International Journal of Information Security (2023) 22:2015–2024
https://doi.org/10.1007/s10207-023-00730-x

REGULAR CONTRIBUTION

Attribute disclosure risk for *k*-anonymity: the case of numerical data

Vicenç Torra[1] · **Guillermo Navarro-Arribas**[2]

Published online: 25 July 2023

Abstract

k-Anonymity is one of the most well-known privacy models. Internal and external attacks were discussed for this privacy model, both focusing on categorical data. These attacks can be seen as attribute disclosure for a particular attribute. Then, *p*-sensitivity and *p*-diversity were proposed as solutions for these privacy models. That is, as a way to avoid attribute disclosure for this very attribute. In this paper we discuss the case of numerical data, and we show that attribute disclosure can also take place. For this, we use well-known rules to detect sensitive cells in tabular data protection. Our experiments show that *k*-anonymity is not immune to attribute disclosure in this sense. We have analyzed the results of two different algorithms for achieving *k*-anonymity. First, MDAV as a way to provide microaggregation and *k*-anonymity. Second, Mondrian. In fact, to our surprise, the number of cells detected as sensitive is quite significant, and there are no fundamental differences between Mondrian and MDAV. We describe the experiments considered, and the results obtained. We define dominance rule compliant and p%-rule compliant *k*-anonymity for *k*-anonymity taking into account attribute disclosure. We conclude with an analysis and directions for future research.

Keywords Data protection · Masking methods · Reidentification · Attribute disclosure · *k*-anonymity · Microaggregation

1 Introduction

Masking methods [1, 2] are one of the tools for data protection when data needs to be shared. So, they are tools for data sharing or data publishing. They provide a way to avoid disclosure while keeping the analytical properties of the data. That is, the goal is to modify the data so that disclosure does not take place, and at the same time, the data is still useful for its use.

There are several methods that have been defined for this purpose. They include additive and multiplicative noise, microaggregation [3], rank swapping, generalization. For each of them there are different variations, some of them addressing different types of disclosure.

✉ Vicenç Torra
vtorra@ieee.org

Guillermo Navarro-Arribas
guillermo.navarro@uab.cat

[1] Department of Computing Sciences, Umeå University, Umeå, Sweden

[2] Department of Information and Communications Engineering, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain

There are mainly two types of disclosure: identity disclosure and attribute disclosure. Identity disclosure is successfully finding someone in a (protected) database. Attribute disclosure is about increasing the knowledge of a particular individual on a particular attribute or variable.

Privacy models [2, 4] are computational definitions of privacy. There are different privacy models depending on the type of disclosure we want to avoid. Differential privacy [5, 6], privacy for re-identification [7], *k*-anonymity [8–10], and some of their variations are the most well-known privacy models. For data publishing *k*-anonymity and local differential privacy are some of the most well-known and used methods. A database is compliant with *k*-anonymity when for each record or individual, there are other $k-1$ records that are indistinguishable for a set of quasi-identifiers. Here, quasi-identifiers refer to those attributes that intruders may know. In other words, *k*-anonymity implies that identity disclosure cannot take place and intruders, taking into account their background knowledge, will retrieve for any individual k possible records. This is equivalent to say that there are anonymity sets of cardinality at least k.

Some extensions of *k*-anonymity were introduced, as avoiding identity disclosure does not avoid attribute disclosure. In particular, when all *k*-indistinguishable records share

the same value for a confidential attribute, any intruder can infer that this very value is the one of any individual in the anonymity set. Privacy models were introduced to avoid this type of disclosure. Among them we find p-sensitivity [11, 12] and l-diversity [13]. All these approaches focuses on attribute disclosure for categorical data. More particularly, they focus on the class label and the inferences an intruder can make based on this information. Up to our knowledge, no analysis has been done so far about attribute disclosure for numerical attributes. We consider that this type of disclosure is also relevant for k-anonymity when data is numerical.

More particularly, in this paper we discuss an attack on k-anonymity for numerical data based on standard tabular data protection privacy models [1, 14–16]. Tabular data summaries of data consists of in terms of a few attributes. Tabular data is typically published by national statistical and economic offices. For example, aggregated salaries (or mean salaries) for each pair (profession, town), or business revenues for each pair (type of business, town). Nevertheless, the values in a cell (e.g., the pair profession, town) can lead to disclosure. Several rules have been proposed for detecting which cells are sensitive. For example, rules check whether there are contributors that can estimate, using their own figures, the contribution of others.

These rules for tabular data have never been considered in the context of microaggregation, while it is clear that the setting is similar (i.e., groups of clusters and data for certain variables), and that this would constitute attribute disclosure. While p-sensitivity and l-diversity were focusing on attribute disclosure for categorical attributes, our focus in this paper is on numerical attributes. In this paper we consider this type of disclosure and show that this type of disclosure is significant for a large range of parameters. In fact, the risk is quite larger than what we were expecting before doing this analysis. We show the results of disclosure for k-anonymized data using MDAV (a microaggregation algorithm) and Mondrian (a well-established method for k-anonymity. As we will see later, the results are similar for both types of algorithms.

The structure of the paper is as follows. In Sect. 2 we introduce the MDAV and Mondrian methods used to provide k-anonymized data, we provide some preliminaries about attribute disclosure, and describe the rules used in tabular data protection to detect that cells are sensitive. In Sect. 3 we introduce our methodology and experiments. We conclude in Sect. 4 with an analysis and directions for future research.

2 Preliminaries

In this section we briefly review two masking methods to provide k-anonymity. They are microaggregation and Mondrian. Then, we review rules used for tabular data to detect sensitive rules.

2.1 Masking methods to provide *k*-anonymity

Microaggregation [3, 17, 18, 28] is one of the methods used to provide k-anonymity. It is based on clustering. Given a database X, the procedure consists of building small clusters (each with at least k records) and then replacing each of the records by the cluster center.

As all the records that belong to the same cluster are replaced by the cluster center, and there are at least k records in each cluster, k-anonymity is satisfied.

Microaggregation is defined as an optimization problem with constraints. The objective function is as in k-means. That is, we consider cluster centers, and records should be assigned to the nearest cluster center. Constraints are defined so that each record is assigned to one and only one cluster, and in addition, each cluster should have at least k records assigned to it (and at most $2k$).

Univariate microaggregation stands for microaggregation on a file with a single variable. Multivariate microaggregation when data is in a n-dimensional space with $n > 1$. Multivariate microaggregation is an NP-problem [19]. Because of that, heuristic methods have been developed. MDAV [3, 20] is one of such methods and has been extensively used in the literature. In this work we use MDAV for microaggregation.

Mondrian [21, 25] is another approach to provide k-anonymity. It provides a way to create a partition of the original database in a top-down way. Its definition is recursive. A data set with more than $2k$ records is divided into two parts, each with approximately the same number of records. The process is repeated until each part has between k and $2k - 1$ records. Splitting a set X' consists of selecting an attribute V, a cut point of the domain of V (i.e., $v_0 \in Dom(V)$) so that we have half records in X' smaller than or equal to v_0 and half records in X' larger than or equal to v_0. Once these clusters are built, Mondrian describes the region, or alternatively, we can build the cluster centers of the data in the cluster. We will use this second approach.

2.2 Attribute disclosure in *k*-anonymized data

Attribute disclosure in k-anonymized data has been mainly studied in relation to categorical confidential attributes. In a common scenario of microdata protection, one distinguishes between identifier, quasi-identifier and confidential attributes. Identifier attributes, which can unambiguously identify a single record are commonly removed or encrypted. Quasi-identifiers are attributes that can be linked with external information to reidentify a record. Even if a single quasi-identifier cannot identify an individual, a combination of different quasi-identifiers might. On the other hand, confidential attributes are those, which contain sensitive information on the respondent.

These definitions are usually made in terms of identity disclosure, but they also affect attribute disclosure. That is, a set of quasi-identifiers might not be used to completely reidentify a respondent but could provide new information about it.

When protecting microdata to produce e.g. a k-anonymous dataset, it is relatively common, to apply the protection method to the quasi-identifiers and leave the confidential attribute/s untouched. This introduces several problems which have been widely studied in the literature. Most notably homogeneity and similarity attacks, and skewness attacks.

The homogeneity or similarity attack, happens when all confidential attributes are equal, or semantically similar, for the same anonymity group. This clearly leads to attribute disclosure if the attacker can link a respondent to the protected quasi-identifiers of an anonymity group.

In this scenario several proposals have appeared in the literature. For example, p-sensitivity [12] requires the number of sensitive values in an anonymity set to be at least p. A stronger proposal is l-diversity [13]. It extends this idea to require l well-represented values in each anonymity set, where *well represented*, can be expressed in different terms. Similarly, t-closeness [26], requires the distribution of sensitive values within each anonymity set to be similar to the overall distribution of sensitive values in the dataset. Other works build in the same line, e.g. in [22] the relation between quasi-identifier (masked) and sensitive (non-masked) attributes is also taken into consideration. All these proposals are for categorical attributes.

As far as we know, no work has considered the potential disclosure of aggregated numerical data. Even if all attributes are masked, that is, there is no unmasked confidential attributes we believe there might be attribute disclosure to some extent. This is specially relevant to masking methods that rely in numeric aggregation to provide privacy, but can also be considered for methods that rely in generalization. The main idea is that each contribution to the aggregated (or generalized) value can gain some information when knowing this aggregated value, depending on the concrete contributed value.

This is related to what some authors denote as internal attack, where one of the correspondents of the protected data can gain knowledge of other correspondents by observing the protected data. Moreover the attacker could craft special data values to increase the attribute disclosure.

2.3 Sensitive rules in tabular data

Tabular data publishing consists of releasing aggregates of data built in terms of a few variables. Given a database X, for each combination of values of r variables, either the count (i.e., the number of records in X with such combination) or the aggregate of another variable in X is provided.

Let us consider that for a given cell we have t contributors which provide the values $c_1, \ldots, c_t$. Then, tabular data protection provides several rules to determine that a cell is sensitive. We review them here (see e.g. [1, 14, 16] for details).

The rule (n_r, r_r)-dominance determines that the cell is sensitive when n contributors represent more than the r fraction of the total. If we consider the values c_i ordered in decreasing order, $c_{\sigma(1)} \geq c_{\sigma(2)} \geq \cdots \geq c_{\sigma(t)}$, this rule will detect a cell as sensitive when

$$\frac{\sum_{i=1}^{n_r} c_{\sigma(i)}}{\sum_{i=1}^{t} c_i} > r_r. \tag{1}$$

The rule $p\%$ is stated as follows. A cell is sensitive when an intruder can estimate the contributor within p percent, taking into account the released table. It can be proven that the best estimation is the one of the second largest contributor (i.e., the one which contributes with $c_{\sigma(2)}$) on the largest one. Then a cell is sensitive when

$$\sum_{i=3}^{t} c_{\sigma(i)} < p c_{\sigma(1)}. \tag{2}$$

In this expression we use p as a value in [0,1] instead of a percentage.

For the dominance rule, parameters of $(1, 0.6)$ as well as parameterizations with $n_r = 2$ and with $r_r > 0.6$ have been considered in the literature. For the rule $p\%$, a parameter larger than 60% has also been considered in the literature. Moreover, Hundepool et al [1] recommend the use of $p' = (1-r_r)/r_r$ (and $p\% = 100p'$) as providing a risk assessment similar to the $(2, r_r)$ rule. E.g., for $n_r = 2$ and $r_r = 0.6$, we would have $p = 66\%$. We are using this parameterization in our experiments. Note that this expression is monotonic decreasing with respect to r_r. This is illustrated in Fig. 1.

3 Methodology and experiments

As we have explained in the introduction, our goal is to assess attribute disclosure for a k-anonymous file using the rules for tabular data to detect sensitive cells. As we have seen in the previous section, a sensitive cell is one in which a contributor can make a good guess of the values of one or more contributors. This corresponds, using the terminology used in the attacks on k-anonymity, an internal attack.

In order to make our attack and experiments clear, we will introduce the following notation. We denote the original file by X and the k-anonymous masked file X'. Then, $X' = \rho(X)$ for a masking method ρ. Both files X and X' consist

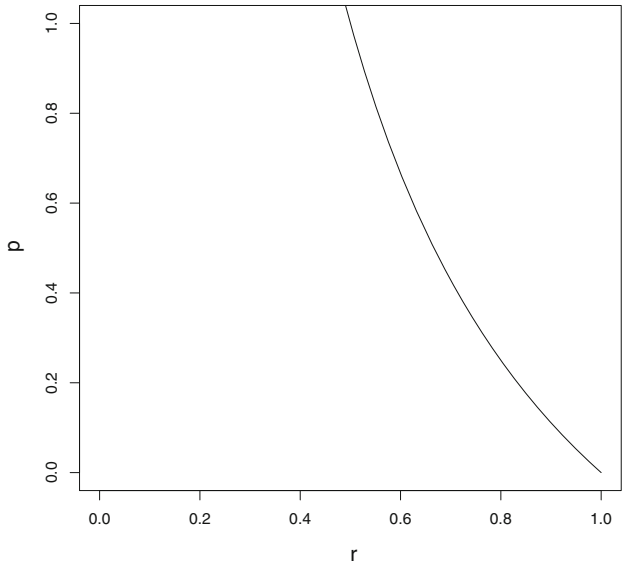

Fig. 1 Relationship between p and r_r which provides a risk assessment for the rule $p\%$ similar to the one of $(2, r_r)$-dominance rule

of m records and n variables. That is, each $x \in X$ is a n-dimensional vector. We consider numerical data. So, x is in $\mathbb{R}^n$.

Then, a k-anonymous file consists of n_c clusters each of them with at least k records and at most $2k$ records. Note that most algorithms provide $n_c = \lfloor m/k \rfloor$ clusters for m the number of records of the file. As there are n_c clusters and each cluster is described in terms of n attributes, this means that we have $n_c \cdot n$ cells. I.e., each cell is defined for a pair cluster, attribute.

Then, given a cell, using the notation above, we have the contributions $c_1, \ldots, c_t$ assigned to the cell. In other words, we have t records assigned to cluster $j \in \{1, \ldots, n_c\}$ and the values of these t records for the attribute $v \in \{1, \ldots, n\}$ are these values $c_1, \ldots, c_t$. Naturally, $t \geq k$ and usually $t < 2k$. Then, the value that represents this cell is $(1/t) \sum_{i=1}^{t} c_i$, which, when t is known, is equivalent to knowing $\sum_{i=1}^{t} c_i$.

3.1 Methodology

For a given data set X' with n attributes that consists of n_c clusters generated using a k-anonymity procedure, we have $n_t = n_c \cdot n$ cells to consider. Given the parameters (n_r, r_r) for the (n_r, r_r)-dominance rule and the parameter $p\%$ for the rule $p\%$ we can count the number of sensitive cells. This number of sensitive cells is denoted by n_{n_r,r_r} and n_p, respectively. We will use this number of sensitive cells (absolute number of sensitive cells) but also the proportion with respect to the total number of cells. The latter corresponds to $n_{n_r,r_r}/n_t$ and n_p/n_t. We consider this proportion as the total number of cells n_t can change significantly when we consider different number of clusters and attributes.

Table 1 Data files used in the experiments, together with the number of records and attributes considered. The categorical class in the Iris data set was translated to a numerical attribute in our research

File name	n. attributes	n. records
Concrete	9	1030
Abalone	9	4177
f1080	13	1079
Iris	5	150
Ionosphere	35	351
Adult	6	48,842

We have considered two alternative methods ρ to generate X' from X. They are microaggregation using MDAV and Mondrian. We have used our own implementations for these algorithms.

3.2 Parameters and files

We have considered different parameters for k (minimal number of records in a cluster), (n_r, r_r) parameters for the (n_r, r_r)-dominance rule, $p\%$ for the rule $p\%$. We have considered a large number of different values of k from 1 to 125. More particularly, we have considered $k = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 21, 22, 23, 24, 25, 30, 40, 50, 75, 100, 125\}$. We have considered (n_r, r_r) with values of $n_r = 1, 2$ and r_r equal to 0.6, 0.7, and 0.8. With respect to $p\%$, we have considered $p = 60\%, 66\%, 70\%$.

In addition, we have considered different data files which consists of different number of records and attributes. In particular, with respect to data files, we have considered the following six data sets: concrete, abalone, 1080, iris, ionosphere, and adult. All of them except 1080 are available at the UCI repository [23]. 1080 data set has been used previously in the data privacy literature and is available in R package sdcMicro [20]. Description of these files in terms of the number of numerical attributes and the number of records can be found in Table 1.

All experiments were done in Python in a regular laptop. We have implemented our own versions of MDAV and Mondrian. Code will be available in [27].

3.3 Experiments

We have considered different parameterizations to observe their effects on the number of cells sensible.

For our surprise, the first observation from our initial experiments using MDAV and a single file X was that the number of cells that are sensible is very high when k is relatively small. A value of k equal to 5, which is considered by some as a good value in k-anonymity, poses a significant risk when the number of variables is as few as 9. For illustration

Table 2 Number of sensible cells for abalone and iris data set, when data is masked with MDAV and different parameters *k*

	Abalone			Iris		
	Dominance rule	Rule $p\%$	n. cells	Dominance rule	Rule $p\%$	n. cells
1	750	0	750	37,591	0	37,593
2	375	375	375	18,791	18,782	18,792
3	250	13	250	12,519	909	12,528
4	10	0	185	473	10	9396
5	4	1	150	44	6	7515
6	1	0	125	6	5	6264
7	1	1	105	5	5	5364
8	0	0	90	5	4	4698
9	0	0	80	5	4	4176

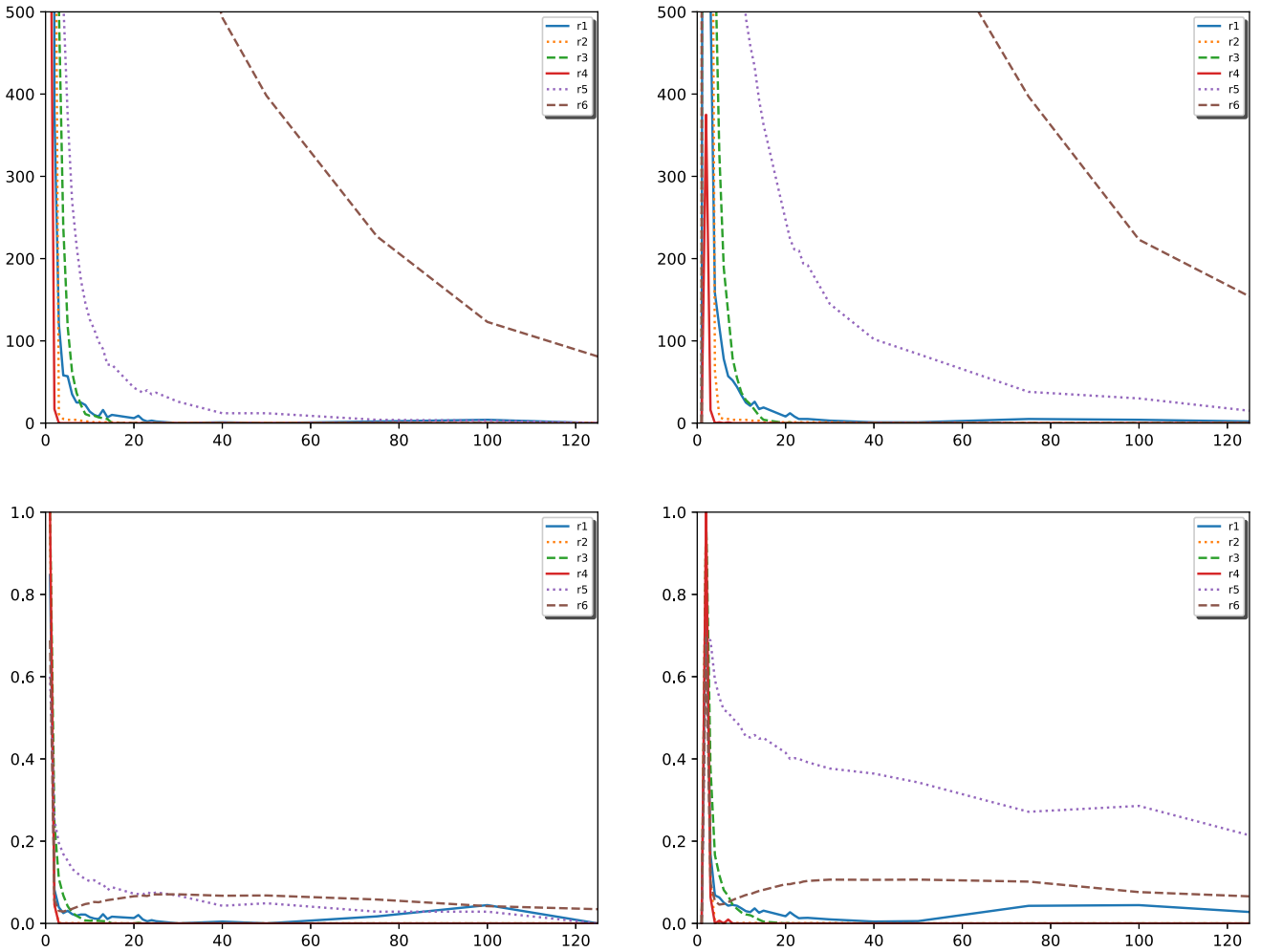

Fig. 2 Attribute disclosure risk for concrete, abalone, 1080, iris, and ionosphere data sets. Microaggregation using MDAV with k={1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 21, 22, 23, 24, 25, 30, 40, 50, 75, 100, 125}. Risk using (n, r)-dominance rule with (1, 0.6) (left) and rule $p\%$ (right) with $p = 66.6\%$, both absolute counts of cells (top) and normalized by the number of all cells (bottom). In the figure, r1 to r6 correspond to concrete, abalone, 1080, iris, ionosphere, and adult data sets

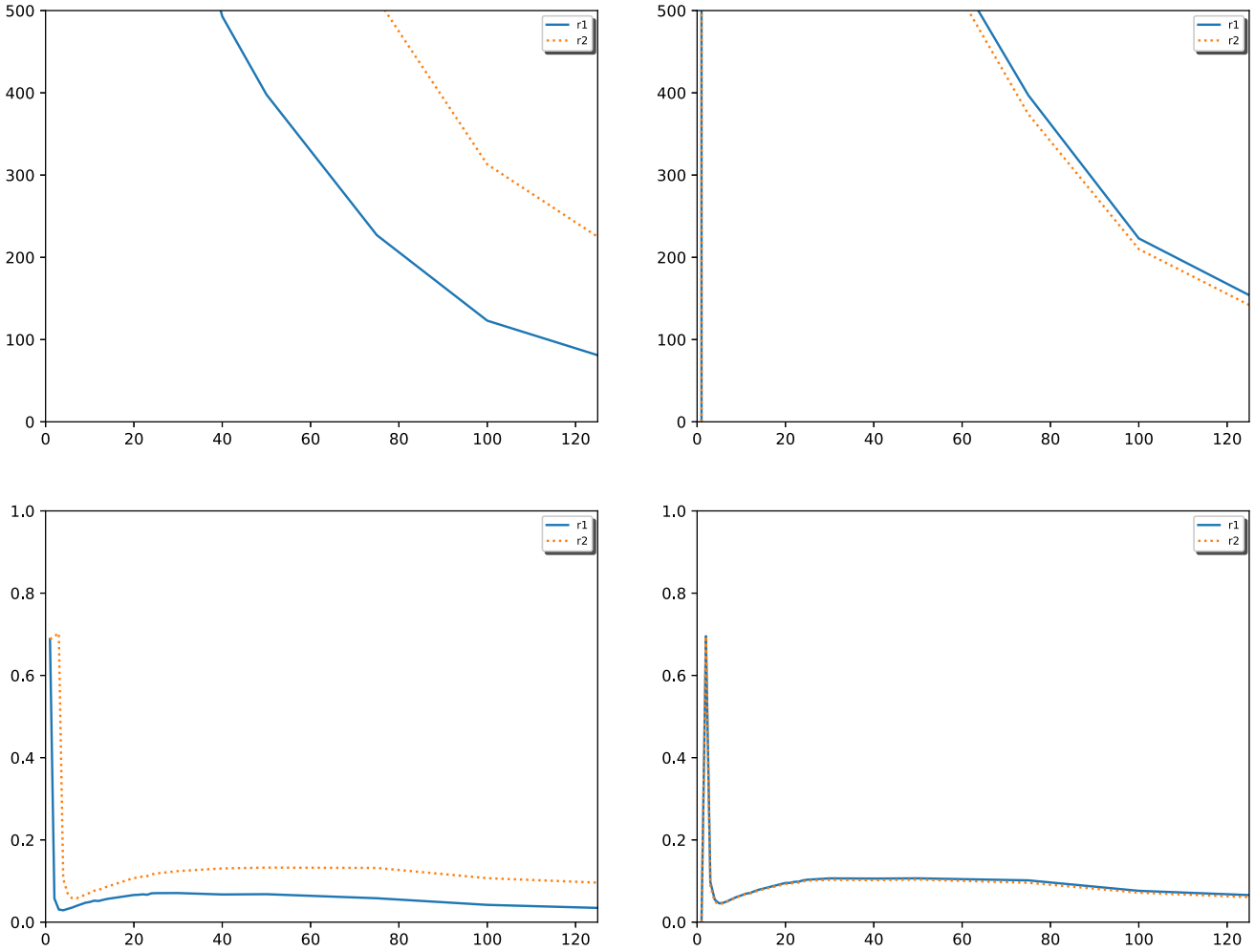

Fig. 3 Attribute disclosure risk for concrete, adult data set comparing $(n = 1, r = 0.6)$-dominance rule (r1) and $(n = 2, r = 0.6)$-dominance rule (r2)—on the left—and rule p%=60% (r1) and p%=66% (r2)—on the right, both absolute counts of cells (top) and normalized by the number of all cells (bottom)

of these findings, we include in Table 2 the number of sensitive cells using $(n_r, r_r) = (1, 0.6)$ in the dominance rule and $p\% = 60\%$ for the rule $p\%$ for the iris data set (5 attributes) and for the abalone data set (9 attributes), for different values of k.

This initial observation triggered the extensive analysis of a large number of parameters k and considering data sets with different number of variables. In addition, while the original data set was masked using MDAV, we then considered Mondrian as a completely different approach to also provide k-anonymity. We show later that the results of using Mondrian are comparable to those using MDAV.

First, we illustrate in Fig. 2, the results related to risk for the six data sets considered, using different parameters k and $(n_r, r_r) = (1, 0.6)$ and $p = 66.6\%$ for the sensitive rules. We provide both absolute counts and relative counts. We can observe in the figures that, in general, the larger the k the smaller the risk, both in absolute terms and relative terms.

Expression 1 given above shows that the dominance rule will increase the number of sensitive cells when n_r is increased. Similarly, the number of cells decreases, when r_r increases. Similarly, Expression 2 shows that increasing p will increase the number of cells detected as sensitive. We can observe this relationship in Fig. 3 for the adult data set.

We have also considered how the number of variables influence the number of sensitive cells. For this, we have considered some of the data files above with some of the parameterizations for the sensitive rules. Then, we have applied these same parameters to the n_i first variables in the file. We have considered all cases from 1 to the number of variables in the file. We have considered the two rules above. We observe that (naturally) the larger the number of variables considered, the larger the number of sensitive cells (as there are more cells to consider). Besides of that, this relationship also appears (although not so clearly stated) when we consider the relative number of sensitive cells. Figure 4

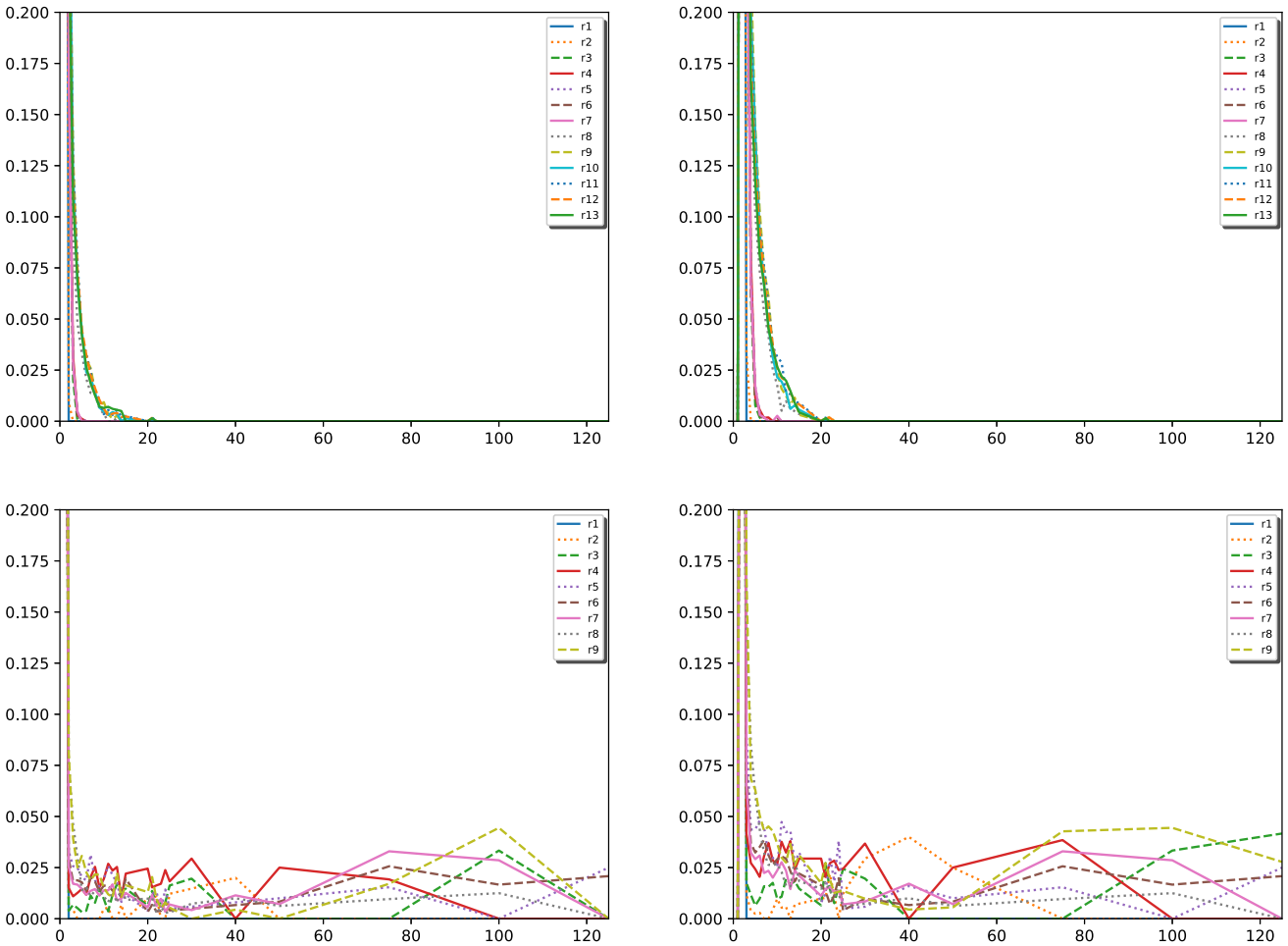

Fig. 4 Attribute disclosure risk for 1080 (top) and concrete (bottom) for $(n = 1, r = 0.6)$-dominance rule (left) and p%=66% rule (right).Number of cells normalized by the total number of cells. r1 to r9 and r1 to r13 correspond to the number of attributes considered. r1 corresponds to 1 attribute, r2 to 2 attributes, etc

illustrates the relative proportion of sensitive cells for both 1080 (top) and concrete (bottom) data sets. We include the results for both dominance (left) and $p\%$ rule (right). We use here an y scale of [0,0.2] to better visualize the results.

As we have explained above, we have considered two approaches for achieving k-anonymity: microaggregation using MDAV and Mondrian. We have observed that, in general, for the different data sets and parameterizations, the results are quite similar. In Fig. 5 we provide the results for two of the data sets (abalone and ionosphere) for which we have found not so similar results. For these data sets, Mondrian seems to behave a little bit better with respect to the attribute disclosure risk.

4 Analysis and conclusions

The literature has discussed attribute disclosure for k-anonymity when data is categorical. Some privacy models as p-diversity and l-diversity were introduced to formally define the associated privacy risk. No previous work has considered the potential disclosure of aggregated data. In this paper we have shown that k-anonymity for numerical variables can also lead to attribute disclosure. We have proposed to use the tools provided for tabular data protection for this purpose. We have shown that the (n_r, r_r)-dominance rule and the rule p% permits to assess this type of risk, and determine when a numerical microaggregated file satisfies an appropriate privacy requirement.

It was completely unexpected to us that the number of cells detected as unsafe was so large. Note that the basic idea in microaggregation is that records assigned to a cluster are similar, and, thus, one would expect that such values

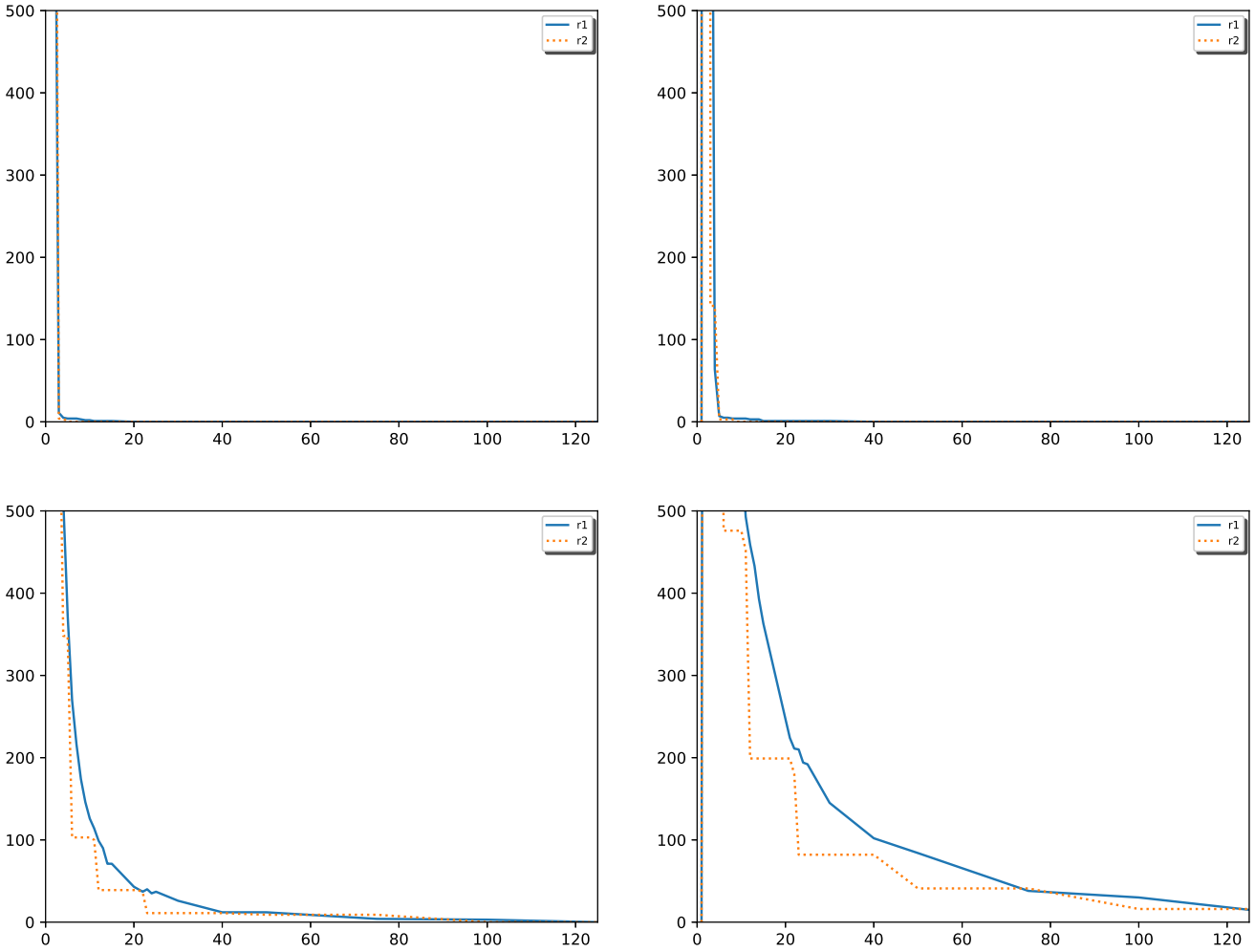

Fig. 5 Attribute disclosure risk for abalone (top) and ionosphere (bottom) for ($n = 1, r = 0.6$)-dominance rule (left) and p%=66% rule (right). Number of cells normalized by the total number of cells. r1 corresponds to the results with MDAV and r2 to the results using Mondrian

are, thus, also similar. Similar values do not imply any risk. Observe that when $c_i = c_j$ for all $i \in \{1, \ldots, t\}$ then the cell is absolutely safe. This unexpected results triggered our experiments. That is, considering different data sets and two algorithms.

More particularly, we have analyzed k-anonymity provided by two different approaches: microaggregation with MDAV and Mondrian. The first one is based on clustering, and the second one builds a partition using a greedy algorithm. We have shown that both approaches lead to similar results and we cannot state that one is better than the other with respect to attribute disclosure, as defined here. As most k-anonymity methods follow one of these two approaches, we do not expect them to perform in an absolutely different way.

Our analysis is based on the results of six data files of different sizes. The results on these six data files are consistent. We have observed the effects of the parameters used in microaggregation in the risk analysis. We have seen that the larger the k, the smaller the risk. This is an expected result, as cells with small number of contributors are sensitive.

The analysis also shows that the larger the number of attributes, the largest the risk in terms of number of cell that are sensitive. This is a natural consequence of the curse of dimensionality in clustering. It is known that the larger the number of variables the larger the distance between pairs of points. This naturally implies that for any variable we will find in a cluster an element that is far away from the others. This implies that the contribution to the mean of this cluster will be an important percentage of the total.

We consider that users and implementers of k-anonymity need to be aware of this type of risk. We think that it is also useful to give names for k-anonymity when attribute disclosure in the above sense does not take place. We call them dominance rule compliant k-anonymity, and p%-rule compliant k-anonymity.

This work has open some research directions. In particular, about providing ways to achieve k-anonymity so that attribute

disclosure does not takes place. We consider the following approaches to avoid this risk.

- Use larger k. We have shown that, in general, the larger the k the lower the risk. A k equal to 5 seems to be too low to avoid this type of risk. In this case, for a large number (and a large proportion) of cells, there are dominant contributors. Note that $k = 5$ is one of the recommended values. A rule of thumb is to use $k = 8$ or $k = 10$, which seems to be much safer. This would provide safer data but unless the number of sensitive cells is zero, some attribute disclosure risk will still be present in the published file.
- Remove sensitive cells. In tabular data, suppression is one of the standard approaches for sensitive cells. This approach can be useful here. Suppressed cells also leak information (i.e., that there was a contributor with a significant share). Therefore, secondary suppression may be required to avoid this type of attacks and the corresponding disclosure. Tabular data also considers secondary suppression. In our case, this approach would correspond to suppress a safe cell for each unsafe one. In this way, we would provide dominance rule compliant or $p\%$-rule compliant k-anonymity.
- Dynamic k-anonymity. Compute the risk of each cell when clusters are formed, and increase the size of the cluster to reduce the risk, or remove records that cause the cell to be marked as sensitive.
- Revise the formal definition of microaggregation. Microaggregation is formalized as an optimization problem with constraints. The objective function is that clusters are as similar as possible as the cluster center. The constraints include that the number of records in each cluster is within k and $2k$. This problem can be reformulated to take into account attribute disclosure risk, but also that clusters need to have some pre-established diversity.

The last two items are directions for future research.

Acknowledgements This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The Second author acknowledges the support of the Spanish Ministry of Science and Innovation (project PID2021-125962OB-C33).

Author Contributions VT and GN wrote the main manuscript. All authors reviewed the manuscript.

Funding Open access funding provided by Umeå University.

Data availability The data sets used for the experiments are publicly available. References are included in the paper. Code will be made available if the paper is accepted.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article. The authors declare no competing interests.

Human or animal rights Research has neither involved human participants nor animals.

References

1. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., de Wolf, P.-P.: Statistical Disclosure Control. Wiley, Hoboken (2012)
2. Torra, V.: Guide to Data Privacy: Models, Technologies, Solutions. Springer, Berlin (2022)
3. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans. Knowl. Data Eng. **14**(1), 189–201 (2002)
4. Capitani, D., di Vimercati, S., Foresti, S., Livraga, G., Samarati, P.: Data privacy: definitions and techniques. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **20**, 793–818 (2012)
5. Dwork, C.: Differential privacy. In: Proc. ICALP 2006, LNCS, pp. 1–12 (2006)
6. Dwork, C.: Differential privacy: a survey of results. In: Proc. TAMC 2008, LNCS, pp. 1–19 (2008)
7. Jaro, M.A.: Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. J. Am. Stat. Assoc. **84**(406), 414–420 (1989)
8. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**(6), 1010–1027 (2001)
9. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. SRI Intl. Tech, Rep (1998)
10. di Vimercati, S.D., Foresti, S., Livraga, G., Samarati, P.: k-anonymity: from theory to applications. Trans. Data Privacy **16**, 25–49 (2023)
11. Truta, T.M., Campan, A., Sun, X.: An overview of P-sensitive k-anonymity models for microdata anonymization. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **20**(6), 819–838 (2012)
12. Truta, T. M., Vinay, B.: Privacy protection: p-sensitive k-anonymity property. In: Proc. 2nd Int. Workshop on Privacy Data management (PDM 2006), p. 94 (2006)
13. Machanavajjhala, A., Gehrke, J., Kiefer, D., Venkitasubramanian, M.: L-diversity: privacy beyond k-anonymity. In: Proc. of the IEEE ICDE (2006)
14. Castro, J.: Minimum-distance controlled perturbation methods for large-scale tabular data protection. Eur. J. Oper. Res. **171**, 39–52 (2006)
15. Castro, J.: On assessing the disclosure risk of controlled adjustment methods for statistical tabular data. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **20**, 921–942 (2012)
16. Duncan, G.T., Elliot, M., Salazar, J.J.: Statistical Confidentiality. Springer, Berlin (2011)

17. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Min. Knowl. Discov. **11**(2), 195–212 (2005)
18. Laszlo, M., Mukherjee, S.: Minimum spanning tree partitioning algorithm for microaggregation. IEEE Trans. Knowl. Data Eng. **17**(7), 902–911 (2005)
19. Oganian, A., Domingo-Ferrer, J.: On the complexity of optimal microaggregation for statistical disclosure control, statistical. J. United Nations Econ. Comm. Eur. **18**(4), 345–354 (2000)
20. Templ, M.: Statistical disclosure control for microdata using the R-package sdcMicro. Trans. Data Privacy **1**(2), 67–85 (2008)
21. LeFevre, K., DeWitt, D. J., Ramakrishnan, R.: Multidimensional k-anonymity. Technical Report 1521, University of Wisconsin (2005)
22. Rebollo-Monedero, D., Forné, J., Soriano, M., Puiggalí Allepuz, J.: k-Anonymous microaggregation with preservation of statistical dependence. Inf. Sci. **342**, 1–23 (2016)
23. Bache, K., Lichman, M.: UCI machine learning repository. University of California, School of Information and Computer Science, Irvine, CA (2013)
24. Clifton, C., Tassa, T.: On syntactic anonymity and differential privacy. In: Proc, ICDE Workshops (2013)
25. LeFevre, K., DeWitt, D. J., Ramakrishnan, R.: Incognito: efficient full-domain K-anonymity. In: SIGMOD (2005)
26. Li, N., Li, T., Venkatasubramanian, S.: T-closeness: privacy beyond k-anonymity and l-diversity. In: Proc. of the IEEE ICDE (2007)
27. http://www.mdai.cat/code
28. Torra, V.: Microaggregation for categorical variables: a median based approach. In: Proc. Privacy in Statistical Databases (PSD 2004), Lecture Notes in Computer Science, pp D162–174 (2004)