An Algorithm for *k*-anonymity-based Fingerprinting

Sebastian Schrittwieser[1], Peter Kieseberg[1], Isao Echizen[2], Sven Wohlgemuth[2],
Noboru Sonehara[2], and Edgar Weippl[1]

[1] SBA-Research, Austria
sschrittwieser,pkieseberg,eweippl@sba-research.org
[2] National Institute of Informatics, Japan
iechizen,wohlgemuth,sonehara@nii.ac.jp

Abstract. The anonymization of sensitive microdata (e.g. medical health records) is a widely-studied topic in the research community. A still unsolved problem is the limited informative value of anonymized microdata that often rules out further processing (e.g. statistical analysis). Thus, a tradeoff between anonymity and data precision has to be made, resulting in the release of partially anonymized microdata sets that still can contain sensitive information and have to be protected against unrestricted disclosure. Anonymization is often driven by the concept of k-anonymity that allows fine-grained control of the anonymization level. In this paper, we present an algorithm for creating unique fingerprints of microdata sets that were partially anonymized with k-anonymity techniques. We show that it is possible to create different versions of partially anonymized microdata sets that share very similar levels of anonymity and data precision, but still can be uniquely identified by a robust fingerprint that is based on the anonymization process.

Keywords: k-anonymity, fingerprinting, generalization, algorithm

1 Introduction

When releasing microdata containing sensitive information such as medical health data for research purposes, a tradeoff between data privacy and data quality has to be made. On the one hand, completely anonymized data records are often too generalized to be useful for further processing (e.g. statistical analysis), on the other hand, anonymization is desirable and often even demanded to achieve regulatory compliance such as Directive 95/46/EC in the European Union or the Health Insurance Portability and Accountability Act (HIPAA), which regulates the processing of medical health data in the United States. k-anonymity [8] is a technique that allows defining a fine-grained level of anonymity for microdata. Although partially anonymized data is usually protected by special usage restrictions, both technical and organizational misuse (e.g. unauthorized disclosure) can not be eliminated. Fingerprinting is a passive form of security, meaning that it can help to identify the source of disclosure after it took place. In this paper, we take our idea of using k-anonymity techniques for fingerprinting partially anonymize microdata presented in [6] and introduce an algorithm for the generation of a batch of partially anonymized microdata sets that share the same levels of anonymity and similar levels of data precision while being uniquely identifiable by a robust fingerprint. A typical application scenario for our approach is the release of partially anonymized microdata sets to multiple receivers (e.g. research institutions, universities, etc.). Instead of delivering identical copies, for each receiver the microdata is anonymized in a slightly different way. The anonymization process inherently generates a fingerprint that is contained in every single record of a given set and unique for each set. Thus, in the case of unauthorized data disclosure, it is possible to identify the source of the leak based on the fingerprint.

The rest of the paper is structured as follows. Section 2 summarizes the concept of k-anonymity and our idea of fingerprinting partially anonymized microdata presented in [6]. Our proposed algorithm is explained in Section 3.1. In Section 4, we evaluate security aspects, including the problem of inference attacks, and finally, we conclude in Section 5.

2 Related Work

2.1 *k*-anonymity

Sweeney [5,8] showed that even after removing uniquely identifying attributes (e.g., name or social security number) from medical health data, people can still be identified by so-called quasi-identifiers (QI), i.e. attributes such as ZIP code, birthdate, and sex that can be linked in order to break anonymization. Her introduced concept of k-anonymity is a widely adopted anonymization technique. The k-anonymity criterion is satisfied, if each record is indistinguishable from at least *k-1* other records with respect to the quasi-identifiers. Hence, quasi-identifying attributes must have the same values within an equivalence class, so that it is impossible to uniquely link a person to a specific record within the class. By raising the value of k, high levels of anonymity can be achieved with this anonymization technique, however, to maintain significance of the data, often lower anonymization levels need to be chosen.

Over the past years, several improvements to the idea of k-anonymity were proposed. ℓ-diversity tries to enhance anonymity in cases where little diversity in sensitive attributes occurs by requiring that each equivalence class has at least ℓ well-represented values for each sensitive attribute [3]. Another improvement to k-anonymity is t-closeness that requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the original table [2]. Both, ℓ-diversity and t-closeness are additional, tightened criteria to k-anonymity, but do not replace its original idea. Our proposed algorithm is explained using the original k-anonymity criterion. However, an adaption to ℓ-diversity and t-closeness would not require any modifications of the algorithm.

Generalization is the main strategy for achieving k-anonymity. Hereby, data granularity is reduced to unify the data records. Figure 1 shows possible generalizations for two different quasi-identifiers. The characters a and b define the two quasi-identifiers *sex* and *birthdate* that have different levels of generalization (described by the numbers 0 to 2).

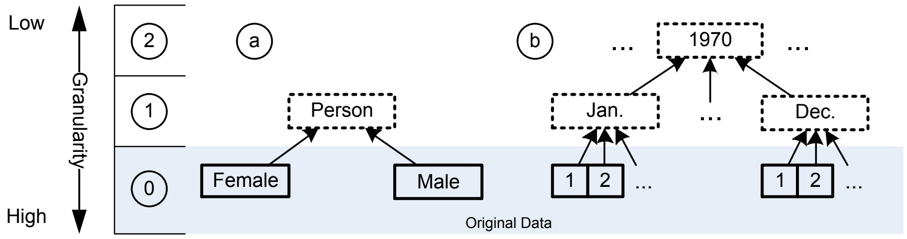


Fig. 1. Generalization strategies for two quasi-identifiers.

2.2 Data precision metrics

Generalization almost always results in the loss of information. To be able to judge the quality of a resulting generalized data set, metrics that make this loss of information measurable need to be devised. We will use this definition of quality for generating the classes of equivalent generalization strategies, since we target at providing all recipients with almost the same data quality.

Samarati Metric The Samarati Metric [4] is defined by simply adding the generalization levels of the quasi-identifiers. This approach has the positive effect that it is very intuitive and easy to calculate, but has a serious drawback: This metric only counts the generalization steps without taking the total number of possible generalizations for an identifier into account.The generalization from {female, male} to {person} results in a much higher information loss than generalizing a date (e.g. birthdate) from a timestamp-granularity to the day-granularity. In the Samarati Metric both operations may yield the same information loss.

Precision Metric[7] contrary to the Samarati Metric, not every generalization possesses the same weight, but this is calculated with respect to the maximum generalization depth possible for a quasi-identifier. In case an identifier possesses more than one level of generalization, the weight of generalizing is defined by

$$\frac{\text{Number of levels generalized}}{\text{Number of possible generalization levels}}$$

Example: For the quasi-identifier *birthdate* (original granularity: timestamp), three levels of generality are defined: {day, month, year}. Thus generalizing birthdate from *timestamp*-level to *day*-level results in a weight of $\frac{1}{3}$. Contrary, generalizing the quasi-identifier *sex* from {female, male} to {person} still results in a weight of 1. One drawback of the Precision Metric lies in the fact that it is completely unrelated to the actual data (a drawback that is valid for the Samarati Metric as well).

*Modified Discernability Metric DM** This is a slightly modified version of the Discernability Metric and was proposed by El Emam et al. in [1]: N denotes the number of classes and n_i the number of elements of the i-th class. Then the Modified Discernability Metric DM* is defined by $DM^* = \sum_{i=1}^{N} n_i^2$.

The big advantage of this metric lies in the fact that the sizes of the resulting classes are incorporated into the data quality.

Our research has shown that in general a perfect metric does not exist and that the practicability of the results heavily depends on the source data, the types of attributes and the generalization patterns. In real life scenarios, the choice for a specific data precision metric has to be based on these parameters.

2.3 Algorithm for finding the optimal solution

All possible generalization strategies can be depicted by the lattice diagram shown in Figure 2. A node in the diagram is generated by generalizing one quasi-identifier by one level, i.e. the diagram shows all generalizations that can be derived by generalizing by one level at once and their relations. For example, a_1 refers to the second generalization step of the quasi-identifier a.

El Emam et. al. [1] proposed an algorithm for calculating the optimal generalization strategy (i.e. the generalization with the highest data precision) with respect to a given metric. The algorithm finds all k-anonymous nodes for each generalization-strategy. The algorithm uses a technique called *predictive tagging* for reducing the workload by being able to tag nodes in the lattice diagram without directly calculating their level of anonymity: Since a metric fulfills the axiom of monotony, the two statements hold true:

- If a node in the diagram is k-anonymous, then all nodes above it in the diagram are at least k-anonymous too, i.e. by further generalizing a k-anonymous classification, the k-anonymity is not lost.
- If a node in the diagram is not k-anonymous, then all nodes below it in the diagram can not be k-anonymous.

When all generalization-strategies are evaluated, the globally lowest node is chosen.

2.4 Fingerprinting

The term watermarking defines techniques that add visible or hidden information (e.g. a copyright notice) to the target data. The important aspect in this

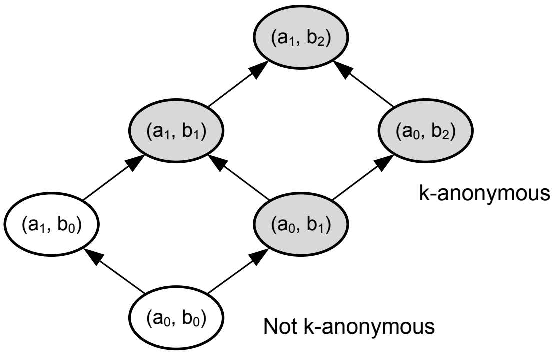


Fig. 2. Generalization strategies for two quasi-identifiers.

definition is that adding a watermark modifies the target data. This modification can either be visible (e.g. a text overlay in an image file) or invisible (e.g. by implementing steganographic techniques) to the user. In both cases, the watermark information is additional data that is combined with the target data. In contrast to watermarking, the definition of fingerprinting is not consistent among the research community. There exist at least two definitions. The first one describes fingerprinting as a subtype of watermarking where a unique watermark (i.e. the fingerprint) is added to each copy of the target data. The second definition distinguishes fingerprinting from watermarking by the source of the fingerprint. While in watermarking, information is actively added, fingerprinting uses intrinsic properties of the data to uniquely differentiate the copies of the data. In both definitions, however, the uniqueness of the fingerprint is the key concept that enables a data owner to uniquely link a data customer to a specific file. Our schema is based on the idea of extracting unique fingerprints from the data structure, thus follows the second definition.

In the past, the concept of fingerprinting microdata was discussed by Willenborg and de Waal [9] and Willenborg and Kardaun [10]. In both approaches, fingerprints are built from combinations of identifying variables in the microdata records and are used for identifying specific records in a set of microdata.

2.5 Extracting Unique Fingerprints from Generalization Patterns

This subsection summarizes our previous work on devising a fingerprinting technique based on k-anonymity [6]. This technique aims at enabling the identification of a source microdata set based on the analysis of a single record. Thus, the fingerprint is the same for every record in a given set and unique for each set. The scenario for our method is the release of sensitive microdata (e.g. medical health records) to multiple receivers (e.g. universities). The idea is that each receiver gets a differently anonymized data set that can be uniquely identified by a single record based on this fingerprint.

When sensitive microdata sets are anonymized using k-anonymity techniques, the choice of generalization levels for the quasi-identifiers influences the value of k and the quality of the anonymized data. In general, data quality decreases and k increases with higher generalization levels as generalization removes information expressiveness of the data and more records look the same with respect to the quasi-identifiers, resulting in larger equivalence classes.

The fingerprints in our approach are formed by the generalization levels of quasi-identifiers used for anonymization. That generalization levels are chosen differently for each released microdata set in a way, that the data quality values of the datasets are all roughly the same. Our proposed algorithm identifies similar anonymization strategies; strategies that generate anonymized datasets that meet a specified level of k and share similar data quality. Figure 3 explains our approach of generating anonymized data sets with different generalization strategies. All candidate data sets above the defined threshold of k are compared and clustered with respect to data quality. Data sets from one cluster are then released to different data receivers. The generalization strategy used for a dataset can be extracted from every single record, thus allowing the identification of the source data set.

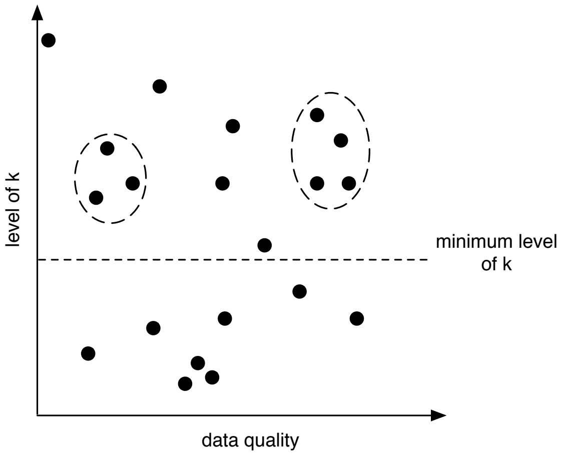


Fig. 3. Clustering similar anonymization strategies

Table 1. Original data and two anonymized sets ($k = 2$)

Original data				First Set			Second Set		
name	sex	birthdate	disease	sex	birthdate	disease	sex	birthdate	disease
Bob	m	19.03.1970	chest pain	F	1970	chest pain	P	03.1970	chest pain
Dave	m	20.03.1970	short breath	M	1970	short breath	P	03.1970	short breath
Alice	f	18.04.1970	obesity	F	1970	obesity	P	04.1970	obesity
Eve	f	21.04.1970	short breath	M	1970	short breath	P	04.1970	short breath

3 Approach

In this section we propose an algorithm for the generation of unique fingerprints for microdata sets and an algorithm for the identification of the source of a data leak.

3.1 Algorithm for Generating Fingerprints

Our approach is based on El Emam's algorithm for calculating the optimal solution (see Section 2.3). Since we need all k-anonymous solutions instead of just the optimal one (i.e. the one with minimum information loss), we cannot rule out the more general solutions of a solution found, but still have to calculate the data precision for each of these nodes.

1. Define a minimum k for the k-anonymity criterion, the minimum and maximum levels of data loss l_{min} and l_{max} and the data precision metric to be used.
2. Define the generalization strategies for each identifier.
3. Calculate the lattice diagram derived from all possible generalizations.
4. Choose a node at middle height and decide whether it is at least k-anonymous.
   (a) In case it is not, rule out all nodes below in the lattice diagram.
   (b) In case it is, mark all nodes above the chosen one as possible solutions.
5. Start with step four for the remaining subgraph, similar to the original algorithm.
6. In case no subgraph is left, start by choosing another initial node at middle height and proceed with step four until all nodes are evaluated.
7. For each at least k-anonymous solution, calculate data precision and the actual k. Remove all solutions with data precision outside the bounds of l_{min} and l_{max}.
8. Classify and cluster the solutions by their data precision.
9. Create "similar" microdata sets based on results in one cluster and distribute them to the recipients.

Following we give further details on some of the steps:

Data precision metrics and maximum levels (Step 1) All recipients shall be provided with (roughly) the same level of data precision. Thus, the method of measuring data precision can have a great impact on the definition of the classes of equivalent anonymization strategies (with respect to data quality). Table 2 gives an overview on the perceived data precision of the data from Table 1 based on all identified generalization techniques (see Figures 1 and 2), with respect to the three metrics discussed in Section 2.2. Additionally, it must be decided beforehand, how big the maximum tolerable data loss is.

Eliminating nodes (Step 4) Contrary to the algorithm for finding the optimal solution as proposed in 2.3, the nodes above the chosen one cannot be ruled out in case of 4.b (i.e. the node represents a k-anonymous generalization), since we need all solutions, not only the optimal one.

Table 2. Impact of different metrics

Node	Sam	Prec	DM*	k
(a_0, b_0)	0	0	4	1
(a_1, b_0)	1	1	4	1
(a_0, b_1)	1	0.5	8	2
(a_0, b_2)	2	1	8	2
(a_1, b_1)	2	1.5	8	2
(a_1, b_2)	3	2	16	4

Clustering the solutions (Step 8) In step eight the solutions are clustered by data loss. As discussed in Section 2.2, the data precision metric can have a great impact on the measured data loss and thus on the resulting clustering. Some metrics lead to finer grained results, thus no real clustering can be achieved by using the equality-function (see the case using the precision metric in Table 2). Here, a reasonable form of discretization (e.g. rounding or defining intervals) needs to be applied in order to generate usable classes.

Final distribution (Step 9) Before distribution, a list containing all assignments from generalization patterns to recipients is generated. We will later refer to this list as the *pattern-list.*

In section 3.3 we provide a detailed step-by-step-walkthrough of this algorithm.

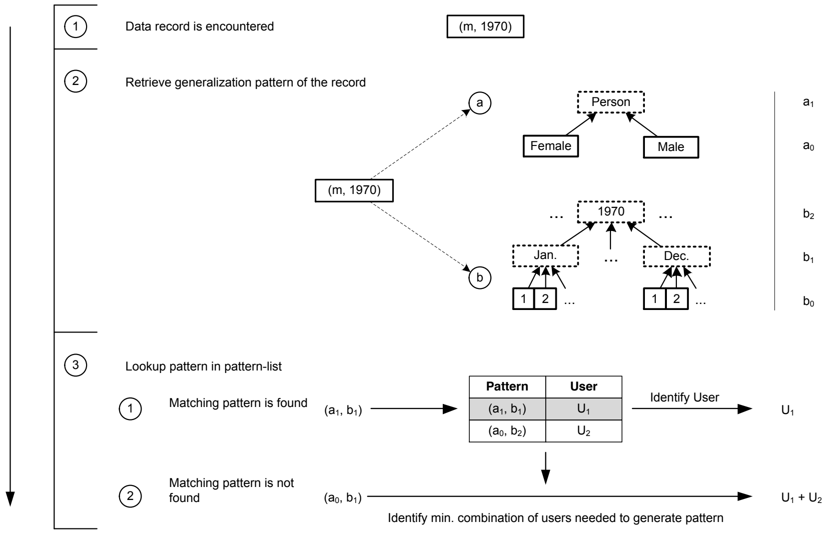
3.2 Principle of the identification of data leaks

In this section we introduce the algorithm for detecting data leaks (see Figure 4 for an illustration). The scenario is based on a data holder encountering data samples in the wild that he gave to other organizations in an anonymized (and fingerprinted) form.

1. The original data holder encounters some leaked data samples.
2. The underlying generalization pattern of this data is extracted.
3. The pattern is compared to the pattern list.
   (a) In case the pattern directly matches the pattern of the data given to a receiver, this receiver is identified as the source of disclosure.
   (b) Otherwise, calculate the minimum of receivers that possess the knowledge (i.e. data) to be able to generate the encountered data set structure.

Figure 4 explains the approach.

Example Original Data D_0 is anonymized using two different patterns, (a_1, b_1) and (a_0, b_2) thus generating the anonymized data sets D_1 and D_2 respectively. Recipient R_1 is provided with D_1 and R_2 with D_2. If, for example, a data record of the form (a_0, b_2) is disclosed (i.e. in our examples this would be a data


1
Data record is encountered
(m, 1970)
2
Retrieve generalization pattern of the record
(m, 1970)
a
b
Person
Female
Male
1970
Jan.
Dec.
1
2
...
a_1
a_0
b_2
b_1
b_0
3
Lookup pattern in pattern-list
1
Matching pattern is found
(a_1, b_1)
Identify User
U_1
2
Matching pattern is not found
(a_0, b_1)
$U_1 + U_2$
Identify min. combination of users needed to generate pattern

Pattern	User
(a_1, b_1)	U_1
(a_0, b_2)	U_2

Fig. 4. Identifying the source of data leakage based on data-patterns.

record where "sex" is provided at original granularity and "birthdate" at year-granularity), it is easy to identify user U_2 as source of the leaked data, since user U_1 would not be able to provide this much detail on the QI "sex".

3.3 Step-by-step description

For this example, we use the microdata from Table 1. In step one we define the side constraints $k = 2$, $l_{min} \geq 1$, and $l_{max} \leq 2$ (minimal and maximal level of data loss) and choose the Samarati metric for measuring data precision. In step two and three we use the generalization strategy from Figure 1, yielding to the lattice diagram shown in Figure 5.

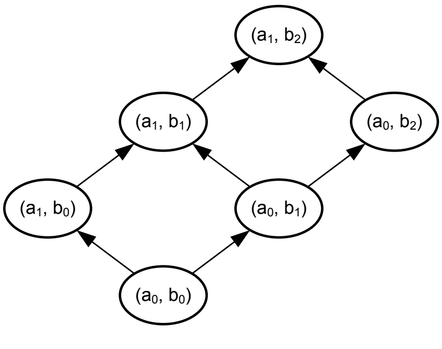


Fig. 5. Lattice diagram after step three.

Analogous to El Emam's algorithm, we then choose a node of middle height as starting node for step four, e.g. (a_1, b_0). This generalization does not fulfill the $2 - \textit{anonymity}$ criterion, thus we can rule out all nodes below (a_1, b_0) (in our example, this applies to one node only: (a_0, b_0)), because they would not fulfill the criterion as well (Step 4a). We now take the resulting subgraph of the chosen generalization path (Step 5) and apply Step 4 to node (a_1, b_1), which fulfills $k = 2$, and thus (by applying Step 4b) the node (a_1, b_2) fulfills the 2-anonymity criterion as well. This leads to the intermediate lattice diagram shown in Figure 6.

Table 3. Generalization (a_1, b_0).

Sex (a)	Birthdate (b)
p	19.03.1970
p	20.03.1970
p	18.04.1970
p	21.04.1970

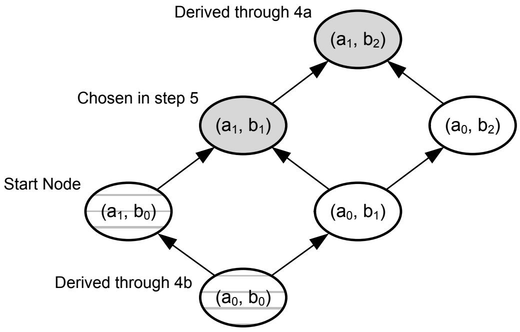


Fig. 6. Lattice diagram after the first generalization path.

As no subgraph is left, we arrive at Step 6 and again choose a starting node (a_0, b_1). It does provide 2-anonymity, so we can continue with Step 4b and and mark node (a_0, b_2) as a possible solution. The resulting lattice diagram is shown in Figure 7.

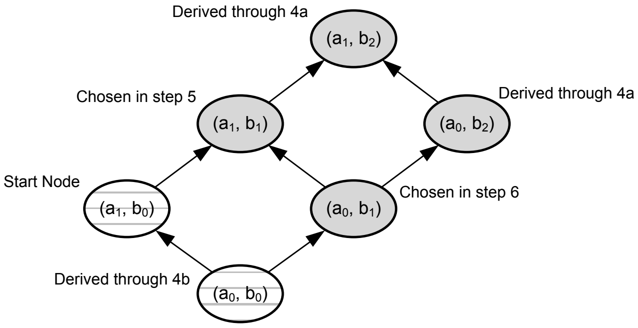


Fig. 7. Lattice diagram after the second generalization path.

Since we now have traversed through the entire lattice diagram, we can proceed with Step 7 and calculate the data precision and the actual k for each node that fulfills the 2-anonymous criterion (see Table 4).

Table 4. Data loss and k for solutions.

Node	Data Loss	k
(a_0, b_1)	1	2
(a_0, b_2)	2	2
(a_1, b_1)	2	2
(a_1, b_2)	3	4

We remove node (a_1, b_2) from the solution set, since the data precision is out of our defined bounds for the data loss l_{min} and l_{max}. The remaining candidate microdata sets are shown in Table 5.

Table 5. Generalization results.

(a_0, b_1)		(a_1, b_1)		(a_0, b_2)	
Sex	Birthdate	Sex	Birthdate	Sex	Birthdate
m	03.1970	p	03.1970	m	1970
m	03.1970	p	03.1970	m	1970
f	04.1970	p	04.1970	f	1970
f	04.1970	p	04.1970	f	1970

Finally, by clustering the remaining data sets by data precision, we derive *generalization clusters*. All solutions in such a cluster will be treated as being equivalent with respect to the data quality provided. In some cases (e.g. when the classes are too small), we have to use ranges instead of exact values as classification criteria. In our example we derived two clusters: $C_1 = \{(a_0, b_1)\}$ with data precision of 1 and $C_2 = \{(a_1, b_1), (a_0, b_2)\}$ with the data precision of 2.

In case the data needs to be sent to two receivers, we choose cluster C_2 and send data anonymized with strategy (a_1, b_1) to the first recipient and data anonymized with strategy (a_0, b_2) to the second one. Additionally, a list containing tuples of data receivers and generalization patterns used is stored by the provider. When data is disclosed, the provider can determine the generalization levels of the quasi-identifier attributes and then compare it to generalization the patterns stored. Thus, it is possible to identify the original owner of the data based on the unique generalization variant.

4 Evaluation

In this chapter we focus on the evaluation of the following aspects:

1. Number of possible fingerprints
2. Robustness of the fingerprints
3. Removing the fingerprint by utilization of complementary releases

Number of fingerprints When utilizing this approach for fingerprinting, it is very important that the owner of the original data is able to generate enough fingerprints, i.e. there must exist enough suitable anonymization patterns so that each recipient can be given a unique dataset. Unfortunately this number depends on various parameters like number of quasi-identifiers and generalization levels, the anonymization level k, the data precision metric in use, and last but not least the actual data itself. Therefore, an exact number cannot be given without analyzing the source data. Still an upper bound can be retrieved by plainly looking at the quasi-identifiers and their respective generalization strategies.

Be N the number of quasi-identifiers and n_i the number of generalization levels of the i-th QI. Then F is an upper bound for the number of different fingerprints with

$$F = \prod_{i=1}^{N} n_i$$

Robustness of a fingerprint In this paragraph we discuss how fingerprints devised by our method can be removed or changed in a way to avoid detection of the disclosing participant. Since the actual granularity of the data is used for fingerprinting, every removal must incorporate changing the generalization patterns. Since the attacker is not in the possession of finer grained data, the only reasonable way would lie in further generalization of one or more quasi-identifiers (actually the attacker could also "invent" finer granulations than he possesses, thus faking the data. Still, this would result in wrong, thus worthless data). This approach results in two major drawbacks:

1. In order to get undetectable, the disclosing party must be in the possession of the generalization strategy that was used for the other recipients, i.e. if U_1 received data with the pattern (a_1, b_2) and U_2 received (a_2, b_1), reducing U_2's granularity to (a_3, b_1) does not avoid U_2's detection.

2. Even if detection is avoided by the disclosing user, the data quality is reduced significantly. In our example, U_1 would at least need to generalize the data to the form (a_2, b_2) which is much more general than the data U_1 would be able to disclose.

Removing fingerprints through complementary releases Collusion attacks utilizing complementary releases can pose a severe threat to the anonymization of the data sets. Still, our fingerprint can lead to detection of the leaking parties, although it can not be guaranteed anymore.

The following examples shows detection of two collaborating leaking parties. The quasi-identifiers birthdate and sex are generalized like in the previous examples and zip code is granulated on two levels.

Table 6. Original data and three generalizations.

data set	name	sex	birthdate	zip code	disease
original data set	Bob	m	18.03.1970	1004	chest pain
	Dave	m	19.03.1970	1015	short breath
	Alice	f	20.04.1970	1004	obesity
	Eve	f	21.04.1970	1015	short breath
anonymized set 1	-	p	1970	1004	chest pain
	-	p	1970	1015	short breath
	-	p	1970	1004	obesity
	-	p	1970	1015	short breath
anonymized set 2	-	p	03.1970	100X	chest pain
	-	p	03.1970	101X	short breath
	-	p	04.1970	100X	obesity
	-	p	04.1970	101X	short breath
anonymized set 3	-	m	1970	100X	chest pain
	-	m	1970	101X	short breath
	-	f	1970	100X	obesity
	-	f	1970	101X	short breath

Revelation of the record {`03.1970, p, 1015, short breath`} reveals data sets one and two as sources: The month of birth could only be extracted from data set one, the zip code in this granulation only from data set two.

5 Conclusion

In this paper, we introduced an algorithm for fingerprinting partially anonymized microdata sets. The main idea is to take the different generalization patterns that can be used to achieve k-anonymity. Our approach is based on an algorithm for finding the optimal solution for the k-anonymity criterion and generates groups of microdata sets that share similar levels of anonymity and data precision.

In contrast to previous work, our fingerprinting method does not aim at identifying a single record from a database, but the identification of the source data set by just analyzing one record out of it. Every single record of the data set stores the same fingerprint that is unique and identifying for the data set it was take from. A typical application scenario for our approach is the release of microdata sets for research purposes to multiple receivers (e.g. universities). In the case of data disclosure, the source can be identified by even a single record.

Future work will focus on the threat of collusion attacks. We aim at constructing collusion-free data sets, i.e. data sets that do not allow to thwart the k-anonymity criterion by combining them.

References

1. K. El Emam, F. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, et al. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5):670, 2009.
2. N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.
3. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.
4. P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13:1010–1027, 2001.
5. P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. *Technical Report SRI-CSL-98-04, Computer Science Laboratory, SRI International*, 1998.
6. S. Schrittwieser, P. Kieseberg, I. Echizen, S. Wohlgemuth, and N. Sonehara. Using Generalization Patterns for Fingerprinting Sets of Partially Anonymized Microdata in the Course of Disasters. In *Workshop on Resilience and IT-Risk in Social Infrastructures (RISI 2011)*, 2011.
7. L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.
8. L. Sweeney et al. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, 10(5):557–570, 2002.
9. L. Willenborg and T. De Waal. *Statistical disclosure control in practice.* Springer Verlag, 1996.
10. L. Willenborg and J. Kardaun. Fingerprints in Microdata Sets. In *Joint ECE/EUROSTAT Work Session on Statistical Data Confidentiality.* Working Paper No. 10, 1999.