain these points, let us reconsider a probability model from [24]. Assume a data record belongs to an EC, which has $|EC_S|$ and $|EC_P|$ records in microdata sets $DS''_{N'}$ and $P''_{L'}$, respectively. For a prosecutor attacker, the probability of correctly linking the individual with a record from the sample microdata set is

$$Pr(\text{correct linkage}|\text{being in } DS''_{N'}) = 1/|EC_S|.$$

In this case, sample uniqueness, captured by $EC_S$ above, is important. This measure corresponds to the measure of Prosecutor Record at Risk in ARX, where the maximum prosecutor risk is $\frac{1}{\text{the smallest } |EC_S|}$.

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

$$\begin{aligned} Pr(\text{correct linkage, being in } DS''_{N'}|\text{ being in } P''_{L'}) &= \\ Pr(\text{correct linkage } |\text{ being in } DS''_{N'}, \text{being in } P''_{L'}) &\times \\ Pr(\text{being in } DS''_{N'}|\text{ being in } P''_{L'}) &= \\ Pr(\text{correct linkage } |\text{ being in } DS''_{N'}) &\times \\ Pr(\text{being in } DS''_{N'}|\text{ being in } P''_{L'}) &= \\ \frac{1}{|EC_S|} \times \frac{|EC_S|}{|EC_P|} = \frac{1}{|EC_P|}. & \end{aligned}$$

This measure corresponds to the measure of Journalist Average Risk in ARX, which is the average of this vale for all ECs.

In both journalist and marketeer attacker cases, the population uniqueness, captured by $|EC_P|$ above, is important. Assuming that $|EC_S| \leq |EC_P|$, the worst-case scenario is the prosecutor attacker, i.e., the sample uniqueness. If we are sure that the attacker is unsure about the victim being in the sample data set, then population uniqueness is important.

## VI. Experiments

In this section, we describe 5 data protection cases that are applicable to Open Data settings. (Note that here we do not claim that these cases represent all possible cases.) From Case I to Case V, we tighten our assumptions on the background information that is available to intruders step-wise and observe the data utility and data disclosure behaviors, based on a number of measures defined in literature. Cases I-IV are based on the normative heuristics, while Case V is based on the formal $\epsilon$–differential privacy model as implemented in ARX.

### A. *Case I*

As the baseline, we consider a microdata with personal information, including identifying attributes. For our experiment, we choose the publicly available Adult data set. It is an excerpt of 32,561 records from the 1994 US census database. The data set is often used in similar studies, like [22][24][25]. As part of data preparation, we consider attribute "hhid" of the Adult data set as an EID, discard attribute "fnlwgt" as it does not convey much information for our purpose, and discard education level in numbers because it is another form of attribute education in categories. This baseline case is subject to personal data disclosures due to intrinsic aspects.

### B. *Case II: Basic Protection Against Intrinsic Risks*

In this case, the EID of the microdata set (e.g., the "hhid" in the case of the Adult data set) is removed, but the other attributes are unchanged. In the past, many practitioners used to characterize this case as anonymized data. Often, the set of explicit identifiers is removed (i.e., filtered), replaced with an unrecognizable value (i.e., masked/suppressed), or replaced with a unique and unrecognizable value (i.e., pseudonymized in a technical sense). Removal, suppression or pseudonymization of EIDs is considered as the first step of applying SDC methods. This first step eliminates the intrinsic risks of personal data disclosures in a microdata set, but is still vulnerable to personal data disclosures due to extrinsic factors [6].

### C. *Case III: Protection Against Data Linkage by Externs*

In this case, a microdata set is without EIDs, but with generalized QIDs (and/or suppressed records), and with untransformed SATs and NATs. The set of the QIDs chosen for Case III is {age, workclass, occupation, race, sex, native-country}, for which the $k$-anonymity is applied. In our experiments we do not modify SATs (by applying, for example, $l$-diversity or $t$-closeness) to contain the complexity of this presentation. Normally, the values of QIDs are transformed by, for example, generalization (e.g., exact ages are changed to age intervals), suppression (e.g., the gender attribute values are replaced with a specific character), or perturbation (e.g., random values are added to the body weight attribute values). Still, this case is subject to extrinsic risks, like record linkage by those who have access to the original microdata set (like data controllers).

A disadvantage of operating according to Case III is that, as the background information increases due to Big Data, the set of QIDs available to intruders expands. Bargh et al. [11] argue that data anonymity in the GDPR sense can be achieved if the data disclosure 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 anonymization. Data disclosure risks may increase over time due to availability of other data sets and changing environment conditions. Thus, the currently anonymous data may become personal data in the future. This implies that an applied privacy protection mechanism, which results in an anonymous data set currently, may not do so in the future.

### D. Case IV: Protection Against All Parties

In this case, a microdata set, which is without EIDs, is protected by considering all other attributes as QIDs. Thus, aligning with the previous case, the microdata set is protected by applying $k$-anonymity to all attributes that are considered as QIDs. To this end, the QIDs are generalized and some records are suppressed. There are no guarantees that data disclosure risks will not take place, as this case is still subject to some extrinsic risks [7][26].

### E. Formal Protection with $\epsilon$-Differential Privacy

In this case, the original microdata set $DS_N$ is stripped off from EIDs, and the result is protected by applying a method that conforms to $\epsilon$-differential privacy definition. For this method, there is a formal guarantee that data disclosure risks will not take place, provided that the definition of $\epsilon$-differential privacy is adopted. Note that this definition of privacy is other than that considered for Cases III and IV. Further, this Case V is subject to some extrinsic risks [7][26].

The tool ARX offers a method for applying $\epsilon$-differential privacy to microdata sets as proposed in [26]. According to this method, first the records of $DS_N$ are presampled and subsequently $k$-anonymity is applied to all remaining records, while considering all attributes as QIDs, to result in $\epsilon$-differential privacy, as described in [27]. The $k$-anonymity is applied to the sampled records by generalizing the values of QIDs and suppressing those records that belong to the ECs with less than $k$ records. To this end, the overall privacy budget $\epsilon$ is split up into two parts: (a) $\epsilon_{\text{anon}}$ (denoted by $\epsilon_a$ here), used by the anonymization operator, and (b) $\epsilon_{\text{search}}$ (denoted by $\epsilon_s$ here), used by the search strategy. The method proposed in [27], i.e., the SafePub method, satisfies ($\epsilon = \epsilon_a + \epsilon_s, \delta$)–differential privacy, where we should specify:

- $\delta$, which is recommended to be $\frac{1}{N} < \delta < 10^{-4}$ (where $N$ is the size of the input data set). We chose $\delta = \frac{1}{N}$ (based on the recommendation in Section 7.2 in [27]);
- Parameter Steps, which is the number of iterations performed by the search strategy. We chose Steps = 300 (based on the recommendation in Section 8.4 in [27]).
- In order to choose $\epsilon = \epsilon_a + \epsilon_s$ and, consequently $\epsilon_a$ and $\epsilon_s$, we chose $\epsilon_s = 0.1$ according to the recommendation in Sections 8.4, 8.5 and 8.6 in [27]. Note that the values of Steps and $\epsilon_s$ are related. Consequently, the value of $\epsilon_a$ can be varied between 0.1 and 2, as done in Fig 15 in [27]. In turn, $\epsilon = \epsilon_a + 0.1$ varies accordingly.

The relation of SafePub to $k$-anonymity is described in [26].

## VII. Results and Discussions

In order to run our experiments, we have used ARX's API and implemented a layer to specify and execute a series of experiments. We have used this layer to collect all the statistical measurement results for Cases I-V, see Section VI, over a range of parameters. In this section, the results of our experiments from Cases III-V are presented. The main privacy model in our experiments is $k$-anonymity. To derive the results based on the formal approach (i.e., Case V), our program explores a set of values for $\epsilon$ and $\delta$ and extracts the corresponding $k$'s. The set of extracted $k$'s is used to perform the experiments using $k$-anonymity model for the cases III-IV.

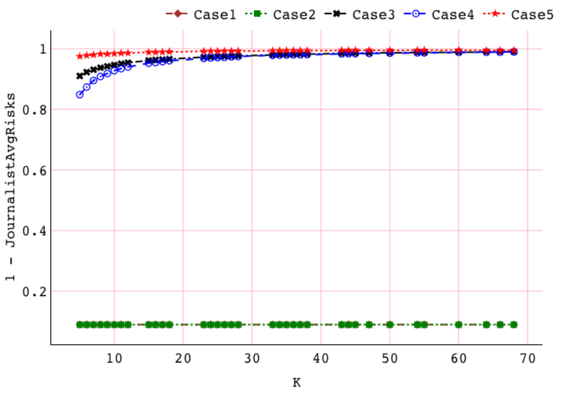


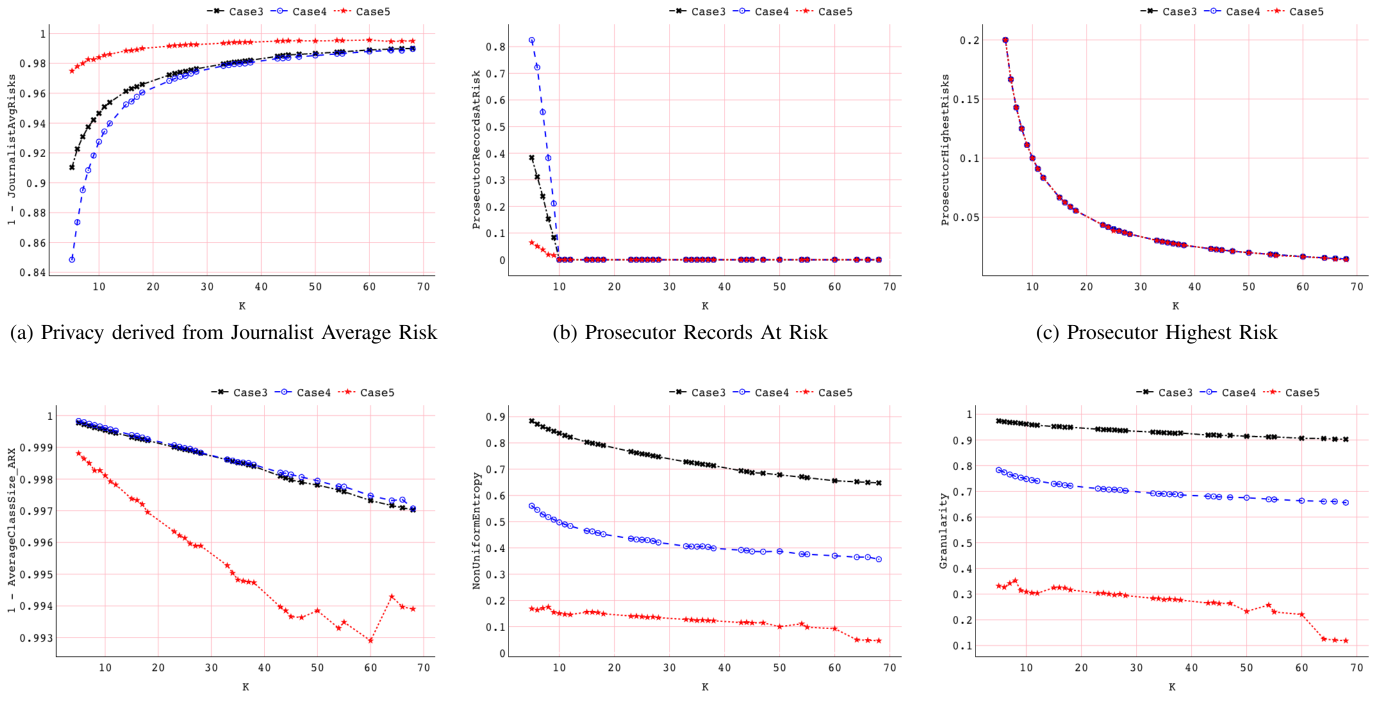
Figure 4. All cases Privacy derived from Journalist Average Risks

The results are visualized in two categories: (1) The behaviour of the risk measures and the utility measures per different values of $k$ (see Figure 5), and (2) The behaviour of data utility versus privacy (see Figure 6). For each category there is a set of graphs that represent the results of our anonymization experiments. All our experiments are performed with the maximum suppression limit and minimum generalization factor as provided in ARX [8]. To present the results in Figure 5 and Figure 6, Cases I and II are omitted as these two cases either had the maximum or the minimum values. Having visualized these two cases would have suppressed the behaviors of Cases III-V. As an example, see Figure 4 where cases I and II show the minimum privacy.

In Figure 5a, the privacy measure of output data is presented. Increasing the value of $k$ yields higher privacy (in this case, lower journalist average risks) which is validated also in Figure 5b and Figure 5c; i.e., the number of records at risk and the highest risk are decreasing. As expected, comparing case III with the other cases, both Figure 5a and Figure 5b justify that even with lower $k$'s, $\epsilon$-differential privacy (Case V) outperforms by a big margin. However, the higher privacy performance in Case V has a drawback of losing quality as depicted in Figure 5d and reaffirmed by Figures 5e and 5f.

The second category of the results, as shown in Figure 6, presents the combined behaviour of a data utility/quality measure versus a data privacy measure in three graphs. As we see from these graphs and as expected, when privacy increases the utility of the data decreases. Interestingly, we can see that Case V operates near the best privacy area, with the lowest data quality. This has to do with the presampling inherent to the $\epsilon$-differential privacy method used, which results in lower disclosure risks and lower data utility values. Comparing the performances of Cases III and IV, we observe that the former behaves closer to the higher values of the data risk and data utility measures. This has to do with having a fewer number of attributes acting as QIDs and being generalized.

For Open Data purposes and from the perspective of protecting privacy, Case V operates better than Case IV, and Case IV operates better than Case III. This performance comes with the cost of having lower data utility, respectively. One should also note that the privacy property that is realized in Cases III and IV differ from that realized in Case V. While the former is based on a normative property (i.e., to prevent record linkage when the transformed microdata is linked with



(a) Privacy derived from Journalist Average Risk

(b) Prosecutor Records At Risk

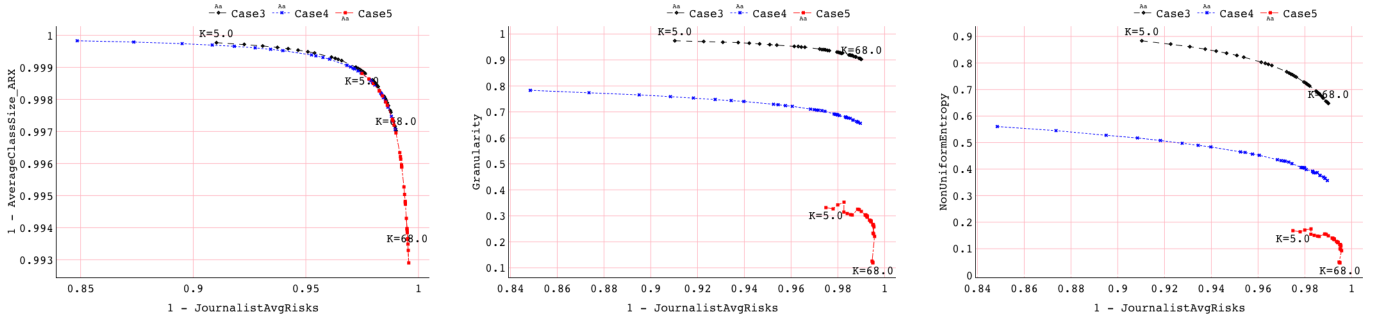
(c) Prosecutor Highest Risk

(d) Quality derived from Average Class Size

(e) Non Uniform Entropy

(f) Granularity

Figure 5. Risks and Utility Measurements for $K \in [5, 68]$



(a) Quality: derived from AverageClassSize per Privacy: derived from JournalistAvgRisks

(b) Quality: Granularity per Privacy: derived from JournalistAvgRisks

(c) NonUniformEntroy per Privacy: derived from JournalistAvgRisk

Figure 6. Quality per Privacy Measurements

other microdata sets), the latter is based on a formal definition of privacy (i.e., the definition of $\epsilon$-differential privacy that guarantees the presence or absence of the (personal) data of an individual in a data set does not have an observable impact on the output of an analysis/computation over that data set). Whether this definition of privacy is comprehensive and adequate, specially in Open Data settings, is not established scientifically and/or adopted within privacy regimes and data protection regulations.

SDC-based anonymization does not provide a full guarantee against personal data disclosures, nevertheless, applying it is necessary for realizing compliance to the principle that personal data should be processed fairly (see Article 5(1-a) of GDPR). This fairness asks for putting sufficient efforts to protect personal data in a given context. Therefore, we argue that choosing all attributes as QIDs in Case IV or some attributes as QIDs in Case III is the least amount of data anonymization efforts needed to protect personal data according to the normative property of privacy discussed above. But the question remains if this will be sufficient, and will be seen as such, for instance, in the light of the GDPR fairness principle. In the environments where the formal definition of $\epsilon$-differential privacy prevails, then applying the formal model as in Case V can (or even must) be considered. Note that in this work we kept our presentation simple and did not apply complementary data protection methods like $l$-diversity and $t$-closeness. In a practical setting, one should consider applying these techniques to contain disclosure risks at an acceptable level based on, among others, the principle of fair processing of data.

## VIII. Conclusion

In this contribution, we analyzed the consequences of two cases in Open Data settings, namely having access to and having no access to original microdata sets, in terms of data utility. To this end, we applied SDC technologies in a number of steps to minimize privacy risks while maintaining data utility. Opting for Case III, where the opened data might potentially be re-identifiable for parties with the original microdata sets (like the data controller), can yield higher data quality than that in Case IV where the microdata is protected against such parties. On the other hand, opting for Case IV is an attempt to make the transformed microdata anonymous for everybody (e.g., the data controller). For Open Data purposes and from the perspective of protecting privacy, Case V operates better than the other cases. This performance of Case V comes with the cost of having lower data utility relatively to the other cases. We noted that the formal definition of privacy behind Case V is not established widely within privacy regimes and data protection regulations.

In this study, we clarified the difference among a number of solution directions for protecting personal data when publishing microdata sets to the public according to their implications on disclosure risks and utility of data. Based on the results of this contribution, we believe, legal experts, legislators and policymakers can make an informed choice among these options or foresee new solution directions based the here adopted approach.

In the future, we intend to apply the experiments to more real-world data sets. Also, we will explore how to embed the desires of data consumers and data publishers who would like to publish data effectively, while preserving the privacy of data subjects as much as possible. Further, we aim at extending the tool in the direction of having improved data utility and being user friendly for data controllers.

## References

[1] M. Bargh and S. Choenni, "On preserving privacy whilst integrating data in connected information systems." Academic Conferences and Publishing International, Proc. of the 1st Int. Conf. on Cloud Security Management (ICCSM), October 17-18, Seattle, USA, 2013.

[2] S. Kalidien, S. Choenni, and R. Meijer, "Crime statistics online: potentials and challenges," in *Proceedings of the 11th Annual International Conference on Digital Government Research, Public Administration Online: Challenges and Opportunities, Puebla, Mexico*, May 17-20, 2010, pp. 131–137.

[3] J. Prins, D. Broeders, and H. Griffioen, "iGovernment: A new perspective on the future of government digitisation," *Computer Law Security Review*, vol. 28, no. 3, pp. 273 – 282, 2012.

[4] S. Choenni, M. S. Bargh, C. Roepan, and R. F. Meijer, *Privacy and Security in Smart Data Collection by Citizens*. Cham: Springer International Publishing, 2016, pp. 349–366.

[5] S. W. van den Braak, S. Choenni, R. Meijer, and A. Zuiderwijk, "Trusted third parties for secure and privacy-preserving data integration and sharing in the public sector," in *13th Annual International Conference on Digital Government Research, DG.O, College Park, MD, USA*, June 4-7, 2012, pp. 135–144.

[6] L. Sweeney, "Maintaining patient confidentiality when sharing medical data requires a symbiotic relationship between technology and policy." Artificial Intelligence Laboratory, Massachusetts Institute of Technology, AIWP-WP344, 1997.

[7] N. Li, W. Qardaji, and D. Su, "Provably private data anonymization: Or, k-anonymity meets differential privacy," *CoRR*, vol. abs/1101.2604, 01 2011.

[8] Arx data anonymization tool. https://arx.deidentifier.org/, Date visited: Feb. 2020.

[9] K. O. M. Elliot, F. Mackey and C. Tudor. The anonymisation decision-making framework, technical report by uk anonymisation network (ukan). https://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf, Date visited: Feb. 2020.

[10] K. E. Emam and B. Malin. Concepts and methods for de-identifying clinical trial data. https://www.ncbi.nlm.nih.gov/books/NBK285994/, Date visited: Feb. 2020.

[11] M. S. Bargh *et al.*, "Opening privacy sensitive microdata sets in light of GDPR," in *20th Annual International Conference on Digital Government Research, DG.O, Dubai, United Arab Emirates*, June 18-20, 2019, pp. 314–323.

[12] K. Nissim and A. Wood, "Is privacy privacy?" *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2128, p. 20170358, 2018.

[13] A. Wood *et al.*, "Differential privacy: A primer for a non-technical audience," vol. 21, no. 17. Vanderbilt Journal of Entertainment and Technology Law, 2018, p. 209.

[14] K. Nissim *et al.*, "Bridging the gap between computer science and legal approaches to privacy," vol. 31, no. 2. Harvard Journal of Law Technology, 2018, pp. 687–780.

[15] M. S. Bargh, R. Meijer, and M. Vink, "On statistical disclosure control technologies: For enabling personal data protection in open data settings," Research and Documentation Center WODC, The Hague, The Netherlands, Tech. Rep., 2018.

[16] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," vol. 42, no. 4, p. 14, 2010.

[17] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography, Third Theory of Cryptography Conference, TCC, New York, NY, USA*, March 4-7, 2006, pp. 265–284.

[18] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proceedings of the 22nd International Conference on Data Engineering, ICDE, Atlanta, GA, USA*, April 3-8, 2006, p. 25.

[19] A. Gionis and T. Tassa, "k-anonymization with minimal loss of information," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 2, pp. 206–219, 2009.

[20] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada*, July 23-26, 2002, pp. 279–288.

[21] F. Prasser, J. Eicher, R. Bild, H. Spengler, and K. A. Kuhn, "A tool for optimizing de-identified health data for use in statistical classification," in *IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, June 2017, pp. 169–174.

[22] F. Prasser, F. Kohlmayer, and K. A. Kuhn, "A benchmark of globally-optimal anonymization methods for biomedical data," in *IEEE 27th International Symposium on Computer-Based Medical Systems, New York, NY, USA*, May 27-29, 2014, pp. 66–71.

[23] K. El Emam, F. K. Dankar, A. Neisa, and E. Jonker, "Evaluating the risk of patient re-identification from adverse drug event reports," *BMC medical informatics and decision making*, vol. 13, no. 1, p. 114, 2013.

[24] F. Prasser, F. Kohlmayer, and K. A. Kuhn, "The importance of context: Risk-based de-identification of biomedical data," *Methods of Information in Medicines*, vol. 55, no. 4, pp. 347–355, 2016.

[25] F. K. Dankar, K. El Emam, A. Neisa, and T. Roffey, "Estimating the re-identification risk of clinical data sets," *BMC medical informatics and decision making*, vol. 12, no. 1, p. 66, 2012.

[26] N. Li, W. H. Qardaji, and D. Su, "On sampling, anonymization, and differential privacy or, *k*-anonymization meets differential privacy," in *7th ACM Symposium on Information, Computer and Communications Security, ASIACCS, Seoul, Korea,*, May 2-4, 2012, pp. 32–33.

[27] R. Bild, K. A. Kuhn, and F. Prasser, "Safepub: A truthful data anonymization algorithm with strong privacy guarantees," *PoPETs*, no. 1, pp. 67–87, 2018.