Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200149

Efficient Protection of Health Data from Sensitive Attribute Disclosure

Raffael BILD [a,1], Johanna EICHER [a] and Fabian PRASSER [b,c]
[a] *University hospital rechts der Isar, Technical University of Munich, Germany*
[b] *Charité - Universitätsmedizin Berlin, Berlin, Germany*
[c] *Berlin Institute of Health (BIH), Berlin, Germany*

Abstract. Biomedical research has become data-driven. To create the required big datasets, health data needs to be shared or reused out of the context of its initial purpose. This leads to significant privacy challenges. Data anonymization is an important protection method where data is transformed such that privacy guarantees can be provided according to formal models. For applications in practice, anonymization methods need to be integrated into scalable and robust tools. In this work, we focus on the problem of scalability.

Protecting biomedical data from inference attacks is challenging, in particular for numeric data. An important privacy model in this context is t-closeness, which has also been defined for attribute values which are totally ordered. However, directly implementing a scalable algorithmic representation of the mathematical definition of the model proves difficult. In this paper we therefore present a series of optimizations that can be used to achieve efficiency in production use. An experimental evaluation shows that our approach reduces execution times of anonymization processes involving t-closeness by up to a factor of two.

Keywords. data protection, anonymization, inference attacks, scalability

1. Introduction

Biomedical research, e.g. in the field of precision medicine which tailors healthcare to characteristics of individuals, is increasingly data-driven and leveraging methods from data science such as machine learning [1]. However, when creating the required big datasets, stringent privacy protection is mandated by laws and regulations. Hence, a wide range of safeguards has to be applied, including organizational and technical measures.

Data anonymization is an important building block for implementing technical privacy protection. The basic idea is to transform data in such a manner that formal guarantees, e.g. regarding the risk of singling out, linkage or inference, can be provided [2]. These formal guarantees are captured by so called *privacy models*. t-Closeness is a state-of-the-art model for protecting data from inference attacks. The model requires that the distribution of sensitive attribute values in a set of indistinguishable data records is not too different from the distribution of sensitive information in the overall dataset [3].

[1]Corresponding Author: Raffael Bild, Institute of Medical Informatics, Statistics and Epidemiology, University Hospital rechts der Isar, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany; E-mail: raffael.bild@tum.de.

2. Objective

The ARX Data Anonymization Tool is among the few software solutions for quantitative data anonymization, that have found wide-spread adoption. ARX focuses on data transformation methods which have been specifically recommended for applications to health data and it implements models for protecting data from singling out, linkage and inference [4]. t-Closeness is amongst the models supported.

t-Closeness has been specified in different variants that apply to variables with different scales of measure. One of these variants focuses on variables which are totally ordered. This model is particularly relevant in practice, as it is one of the few privacy models which have been proposed for protecting sensitive numeric variables.

When using ARX to protect complex datasets using the t-closeness model, however, we realized that the initial implementation is not scalable. Upon further inspection, we realized that directly implementing a scalable algorithmic representation of the mathematical definition of the model t-closeness proves difficult in general. In this paper we therefore present a series of optimizations that we have developed to achieve efficiency in productive use. All of them have been integrated into ARX.

3. Methods

3.1. Problem Definition

t-Closeness is a condition that applies to equivalence classes, i.e. groups of records which are indistinguishable regarding attributes that could be used for linking records. Let $P(e) = (p_1, p_2, ..., p_m)$ be the relative frequency distribution of sensitive values in a given equivalence class e and let $Q = (q_1, q_2, ..., q_m)$ be the relative frequency distribution of sensitive values in the whole dataset. $D[P(e), Q]$ is the distance between the distributions $P(e)$ and Q [3]. It is defined as follows [3]:

$$D[P(e), Q] = \frac{1}{m-1} \sum_{i=1}^{m} \left| \sum_{j=1}^{i} (p_j - q_j) \right|.$$

A dataset fulfills t-closeness with numerical ground distance if for all equivalence classes e, $D[P(e), Q] \leq t$ holds.

	Age	Sex	LoS	AdmQrtr	Charge	
Class e1	[25, 50[	Male	[1, 5[	3	50.000	Sensitive attribute
	[25, 50[	Male	[1, 5[	3	60.000	
Class e2	[50, 75[	Female	[5, 10[	1	60.000	
	[50, 75[	Female	[5, 10[	1	60.000	
	[50, 75[	Female	[5, 10[	1	70.000	

Figure 1. Example discharge dataset. "LoS" = Length of stay, "AdmQrtr" = Admission quarter.

Figure 1 shows an example dataset with a sensitive attribute "Charge" and two equivalence classes $e1$ and $e2$ defined by the values of the other attributes. The distribution of sensitive values is $Q = (\frac{1}{5}, \frac{3}{5}, \frac{1}{5})$ since the values 50.000, 60.000 and 70.000 appear 1, 3 and 1 times in the whole dataset, respectively. For the equivalence classes, we get $P(e1) = (\frac{1}{2}, \frac{1}{2}, 0)$ and $P(e2) = (0, \frac{2}{3}, \frac{1}{3})$. Consequently, we have

$$D[P(e1),Q] = \frac{1}{2}(|\frac{1}{2}-\frac{1}{5}|+|\frac{1}{2}-\frac{1}{5}+\frac{1}{2}-\frac{3}{5}|+|\frac{1}{2}-\frac{1}{5}+\frac{1}{2}-\frac{3}{5}-\frac{1}{5}|) = \frac{1}{4},$$
$$D[P(e2),Q] = \frac{1}{2}(|-\frac{1}{5}|+|-\frac{1}{5}+\frac{2}{3}-\frac{3}{5}|+|-\frac{1}{5}+\frac{2}{3}-\frac{3}{5}+\frac{1}{3}-\frac{1}{5}|) = \frac{1}{6}.$$

Hence, we can conclude that the dataset satisfies t-closeness with $t = max\{\frac{1}{4},\frac{1}{6}\} = 0.25$.

A straight-forward implementation of t-closeness for fully ordered attributes would implement this by checking if the following inequality holds:

$$|r_1|+|r_1+r_2|+|r_1+r_2+r_3|+...|r_1+...+r_{m-1}| \leq t(m-1)$$

Each r_i has the form $r_i = p_i - q_i$, where p_i is the frequency of the attribute value number i in the currently considered equivalence class e of the transformed data set, and q_i the frequency of the attribute value number i in the entire input dataset.

As we will show in Section 4 this process is highly inefficient. The main reason is that it needs to iterate over all sensitive attribute values contained in the overall dataset. Given that this process needs to be executed for all equivalence classes, the worst-case complexity is $O(n^2)$ where n is the number of records in the dataset.

3.2. Optimization Approaches

In this section, we present three optimizations that we used to improve our initial, straight-forward implementation of the model.

Optimization 1 – Fibonacci hashing: The first optimization adresses the implementation level. One of the most time-consuming aspects of implementing a check for t-closeness is to dynamically group the sensitive attribute values in each class to determine their frequency. The standard data structure used for this purpose are hash tables. ARX already used an efficient implementation provided by the High Performance Primitive Collections for Java library [5]. However, these collections are still much more complex than required, as they for example support updating the data stored in a map. We therefore implemented a simplified hash table using Fibonacci hashing based on the golden ratio to reduce the number of CPU cycles required for adding and querying elements.

Optimization 2 – Check pruning: The second optimization addresses the mathematical definition of the model. Let us assume that the attribute values number $1...k$ in the equivalence class currently under consideration do not occur at all, then $p_1 = p_2 = ... = p_k = 0$ holds and the first k summands in the above condition have the form:

$$|-q_1|+|-q_1-q_2|+...+|-q_1-...-q_k|.$$

This partial sum depends only on the input dataset and can therefore be precalculated for every possible value of k before the individual equivalence classes of the transformed dataset are checked.

These precalculations can be performed in an initialization step, for ascending values of k until the corresponding subtotal is greater than the threshold, i.e. the following holds:

$$|-q_1|+|-q_1-q_2|+...+|-q_1-...-q_k| > t(m-1).$$

Let us denote the smallest value of k for which this inequality is fulfilled with x.

When checking whether t-closeness holds for a specific equivalence class, as a first step, we calculate the smallest index of any attribute value occurring in the class. We call this index y. When $y > x$ it can be inferred that the following summands are included in the relevant sum:

$$|-q_1| + |-q_1-q_2| + ... + |-q_1-...-q_x|.$$

It follows that the threshold $t(m-1)$ will definitely be exceeded and computations can already be stopped at this point (concluding that privacy guarantees are not fulfilled).

Optimization 3 – Summand pruning: The third optimization adds an additional pruning mechanism using pre-computations.

It can be used in cases where the pruning strategy described above is not applicable, i.e. if $y \leq x$ holds. It works by starting the summation at position y, using an appropriate sum which has been pre-calcuated ahead of time for all $k \leq x$ as a starting point:

$$|-q_1| + |-q_1-q_2| + ... + |-q_1-...-q_k|.$$

3.3. Experimental Design

To evaluate our approach, we used the following data from registries and health surveys from the U.S.: 100,937 records about traffic accidents from the NHTSA Fatality Analysis Reporting System (FARS), 539,253 records from the American Time Use Survey (ATUS) and 1,193,504 records from the Integrated Health Interview Series (IHIS). Moreover, we analyzed a subset of a synthetic discharge dataset which is particularly hard to protect from sensitive attribute disclosure (SPD) [6]. We also included two de-facto standard datasets for benchmarking anonymization methods: 30,162 records from the 1994 U.S. Census (ADULT) and 63,441 records from the 1998 KDD competition (CUP). For a detailed specification of the datasets we refer to [7].

We anonymized the datasets with attribute generalization and record suppression to produce output datasets which fulfill *t-closeness* for fully ordered attributes. We varied the risk threshold t ($0.5 \geq t \geq 0.1$) to study the effect of our optimizations on different parameterizations. As a baseline, we used our original, unoptimized implementation. In the software, both pruning strategies are combined into a common implementation and we therefore present one measurement capturing both of them.

4. Results

Figure 2 shows the results of our experiments. As can be seen, our optimizations improved execution times by up two a factor of more than two. Each optimization had a positive effect in all setups while the degree of effectiveness of each optimization varied between setups. Pruning was possible in between 69% and 99% of the checks performed, but the effect on execution times varied.

The differences in the impact of optimizations on execution times can be explained by considering the distribution of sensitive attributes values in the different datasets. The impact was lower for datasets with a small number of distinct values, i.e. ADULT (7), ATUS (7) and IHIS (10), higher for datasets with more distinct sensitive values, i.e. CUP (81) and FARS (20). We measured the strongest effect for the SPD dataset, which has 101 different sensitive attribute values. When the number of sensitive attribute values is high, execution times are also higher, implying the our optimizations are more effective exactly in the cases where they are needed the most.

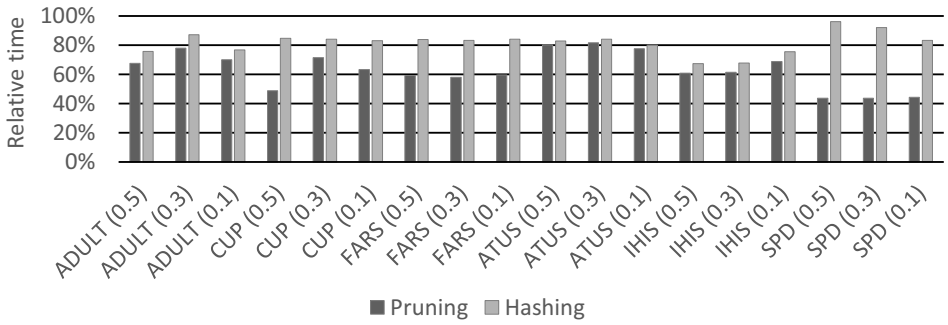

Figure 2. Execution times relative to the original implementation for various datasets and risk thresholds.

5. Conclusion

Due to its usability, flexibility and scalability, ARX is actively used in many areas, including commercial big data analytics platforms, medical research projects, clinical trial data sharing and for training purposes. An important reason for ARX's scalability are the many optimizations that have been integrated into the software. In prior work we have, for example, presented methods to improve the scalability of optimization algorithms for trading off risks vs. utility [8] and a versatile optimized runtime environment for anonymization algorithms [9].

In this paper, we have presented an optimization affecting a specific and important privacy model only. Our approach addresses the implementation level as well as the mathematical definition of the model implemented. Our solution utilizes high-performance data structures as well as pre-computation techniques. The model is particularly relevant in practice, as it is one of the few approaches which can be used to protect sensitive numeric data.

References

[1] V. Gligorijević et al., Integrative methods for analyzing big data in precision medicine, *Proteomics* **16** (2016), 741–758.

[2] B. C. M. Fung, K. Wang, A. W.-C. Fu and P. S. Yu, *Introduction to privacy-preserving data publishing: Concepts and techniques*, CRC Press, 1st ed., 2010, ISBN 9781420091489.

[3] N. Li, T. Li and S. Venkatasubramanian, t-closeness: Privacy beyond k-anonymity and l-diversity, *2007 IEEE 23rd International Conference on Data Engineering*, IEEE, 2007, 106–115.

[4] F. Prasser and F. Kohlmayer, Putting statistical disclosure control into practice: The ARX data anonymization tool, *Medical data privacy handbook*, Springer, 2015, 111–148.

[5] S. Osinski and D. Weiss, HPPC: High Performance Primitive Collections for Java, `https://labs.carrotsearch.com/hppc.html`, accessed: October 10, 2019.

[6] D. Sánchez, S. Martínez and J. Domingo-Ferrer, Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata", *Science* **351** (2016), 1274–1274.

[7] F. Prasser, F. Kohlmayer and K. A. Kuhn, A benchmark of globally-optimal anonymization methods for biomedical data, *2014 IEEE 27th International Symposium on Computer-Based Medical Systems*, IEEE, 2014, 66–71.

[8] F. Prasser, F. Kohlmayer and K. A. Kuhn, Efficient and effective pruning strategies for health data de-identification, *BMC medical informatics and decision making* **16** (2016), 49.

[9] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper and K. A. Kuhn, Highly efficient optimal k-anonymity for biomedical datasets, *2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, 2012, 1–6.