A Review of Frequency Table Disclosure Control from a Microdata Perspective

Alexander Latenko[1], Mortaza S. Bargh[2], Susan van den Braak[3], Marco Vink[4]
[1-4] Research and Documentation Centre, Ministry of Justice and Security, The Hague, The Netherlands
[2] Rotterdam University of Applied science, Research Center Creating 010, Rotterdam, The Netherlands
Email: [1] a.latenko@wodc.nl [2] m.shoae.bargh@wodc.nl [3] s.w.van.den.braak@wodc.nl [4] m.e.vink@wodc.nl

***Abstract*—Protecting personal data is a key requirement for properly sharing and opening data. With growing concerns regarding privacy, it is important to ensure that the personal data of individuals is not compromised or made public in Open Data initiatives. For the most part, the personal data protection fields for microdata and tabular data have been researched separately. This separation has caused both fields to have much overlapping research, particularly concerning the privacy and utility of the respective data types. This overlapping research, however, has not been well integrated between the fields. Recently, there have been developments and improvements for protecting microdata that are not being applied to the field of tabular data protection. In this work, the association between microdata and tabular data is formalized and used to link the personal data disclosure risks and the personal data protection models that can be applied to both microdata and tabular data.**

Keywords–Data Protection; Disclosure Scenarios; Frequency Tables; Statistical Disclosure Control.

I. INTRODUCTION

Within the process of opening and sharing data, Statistical Disclosure Control (SDC) is applied to reduce the risk of privacy disclosures for individuals while preserving the quality and utility of the data. Minimizing the risk of privacy disclosures is an essential step that needs to be performed in order to adhere to privacy regulations, such as the EU's General Data Protection Regulation (GDPR). As essential as it is for a data controller, i.e., the entity that opens the data, to provide sufficient guarantees of privacy, it is perhaps just as essential for a data user to be provided with similar guarantees of the quality of data. There are different reasons for opening or disseminating data, including, among others, improving transparency and enabling (scientific) research. Census tables are an example of opening data for transparency, where the information in those tables influences public perception and therefore should be as informative as possible. Opening data does not only facilitate research, but has become increasingly necessary for academic work to be acceptable for publication in certain journals [1].

SDC solutions are non-trivial in practical settings as the identification of potential sources of disclosure is a difficult task. This becomes clear from the recent cases where data subjects, the individuals present in the data, were first de-identified (anonymized), but were later re-identified by researchers [2]. Even when SDC methods have been applied on data, re-identification is still sometimes possible. Numerous cases have been discovered, including the infamous cases of disclosure in the microdata of taxi rides from NYC [3] and tabular data containing sensitive health information [4].

To prevent re-identification, an initial identification of the sources and causes of personal data disclosures is required. As such, this work contributes by providing a taxonomy for data disclosures when opening tabular data. Models such as t-closeness [5] and differential privacy [6] have been introduced to provide certain levels of privacy. Such models have mainly been introduced for protecting microdata. However, tabular data and microdata are closely related. We fundamentally formalize the relation between the two data types. This formalization makes it possible to evaluate the relation between SDC models developed for protecting microdata sets and those developed for protecting tabular data sets. Thus, this work improves the unification of the SDC methods and models developed for microdata and tabular data sets. The contribution on this work is focused on frequency tables, which is the most general type of tabular data. The disclosure risks and privacy models for frequency tables mainly hold for other types of tabular data, such as magnitude tables [7]. However, the disclosure risks that affect other specific types of tabular data are not considered in this work.

To the best of our knowledge, this is the first work that aims at unifying the privacy models for microdata and those for tabular data, allowing for comparisons between the privacy models. The rest of this work consists of the formalization of microdata and tabular data, specifically frequency tables, in Section II. The concept of disclosure is introduced in Section III, followed by the attacks that cause personal data disclosures in Section IV. An overview of privacy models is presented in Section V. Lastly, Section VI concludes this work and discusses possible future work.

II. DATA ASSOCIATION

In order to unambiguously describe the scenarios where personal data disclosures may take place for tabular data sets, the concept of microdata and tabular data are formalized in this section.

A. Microdata

A microdata set $\mathcal{DS}_M$ comprises N rows, or records, denoted by x^n, where $n = 1, \ldots, N$ and every record x^n corresponds to one individual. Further, every record x^n comprises D attributes. An attribute is denoted by a_i, where $i : 1, \ldots, D$. An attribute a_i has an associated domain of nominal or ordinal values A_i. Domain $A = A_1 \times A_2 \times \ldots \times A_D$ denotes the super domain, which contains all attribute values in $\mathcal{DS}_M$. Every record x^n is defined over A, consisting of attribute values $x_1^n, x_2^n, \ldots, x_D^n$, where $x_i^n \in A_i$, $i : 1, \ldots, D$. Table I is an example of a microdata table.

In the SDC literature for microdata, the set of attributes $\{a_1, a_2, \ldots, a_D\}$ are generally divided into four disjoint sets called: explicit identifiers, quasi identifiers, sensitive attributes, and non-sensitive attributes. *Explicit Identifiers* (EIDs) refer

 ISBN: 978-1-61208-760-3

TABLE I. EXAMPLE OF MICRODATA

	EID	QIDs		SAT	Misc.	
Record	a_1: Name	a_2: Zip	a_3: Age	a_4: Illness	...	a_D
x^1	x_1^1=Jane Doe	x_2^1=2230	x_3^1=15	x_4^1=Cancer	...	x_D^1
.	.	.	.	.		.
.	.	.	.	.	...	.
x^N	x_1^N=...	x_2^N=...	x_3^N=...	x_4^N=...	...	x_D^N

to the set of attributes in the original microdata set $\mathcal{DS}_M$ that structurally and on their own could uniquely identify an individual. Examples of explicit identifiers are an individual's name, home address and unique personal numbers like a 'social security number', 'national health service number', 'voter card identification number', or 'permanent account number'.

Quasi Identifiers (QIDs) refer to the set of attributes in the original microdata set $\mathcal{DS}_M$ that could 'potentially' identify individuals. Identification through QIDs is achieved by using a combination of values, which belong to QID attributes, of a record from the microdata set $\mathcal{DS}_M$. For these values of QID attributes, an intruder, a person or entity that seeks to learn personal information about data subjects, can identify individuals from other known knowledge bases. Knowledge bases can be very specific and personal, such as being acquainted with the data subjects, or general, such as the information from other public data releases. For example, assume that weight, length, hair color, and location are QIDs. Knowing the values of these attributes, an acquaintance may recognize the person uniquely. The QIDs in microdata set $\mathcal{DS}_M$, therefore, capture the so-called background knowledge that intruders have with respect to microdata set $\mathcal{DS}_M$.

Sensitive Attributes (SATs) refer to those attributes that capture privacy-sensitive information about individuals. In the justice domain, for example, this could be the specifics of a crime committed or the remaining duration of a prison sentence, and in the health domain this could be the condition an individual is suffering from. These sensitive attributes are sometimes important for data users for data analytics purposes. Unlike QIDs, SATs are assumed to be unknown outside of the original microdata set $\mathcal{DS}_M$ and, therefore, they are not characterized as background knowledge of intruders.

Non-sensitive Attributes (NATs) refer to all the miscellaneous attributes that are not directly-identifying, quasi-identifying or sensitive in a specific context. For example, someone's favorite color may be considered as a NAT in a microdata set for medical research. We shall use the concepts of EID, QID, SAT and NAT to explain statistical disclosures and the SDC methods for frequency tabular data sets.

B. Frequency Tables

A frequency table, also known as a contingency table, is constructed from a subset of attributes generally referred to as *grouping attributes*. The set of grouping attributes γ, which consists of d attributes, is generally only a small subset of the attributes of the original microdata. As such, the dimension of a frequency table, denoted by d, tends to be (much) smaller than that of the microdata, denoted by D.

A table consists of a number of cells. Every cell C_y, in a tabular data set $\mathcal{DS}_T$, contains an attribute value pattern y, which consists of a combination of grouping attribute values, i.e., $y = \{y_i | y_i \subset A_i, a_i \in \gamma\}$. The set notation has been used here because y_i can represent a single attribute value, for instance, a single age y_i=15 if a_i=Age, or y_i can also be used to represent a set of ages y_i=$\{15, 16\}$. In frequency tables a cell C_y expresses the number of records x^n in the source microdata set $\mathcal{DS}_M$ whose relevant attribute values $x_1^n, x_2^n, \ldots, x_d^n$ fit the attribute value pattern y. Since the records in $\mathcal{DS}_M$ generally have a larger dimension, only a subset of attribute values are counted. For example, in Table I attributes (x_3^n, x_4^n) are counted for all n records when they fit the pattern of (y_3, y_4). Table II is an example of a frequency table with attributes (y_3, y_4) that could have been sourced from the microdata of Table I.

TABLE II. EXAMPLE OF A 2-DIMENSIONAL TABLE CONSTRUCTED FROM TWO ATTRIBUTES IN TABLE I

		a_4(SAT)		m_3=Total
		y_4=Cancer	y_4=HIV	
a_3(QID)	y_3=15	1	2	3
	y_3=16	4	5	9
m_4=Total		5	7	12

In a frequency table, the cell value can be defined as:

$$C_y = |\{x^n | x^n \in \mathcal{DS}_M, \forall y_i \in y : x_i^n \in y_i\}|. \quad (1)$$

In addition to the cells defined in (1), frequency tables also contain total cells, denoted by m_i, that are only comprised of a single attribute value. We refer such cells are *marginals*. Table marginals in tabular data sets are obtained from the projection of the j-ary cube $a_1, a_2, \ldots, a_j$ onto a subset of j attributes, also called j-way marginals [8], where $j < d$. For example, one-way marginal with respect to grouping attribute a_i with value y_i is:

$$m_i(y_i) = \sum_{C_y \in \mathcal{DS}_\mathcal{T}, \text{ given } y_i \text{of } y} C_y.$$

The summation above is over all cells with value y_i. The two-way marginal with respect to attributes a_i and a_j with value y_i and y_j is:

$$m_{i,j}(y_i, y_j) = \sum_{C_y \in \mathcal{DS}_\mathcal{T}, \text{ given } y_i, y_j \text{of } y} C_y. \quad (2)$$

These summations can be used to define marginals for up to d-way marginals.

C. Release

Frequency tables are derived from microdata. Similar to a microdata set, a frequency table provides information about a number of records. A frequency table provides this information about d grouping attributes, while the microdata set $\mathcal{DS}_M$ provides this information for all attributes. When frequency tables have the same dimension as their microdata sources, the data sources contain the same information, albeit being differently structured, which is due to the difference between the definitions of y_i and x_i^n. Compared to frequency tables, microdata generally provides more detailed information, which introduces a higher risk of personal data disclosures.

Tabular data has been used to provide information on a subset of attributes, i.e., the grouping attributes, with those attribute values that are assumed to be interesting for data users. These attributes are selected to be in the y of the table cells. A common example is census data, where the objective is to provide information about the population that is as accurate

 ISBN: 978-1-61208-760-3

as possible and is processed minimally. When the objective is to release information for the sake of transparency, tabular data is a common choice.

III. DISCLOSURE CONCEPTS

In this section, we formalize several concepts that are relevant for characterizing personal data disclosures within tabular data sets.

A. Disclosure Elements in Tabular Data Sets

The elements of a tabular data set, or a table in short, that can be used for disclosing someone's personal information are:

1) Grouping attributes $a_1, a_2, \ldots, a_d$, which represent the dimensions of the table.
2) The table description attribute(s) t_{des}, which results from the table caption or the textual explanations embedded in the paragraphs preceding or succeeding the table.

For example, in Table III, there are two attributes a_1 (gender, being male or female) and a_2 (age, being minor or adult). The table also includes a table description in the table caption, saying that the table is about those arrested (acting as $t_{\text{des},1}$) in January 2019 (acting as $t_{\text{des},2}$). Note that, although we denote it as t_{des} for emphasis, a table description attribute is also an attribute from the microdata source.

TABLE III. NUMBER OF ARRESTS IN JANUARY 2019 (ACTING AS $t_{\text{des},1}$ and $t_{\text{des},2}$, RESPECTIVELY).

a_1 : Gender		male	female	Total
a_2 : Age	minor	11	1	12
	adult	62	37	99
Total		73	38	111

B. Re-identification

In SDC methods for protecting microdata sets, the background knowledge of intruders (so-called intruder's prior), which can be used to re-identify data subjects, is mainly modelled in QIDs. In this section, we outline how this approach can be extended to tabular data sets. To explain the attribute mapping for tabular data sets, we use Table III as an example of a frequency table. The cell with value 1 in the table corresponds to one data record in the microdata set from which the table is constructed. This cell can be specified by grouping attribute values $a_1 =$ female and $a_2 =$ minor, and the table description attribute values $t_{\text{des},1} =$ those arrested and $t_{\text{des},2} =$ in January 2019.

The re-identification of the record, corresponding to a cell with value 1, may take place based on any combination of attributes $a_1, a_2, t_{\text{des},1}$ and $t_{\text{des},2}$, which may potentially act as QIDs. For example, in our data set we have a single female minor who has been arrested in January 2019. That means that any intruder who knows someone that fits those QID values uniquely, can identify the person that is counted in Table III. This will be referred to as the *re-identification* of a cell with value 1. Once an individual has been identified to be uniquely part of a cell, any new information or data related to that cell will thus also be attributed to that individual, as described in the following subsection.

Note that for exact re-identification of an individual, corresponding to a cell with value 1 in a frequency table, it is important that just one person fits in the group specified by the values of the grouping, and description attributes, that act as QIDs. This category is also called the *Equivalence Class* (EC) of those QIDs. This uniqueness of the individual corresponding to the cell with value 1, and within the EC of the corresponding QIDs, can be described based on the concepts of sample uniqueness and population uniqueness [9].

Both sample uniqueness and population uniqueness should be defined based on the values of those attributes that act as QIDs. The underlying assumption is that every cell with value 1 (and corresponding microdata record) in the frequency table can potentially be identified based on the QIDs. The QIDs being the background information available to an intruder to link a cell to an actual individual.

Let $|S_{EC}|$ denote the cell value in a frequency table, i.e., the number of data records of the corresponding microdata set. The EC being determined over the values of QIDs of the frequency table. The value of $|S_{EC}|$ determines the degree of uniqueness of the cell (or of the corresponding records/individuals in the microdata set). If $|S_{EC}| = 1$, then the corresponding cell is unique in the published frequency table. A larger value of $|S_{EC}|$ makes the corresponding cells (or records) less unique.

Sample uniqueness is necessary, but it is not enough for re-identification. With respect to the set of the QIDs, we assume that the frequency table (or the corresponding microdata set) is a sample of a larger population microdata set. In other words, all data records in the sample data set (i.e., the frequency table of the corresponding microdata set) are also in the population microdata set denoted by P, where these sample and population microdata sets have been defined over the same attribute value domains. Therefore, both have the same ECs (i.e., the same patterns of the values for the QIDs). Let $|P_{EC}|$ denote the size of the EC, which is determined over the values of the QIDs of the frequency table. The uniqueness of an individual/record in both microdata sets can be defined by $|S_{EC}| = 1$ and $|P_{EC}| = 1$, which are the sizes of the EC in those microdata sets. We note that:

- Population uniqueness results in sample uniqueness (i.e., if $|P_{EC}| = 1$, then $|S_{EC}| = 1$); and
- Sample uniqueness does not necessarily result in population uniqueness (i.e., if $|S_{EC}| = 1$, then $|P_{EC}| \geq 1$).

In practice, given a sufficient number of QIDs, an intruder can fairly accurately estimate the probability that $|S_{EC}| = 1$ results in $|P_{EC}| = 1$. A recent work has shown that the estimations were possible with more than 95% accuracy when there are 15 QIDs [10].

One should also note that while a data controller can easily validate sample uniqueness by investigating the released frequency table (or the corresponding microdata set), this is not always possible for the population uniqueness. The data controller does not necessarily possess the entire population data.

Note that *population uniqueness is necessary, but it is not enough for exact re-identification*. In addition to population uniqueness, whereby the size of an EC in the population data set is 1 (i.e., $|P_{EC}| = 1$), there should be a unique identifier (e.g., an EID) associated with the corresponding EC from the population microdata set, so that the identities can be linked to the cell with a value 1 in the frequency table.

 ISBN: 978-1-61208-760-3

C. Attribution

Some grouping attributes can act as QIDs, as mentioned in the previous subsections, and other attributes can act as SATs. An attribute is a SAT when it contains information that could potentially be harmful for the associated individual or groups when released [11]. The value of a SAT does not contribute to the identification of an individual the way that a QID does, as the SAT is specific to (i.e., known within) the data set to be released. In combination with external data sets, the grouping and table description attributes acting as QIDs could be used to identify individuals and, consequently, reveal the values of the other grouping attributes for those records/individuals.

Table IV provides 4 example cases of the grouping and table description attributes that could act as QIDs and SATs, based on the example given in Table III. Note that in Table IV, attributes $t_{\text{des},1}$ and $t_{\text{des},2}$ are merged into attribute t_{des} for simplifying the presentation. For each case in Table IV, we assume that the corresponding QIDs result in identification for the cell with value 1 in Table III. This is because there is only a single person described by the QIDs, in those cases, in the sample data sets (i.e., sample uniqueness), it is assumed that the population uniqueness holds as well. Note that this is an illustrative example, in practice, often more QIDs are needed to result in population uniqueness and someone's identification with a high certainty. Furthermore, for the sake of simplicity, we assume that all other attributes per case in Table III are SATs.

TABLE IV. FOUR CASES ILLUSTRATING ATTRIBUTION TO SATs, THROUGH IDENTIFICATION BY QIDs IN TABLE III)

	QIDs (already known facts about the cell contributor in the world)	SATs (new facts known about the cell contributor via Table III)
Case 1	a_1, a_2	t_{des}
Case 2	a_1, t_{des}	a_2
Case 3	a_2, t_{des}	a_1
Case 4	a_1, a_2, t_{des}	. . .?

Disclosure through re-identification on its own may not be an issue, if one guarantees that no more information about the identified individual can been learned. If we examine the SATs column in Table IV, we find that for cases 1-3 we do learn a new attribute value from the released table that describes something extra about the person in the data. Thus, the combination of unique identification and learning a new attribute value about the identified individual can lead to so-called *individual attribution*. Note that the intruder in case 4 may not learn anything new about the data subject, but the case could still be perceived as privacy intrusive because this table aggregates, presents and reaffirms the associations of all grouping attributes and the table description attribute (i.e., a_1, a_2, t_{des}) to the individual.

TABLE V. AN ILLUSTRATION OF GROUP ATTRIBUTION.

Number of arrests ($t_{\text{des},1}$) of minors (a_2) in 2019 ($t_{\text{des},2}$)				
a_1: Gender		Male	female	total
a_3: Crime	hacking	5	5	10
	DUI	15	0	15
Total		20	5	25

Attribution can also occur without identification, for example, intruders can learn something new about a whole group without identifying the individual groups members. This is called *group attribution*. Consider the example in Table V, which is a representation with a more specific set of attributes from Table III. In Table V the grouping attribute a_2 assumes only the 'minor' value, furthermore, the grouping attribute a_3 is included, which only specifies the crime types: hacking and Driving Under Influence (DUI). In Table V, we cannot uniquely identify the records corresponding to the cell with value 5 by knowing the values of just the QIDs a_1 and a_2 because they correspond to 5 records/individuals in this table. Nevertheless, the intruder can learn that the crime type is hacking (i.e., $a_3 =$ hacking), for someone whose QIDs match $(a_1, a_2) =$ (female, minor), without being able to re-identify the exact person from the released table.

IV. Sources of Disclosure

Intruders seek to learn private information about the contributors present in a released table. There are many aspects that matter when trying to assess the risk of disclosure, such as the motivation, means and consequences of a disclosure attempt, which are part of what is called a disclosure scenario [12]. These aspects vary between releases as much as the information within the releases does. Most intruders that are interested in medical data might have little reason to actively seek disclosure risks in tax data, and vice versa. Both groups, however, will use similar attacks to learn about contributors.

Let us assume that the intruder, as background knowledge, has data set D_B, where every record z^n is defined over d'' attributes, some of which are defined from the same attribute domains of the original microdata $\mathcal{DS}_M$. In other words, microdata sets D_B and $\mathcal{DS}_M$ have some attributes in common (only QIDs).

A. Few contributors

When there are few contributors in a cell of a frequency table, the risk of disclosure is fairly high. When there is a single contributor to a cell such that $C_y = 1$, then there is a risk of re-identification. If the intruder has knowledge of the identity of an individual n, whose characteristics fit the pattern y (i.e., the intruders has knowledge of a record $z^{(n)}$ that fits y), and $|P_{EC}| = 1$, then re-identification takes place. Furthermore, if the length of the pattern y is longer than the number of identifying attributes in z^n (i.e., the attributes in z^n acting as QIDs), then we have attribution because the intruder can now learn some attributes that the intruder had not known previously.

A single contributor to a cell is not the only issue. It is possible that the intruder knows some of the individuals that contribute to a cell with a value of more than one, i.e., $C_y > 1$. For example, when the intruder is in the table himself, or when the intruder colludes with some other contributors from the cell to gain information. This information is used to recognize a single individual with certainty, i.e., to subtract the known individuals from a cell with $C_y > 1$ to create a new cell that has $C_y = 1$.

B. Zero Cells

One major source of attribution is the presence of zero cells in the table. Take Table VI as an example, where there are several zero values. When an intruder has in his background

knowledge a record $z^n = (Arrest = yes, Gender = Female, \ldots)$, then the intruder immediately learns that the age of the arrested individual is not greater than 21. This is referred to as negative attribution [13], whereby we learn which values cannot be attributed to an individual.

TABLE VI. EXAMPLE OF SKEWED TABLE OF ARRESTS IN 2019 (a_3)

		a_1:Age				marginal
		< 18	$18-21$	$21-29$	$29+$	m_{gender}
a_2:Gender	female	1	37	0	0	38
	male	3	0	21	10	32
marginal	m_{age}	4	37	21	10	70

Negative attribution is a form of disclosure that occurs commonly, as any zero cell in a table could lead to negative attribution. The impact of negative attribution is generally smaller than exact attribution. Learning that someone is not exactly 15 years old, is less impactful than learning that someone is exactly 16 years old. However, as negative attributions are common, multiple negative attributions can be combined to have close to exact attribution. An intruder may learn that everyone in Table VI had been trialed as an adult, which can only happen for 15-year olds or older, resulting in a negative attribution that no one is younger than 16. Now, an intruder can learn from Table VI that any female in the table is between the ages of 16 and 22.

In some cases, zero cells can cause exact attribution. In the case of Table VI, if the intruders know of an individual between the age of 21 and 30 that had been arrested, the intruders will immediately learn that the gender of the individual is male.

C. Differencing

Unfortunately, disclosure risk does not originate only from having cells with few or no contributors. If it had been so, it would have been possible to adjust only the specific cells with low numbers of contributors. The values that cells represent in a table are dependent on each other. For example, the value in one cell could be the summed total of several other cell values (like marginals). When the value of a cell is adjusted for its protection, it is sometimes possible to use the cells that are related to that cell to find the original value of the cell.

Suppression is a common method for tabular protection that hides the number in the cell with a "*", "NULL", or any other string or symbol. Such a symbol clearly indicates that the value of the cell is suppressed [11]. Let us reconsider Table VI, only this time we suppress the values of the first column by publishing a new cell $C'_{female,<18} = *$, instead of the correct value $C_{female,<18} = 1$. An intruder does not directly know the number of contributors belonging to the cell with pattern $(female, < 18)$. Further, assume that the intruder does know the number of contributors that belong to the gender marginal, i.e, the right most column in Table VI. We can subtract all other cell values from the marginal to retrieve the original value:

$$\begin{aligned} C_{female,<18} &= m_{gender}(female) - C_{female,18\text{-}21} \\ &\quad - C_{female,21\text{-}29} - C_{female,29+} \\ &= 1. \end{aligned}$$

This common linear relation between the cells makes it much harder to determine which cell to adjust. In the case of suppression, for example, the problem becomes NP-complete [14]. Aside from marginals, other cells may also have a linear relation with each other. When a table describes some flows, for instance, the number of patients following a certain treatment, then one common relation is that the number of outputs of the flow equals the number of inputs of the flow subtracted by the number of cases still being in the flow (e.g., receiving a treatment). These types of relations are inherent to the domain and the processes that the data tracks. Generally, some domain knowledge is necessary to identify such relations.

D. Linking

When assessing the disclosure risk of a table, it is not enough to only examine the table itself, as the cells in a table can be linked to the background knowledge of the intruder and other tables that may have been released in the past. Generally, different table releases use different source (micro)data, which makes information linking unlikely. However, several specific types of data releases have been identified that have a high risk of disclosure through linking [16]. Linking increases the number of methods, which includes the ones previously discussed, for personal data disclosure.

One situation where data linkage likely leads to disclosure is when the same data is released multiple times with slight alterations, for example, when changing the values of an attribute. This has been identified for microdata as the *republication problem* [15]. The same problem persists in tabular data releases. When a table is published with, for example, a cell value $C_{age=16\text{-}19}$, it is possible to adjust the value of the grouping attribute age and release a cell value of $C_{age=15\text{-}19}$. This adjustment may be done for a variety of reasons, such as a new legislation making the new age group of interest or when a new method for generating tables is implemented. Even when both cells are safe given the disclosure risk mentioned above, the difference between these releases generates a new cell $C_{age=15}$ which may not be safe. We refer to [16] for more extensive examples of disclosure through linking.

When searching for disclosures in data, intruders can use linking and differencing to create more cells with zero or unique contributors. Subsequently, intruders can try to re-identify individuals contributing to those cells and carry out attribution attacks against those individuals. This search process is visualized in Figure 1.

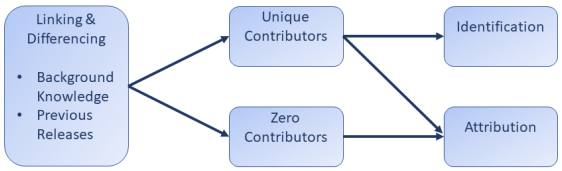

Figure 1. Diagram representing the relation between disclosure sources.

E. Approximate Disclosure

Disclosures do not always happen with complete certainty. If data is appropriately protected, it is possible that none of the disclosure scenarios above occur. When an intruder has some uncertainty in the attribution or re-identification, the intruder could either direct more resources towards confirming

the re-identification or attribution (e.g., via linking the newly discovered information with other data sources). It is possible to have approximate disclosures in re-identification and attribution cases.

Re-identification with sampled data almost never happens with complete certainty. An intruder with access to a sampled table, might potentially discover cells representing only one individual, i.e., $C_y = 1$. The intruder might even know a person that fits pattern y. The risk of disclosure may still be fairly small if the intruder has no knowledge of how large the number of similar individuals in the population is. In other words, the value of the corresponding $|P_{EC}|$ could be one, but can also be let us say 100 (i.e, the individual could be unique or 100 individuals could fit that pattern in the population). In the former, the re-identification with certainty can occur. In certain cases, given uniqueness in the sampled data, it is possible to be fairly certain how many individuals fit that pattern in the population. For example, [17] has shown that 86% of the U.S. population is unique given only a few attributes.

Zero cells commonly cause (negative) group attribution, since intruders can learn that no individual in surrounding groups belongs to the pattern represented by the zero cell. It is also common that, instead of zero, there are only very few individuals belonging to one cell and an overwhelming majority belonging to another cell. Consider the distribution of female arrestees in Table VI, where the majority (i.e., 37) are adults, compared to the single arrestee that was a minor. If an intruder seeks to learn information about a female arrestee, the intruder may conclude with fairly high certainty that she is an adult. In many nations it is common for minors' privacy to be additionally protected by law. Thus, an intruder who learns that his target is likely an adult, might become more motivated to spend more resources in tracking his target, as their information may be more easily accessible (relative to a minor).

Differencing and linking can be applied to find cells with few or zero contributors. This can lead to identification and attribution as depicted in Figure 1. When finding these cells, an intruder may approximate the likelihood that an individual, whom the intruder may know, is associated with some sensitive information. This approximate learning is unavoidable to an extent. However, it is important that intruders do not learn too much sensitive information about individuals either exactly or approximately. For this purpose, various privacy models have been devised for protecting frequency tables.

V. Privacy Models

There are various techniques to protect tabular data from attacks. These techniques adjust the table and the cells by suppressing or rounding cells, restructuring table attributes [11] or adding noise [18]. These techniques aim at having a certain level of privacy. The level of privacy is determined by so-called privacy models, which identify the cells or records that could be at risk of disclosure. The privacy models, although researched independently, are fairly similar for tabular data and microdata. In this section, we aim at clarifying these similarities. Furthermore, note that these privacy models cannot guarantee full protection against all disclosures. But, if they are appropriately applied, they can minimize the likelihood that intruders with reasonable resources are able to disclose sensitive information.

A. Exact Risks

One early and fairly common privacy model in microdata is k-anonymity [19]. This model requires that for any pattern of QIDs, $y = (a_1 = i, \ldots, a_d = j)$, there are at least k individuals that belong to that pattern. Similarly, if we use pattern y to define a cell in the table, then when the privacy model is applied on the microdata it also holds for tabular data: $C_y \geq k$.

One major issue with the k-anonymity privacy model is that it does not protect SATs, even though these are commonly released. k-anonymity ensures for every possible QID pattern to have sufficient individuals. This makes it difficult to identify a single individual and, generally, any SAT published will have values distributed across many QID patterns. However, one common source of disclosure risk arises when a SAT value is skewed towards one group. Take pattern y and extend it with a SAT s, let us take gender as s. In this case, it is possible that $C_{y,s=female} = 0$, whereas $C_{y,s=male} = 15$ (we used this example in Table V). We may not be able to identify an individual based on pattern y from the data, but we learn that any individual that falls under pattern y is a male. This is known as a homogeneity attack and can be reduced with the ℓ-diversity model [20]. ℓ-diversity requires that for all QID patterns there are at least ℓ SAT values (note that this requirement holds for every SAT attribute). In the case of gender, requiring 2-diversity means that $C_{y,s=male} \geq 1$ and $C_{y,s=female} \geq 1$, which would prevent zeroes in the case of two genders. If another sensitive value $j =$ bigender would be possible, then with 2-diversity only two cells have to be nonzero and, in this case, one zero would be allowed in $\{C_{y,male}, C_{y,female}, C_{y,bigender}\}$, thus negative attribution can still occur, but at least ℓ possible values prevent exact attribution without extensive background knowledge.

Both k-anonymity and ℓ-diversity have been developed for microdata protection. Similarly, the minimum frequency rule was developed for protecting frequency tables [21]. The minimum frequency rule restricts all cells to contain at least n individuals, i.e., $C_y \geq n$. This is required for a complete pattern y, this includes all the QIDs and SATs. The minimum frequency rule is a more strict model than ℓ-diversity as it prevents any negative attribution. This provides additional protection but may remove too much information in certain cases. ℓ-diversity can already suffer from too much information loss with multiple SATs due to the curse of dimensionality [22]. In protecting frequency tables, the minimum frequency rule requires removing zero cells, which may reduce the number of grouping attributes in order for a table to uphold the rule.

Whether zeroes are acceptable varies by case, as the impact of negative attribution is severely less than of exact attribution. The minimum frequency rule is the strongest in prevention of zero cells, ℓ-diversity prevents exact attribution through zero cells, and k-anonymity leaves the most information for data users, but also provides the least protection against negative or exact attribution through those cells.

B. Approximate Risks

A data controller may prevent exact identification and attribution of individuals, given that intruders have a limited set of resources for re-identification or attribution. However, approximate attribution in some cases might be sufficient for

 ISBN: 978-1-61208-760-3

an intruder with enough background knowledge, or resources, to learn sensitive information.

A risk of approximate attribution exists when there is a skew between subgroups, e.g., when almost everyone that fits a certain pattern y belongs to a specific age group. This concern has been identified for microdata, which resulted in Entropy ℓ-diversity being introduced [20]. The entropy restriction requires that individuals that fit pattern y are well distributed across the values of their SATs. For example, for a SAT s in the set of sensitive attributes S, entropy can be defined as:

$$Entropy(y) = -\sum_{s \in S} p(C_{y,s}) \log p(C_{y,s}),$$

where $p(C_{y,s})$ is the fraction of contributors that belong to pattern y and sensitive value s. If $Entropy(y) \geq \log \ell$ holds for all patterns y, then the data is considered safe.

Even with entropy ℓ-diversity, it is still possible to learn something about pattern y. For example, if we take a SAT with value s_1 representing some rare disease and value s_2 some common disease, we can take $p(C_{y,s_1}) \approx p(C_{y,s_2}) \approx 0.5$, which will have very high entropy and be considered to have a low risk of disclosure. Assume the fraction of $p(s_2)$ in the data as a whole is very small ≤ 0.001, then intruders can learn that individuals that fit pattern y have an abnormally large chance of having the rare disease. This is known as the skewness attack and can be prevented with the t-closeness model [5]. The t-closeness model requires the distribution of the attribute values $s \in S$, for every pattern y, to be similar to the distribution of s in the data as a whole.

ℓ-diversity and t-closeness have been developed for microdata. For tabular data, a work recently introduced privacy models that includes an entropy constraint for tabular data [23][24]. The entropy constraint in the model is a generalization of ℓ-diversity and t-closeness. Instead of using the entropy of S over a pattern y, this model restricts the entropy of the distribution over C (all cells). Computing the entropy over C, ignoring the grouping of cells by their SATs, and instead, restricting the entropy for all cells regardless of their SATs or QIDs, allows for more protection on the table as a whole. As ℓ-diversity and t-closeness solely restrict the distributions with respect to SATs, they reduce only the most impactful disclosure risks (on SATs) while maintaining more useful information. However, determining which attributes should be considered as QIDs or SATs is not always a trivial task [25].

C. Differential Privacy

A data controller may not always be able to determine the QIDs and SATs accurately, as it is difficult to know what kind of information is out there. A data controller may be aware of intruders with a lot of resources and background knowledge, but may be unable to specify the means these intruders possess exactly. For such cases, differential privacy models have been developed, originally for microdata [6], and later also introduced for tabular data [26]. ϵ-differential privacy requires that the effect that individuals have on the data is limited. Take data sets DS and DS^*, where the difference between the two is a single individual, the result r (the result of an analysis or a query), for both data sets, has to be similar enough such that:

$$\frac{P(r|DS)}{P(r|DS^*)} \leq e^{\epsilon}. \quad (3)$$

The advantage of applying such a model is that data controllers have a theoretical guarantee for containing personal data disclosures with a tuneable parameter ϵ. This guarantees holds regardless of the information an intruder may possess [6]. This property makes ϵ-differential privacy useful for cases where data controllers do not know much about the intruders' background knowledge. Note that this guarantee applies to the definition of ϵ-differential privacy, according to which 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 [6]. Whether this definition of privacy is comprehensive and adequate has not been established yet.

Recently, there has been a significant demand for on-line table generation, which allows the user to query data numerous times, instead of receiving a pre-processed data set. This has advantages for both data user and data controller, but it also comes with the issue that if the data users are also potential intruders, then, they now possess resources for differencing and linking [27][28]. In such environments, especially when data users' queries are minimally controlled, the differential privacy model is indispensable. More syntactic privacy models [29], such as k-anonymity and t-closeness, cannot protect the data against such intruders, unless the queries are restricted or tracked to prevent the intruder's background knowledge from increasing too much.

One issue with differential privacy is that, in order to provide the guarantee (3), a stochastic mechanism is required to transform the data. This transformation is generally achieved by using some distribution of noise [30]. Applying noise or some other stochastic mechanism to transform the data has a probability that the transformed data differs significantly from the original data. This issue in the so-called the range of correctness [29], i.e., the possible original values that the transformed data represents, makes it more difficult to apply the differential privacy model. In cases where it is expected that the published statistics about crime, income, etc., are close to their actual numbers (e.g., when releasing census tables), large potential variations are unacceptable. This issue has been identified in US census data, where due to smaller samples with some outliers, the added variance from the noise could vary between 1000% and 7000% for moderate levels of privacy [31]. There are more variations of differential privacy that vary slightly in the privacy and utility levels but provide similar theoretical guarantees as (3) and require some stochastic mechanism to work as well. For a more extensive view focused on differential privacy, we refer to [30].

VI. Conclusion and Future Work

This study examined privacy models from microdata and tabular literature through a unified formalization. It was found that the personal privacy models in the tabular data literature are more privacy-preserving, and thus require less information to be released than their microdata counterparts. The dimension, i.e., the number of attributes of the data that is published, is generally much larger for microdata than for tabular data. When there are more attributes, intruders may learn more from attribution, additionally, intruders have more attributes at their

 ISBN: 978-1-61208-760-3

disposal for re-identification. As such, having stricter privacy models for tabular data may seem counterintuitive.

From the comparison of the privacy models, it becomes clear that the background knowledge is differently assessed between the two data types. The privacy models for tabular data require the same protection for all cells, regardless of the classes, such as QIDs or SATs. If done correctly, using microdata privacy models allows for releasing more data with minimal increase of privacy risks. However, the process of classifying the attributes into QIDs, SAT, etc., is difficult. Due to tabular data being more aggregated by nature, it generally needs to be less transformed/processed (than its microdata counterpart) to provide similar privacy guarantees. For the same reason, the information in tabular data is more robust against stronger privacy models, which means that adhering to stronger privacy models causes less information loss for tabular data. However, as the dimension of tables increases, we suspect that using attribute classes, such as QIDs and SATs, in data privacy models may become unavoidable. Additional research is required into the loss of information from privacy models for various release purposes and dimensions of tables. A different manner to avoid the process of having to assign attribute classes is by applying differential privacy. One issue of differential privacy for tabular data is that the range of correctness for values can become very large. This can increase the variance tremendously, as shown in previous works, which may be unacceptable for common tabular data releases, such as census tables.

Microdata and tabular data are very similar, however, there are differences in practical release purposes that cause tabular data to generally require more accuracy for data users in their privacy models. A possible future work is to assess how protection methods, i.e., the transformation done on the data, differ between microdata and tabular data. Of interest would be to investigate whether similar differences can be found in protection methods, for tabular data and microdata, and whether they depend on the slight differences in the purposes of release for the respective data types.

REFERENCES

[1] I. Hrynaszkiewicz, M. L. Norton, A. J. Vickers, and D. G. Altman, "Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers," Bmj, vol. 340, 2010, p. c181.

[2] C. Culnane, B. I. P. Rubinstein, and V. Teague, "Health Data in an Open World," arxiv preprint, 2017. [Online]. Available: http://arxiv.org/abs/1712.05627

[3] M. Douriez, H. Doraiswamy, J. Freire, and C. T. Silva, "Anonymizing NYC taxi data: Does it matter?" Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, 2016, pp. 140–148.

[4] G. J. Matthews, O. Harel, and R. H. Aseltine, "Privacy protection and aggregate health data: a review of tabular cell suppression methods (not) employed in public health data systems," Health Services and Outcomes Research Methodology, vol. 16, no. 4, 2016, pp. 258–270.

[5] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," 2007 IEEE 23rd International Conference on Data Engineering, no. 2, 2007, pp. 106—-115.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3876 LNCS, 2006, pp. 265–284.

[7] J. Castro, "Statistical disclosure control in tabular data," Advanced Information and Knowledge Processing, vol. 51, no. November, 2010, pp. 113–131.

[8] B. Barak et al., "Privacy, accuracy, and consistency too," Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2007, p. 273.

[9] M. S. Bargh, R. Meijer, and M. Vink, "On statistical disclosure control technologies," Tech. Rep., 2018.

[10] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," Nature Communications, vol. 10, no. 1, 2019, p. 3069. [Online]. Available: http://www.nature.com/articles/s41467-019-10933-3

[11] A. Hundepool et al., Statistical Disclosure Control. John Wiley & Sons, 2012.

[12] M. Elliot and A. Dale, "Scenarios of attack: the data intruder's perspective on statistical disclosure risk," Netherlands Official Statistics, vol. 14, no. Spring, 1999, pp. 6–10.

[13] M. Elliot, E. Mackey, K. O'Hara, and C. Tudor, The Anonymisation Decision-Making Framework. UKAN, 2016, vol. 1, no. 2006. [Online]. Available: https://fpf.org/brussels-privacy-symposium-final-papers/

[14] J. P. Kelly, B. L. Golden, and A. A. Assad, "Cell suppression: Disclosure protection for sensitive tabular data," Networks, vol. 22, no. 4, 1992, pp. 397–417.

[15] X. Xiao and Y. Tao, "M-invariance: towards privacy preserving re-publication of dynamic datasets," SIGMOD Conference, 2007, pp. 689–700. [Online]. Available: http://portal.acm.org/citation.cfm?id=1247556

[16] B. C. M. Fung et al., "Privacy-preserving data publishing," ACM Computing Surveys (Csur), vol. 42, no. 4, 2010, pp. 1–53. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1749603.1749605

[17] P. Golle, "Revisiting the uniqueness of simple demographics in the US population," Proceedings of the ACM Conference on Computer and Communications Security, 2006, pp. 77–80.

[18] V. Leaver, "Implementing a method for automatically protecting user-defined Census tables," Joint ECE/Eurostat Worksession on Statistical Confidentiality in Bilbao (December 2009), http://www. unece. org/stats/documents/2009.12. confidentiality. htm, 2009.

[19] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 1998.

[20] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy beyond k-anonymity," ACM Transactions on Knowledge Discovery from Data, vol. 1, 2007, p. 3.

[21] A. Hundepool et al., "Handbook on Statistical Disclosure Control," Statistical Disclosure Control, 2010, pp. 1–288.

[22] C. C. Aggarwal and P. S. Yu, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms," 2008, pp. 11–52.

[23] L. Antal, N. Shlomo, and M. Elliot, "Measuring Disclosure Risk with Entropy in Population Based Frequency Tables," 2014, pp. 62–78.

[24] ——, "Disclosure Risk Measurement with Entropy in Two-Dimensional Sample Based Frequency Tables," pp. 1–10.

[25] G. T. Duncan and M. Elliot, Statistical confidentiality, 2010.

[26] Y. Rinott, C. M. O'Keefe, N. Shlomo, and C. Skinner, "Confidentiality and differential privacy in the dissemination of frequency tables," Statistical Science, vol. 33, no. 3, 2018, pp. 358–385.

[27] G. Thompson, S. Broadfoot, and D. Elazar, "Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of St," no. October, 2013, pp. 28–30.

[28] N. Shlomo, L. Antal, and M. Elliot, "Measuring disclosure risk and data utility for flexible table generators," Journal of Official Statistics, vol. 31, no. 2, 2015, pp. 305–324.

[29] C. Clifton and T. Tassa, "On syntactic anonymity and differential privacy," Transactions on Data Privacy, vol. 6, no. 2, 2013, pp. 161–183.

[30] C. Dwork, "Differential privacy: A survey of results," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4978 LNCS, 2008, pp. 1–19.

[31] R. Sarathy and K. Muralidhar, "Evaluating laplace noise addition to satisfy differential privacy for numeric data," Transactions on Data Privacy, vol. 4, no. 1, 2011, pp. 1–17.

 ISBN: 978-1-61208-760-3