Electron Markets (2014) 24:113–124
DOI 10.1007/s12525-014-0154-x

SPECIAL THEME

An algorithm for collusion-resistant anonymization and fingerprinting of sensitive microdata

Peter Kieseberg · Sebastian Schrittwieser · Martin Mulazzani · Isao Echizen · Edgar Weippl

Received: 6 September 2012 / Accepted: 7 February 2014 / Published online: 29 April 2014

Abstract The collection, processing, and selling of personal data is an integral part of today's electronic markets, either as means for operating business, or as an asset itself. However, the exchange of sensitive information between companies is limited by two major issues: Firstly, regulatory compliance with laws such as SOX requires anonymization of personal data prior to transmission to other parties. Secondly, transmission always implicates some loss of control over the data since further dissemination is possible without knowledge of the owner. In this paper, we extend an approach based on the utilization of ***k***-anonymity that aims at solving both concerns in one single step - anonymization and fingerprinting of microdata such as database records. Furthermore, we develop criteria to achieve detectability of colluding attackers, as well as an anonymization strategy that resists combined efforts of colluding attackers on reducing the anonymization-level. Based on these results we propose an algorithm for the generation of collusion-resistant fingerprints for microdata.

Responsible Editor: Sven Wohlgemuth

P. Kieseberg (✉) · M. Mulazzani · E. Weippl
SBA Research gGmbH, Favoritenstraße 16, 1040 Vienna, Austria
e-mail: pkieseberg@sba-research.org

M. Mulazzani
e-mail: mmulazzani@sba-research.org

E. Weippl
e-mail: eweippl@sba-research.org

S. Schrittwieser
St. Pölten University of Applied Sciences,
Matthias Corvinus-Straße 15, St., 3100 Pölten, Austria
e-mail: Sebastian.Schrittwieser@fhstp.ac.at

I. Echizen
National Institute of Informatics,
2-1-2, Hitotsubashi, Chiyoda-ku, 101-8430 Tokyo, Japan
e-mail: iechizen@nii.ac.jp

Keywords Anonymization · Fingerprinting · Collusion-resistance · K-anonymity

JEL classification C88 - other computer software

Introduction

In today's electronic markets personal information becomes increasingly important as a tradable good. Entire business models such as the one of Facebook are built upon the generation of value from the collection, processing and selling of personal data from users. Especially in industries like telecommunication, providers need to handle large sets of structured data containing private information and exchange them with a fixed set of partners, e.g., in the course of interconnection billing or when using services of a clearing house. Due to the high sensitivity of personal data, compliance with laws and regulations such as SOX, HIPAA and others demands the anonymization of personal data prior to forwarding to other parties within a business context. In addition, once data is disclosed to another party, any further dissemination is out of control of the original owner. This raises the demand for tracking mechanisms to identify data leaks. Especially in the last years, the number of incidents including data loss rose dramatically, reaching a new all-.time high in 2012 according to "DataLossDB", a well renowned source for data breaches[1]. The inability of owners to keep control over the spread of data is in particular a threat to anonymization efforts, as the combination of multiple, differently anonymized sets can subvert the anonymization and disclose sensitive information. This attack vector is known as a collusion attack and is able to result in great financial damage to affected companies, on the one hand by direct penalties getting increasingly introduced

[1] www.datalossdb.org (Accessed: December 27th 2013)

by federal bureaus, on the other hand by eventually reducing customer's trust in the company, an asset that is especially important in today's electronic markets with their highly flexible and volatile customer bases.

Watermarking and fingerprinting are techniques which allow owners to add unique marks to their data sets to identify them later on. Identification of leaked anonymized data may be very important in markets dealing with information itself and concerning companies building their business model on retrieval and dissemination of information particles. For these companies, anonymized information holds a certain value, thus making the reliable and unambiguous identification of data leaks of vital importance in order to take legal actions and protect their core business. However, when considering several versions of the same data anonymized with different parameters, inference attacks can be used to identify and remove the digital mark or to corrupt it to render it unrecognizable and thus remove its value as a tracking mechanism.

In previous publications ((Schrittwieser et. al. 2011a) and (Schrittwieser et. al. 2011b)) we presented an idea which combines the anonymization of microdata (i.e., database records) and the generation of a fingerprint in a single step. In fact, the fingerprint is generated intrinsically by anonymizing the data, thus, our schema is based on the idea of extracting fingerprints from the data structure.

In this paper we show how a systematic selection of anonymization strategies prevents collusion attacks targeting the anonymization level as well as the robustness of the fingerprint. In particular, the core contributions presented in this paper are:

1. Description and analysis of collusions against our concept of *k*-anonymity based fingerprinting.
2. Protection of the anonymity level against collusions.
3. Achieving detectability of colluding attackers.
4. Presentation of an algorithm for constructing *k*-anonymity based fingerprints incorporating these results.

Our solutions can be used to keep track of data breaches and to attribute the breaches to the correct party. For example, a large retailer may hand over data to five market research firms. The data is aggregated so that individuals cannot be identified but only groups of 50 people (k=50 in k-anonymity). If two market research firms collude to reduce the level of privacy protection and manage to identify people in a group 10 (k=10) and this data is leaked, then the retailer will be able to identify which two market research firms colluded.

It must be noted that the approach proposed in this paper does not constitute a strategy for defending the underlying anonymization technique against inference attacks utilizing arbitrary external data, thus providing a flexible approach that can be adapted to be used with a wide variety of other methods for anonymization.

Background and related work

Watermarking and fingerprinting

Watermarking defines techniques that add visible or hidden information (e.g., a copyright notice) to the target data. An important characteristic of watermarking is that adding this information modifies the data, either visible or invisible to users. In contrast to watermarking, a consistent definition of fingerprinting does not exist among the research community. A common definition describes fingerprinting as a subtype of watermarking where a unique mark (i.e., the fingerprint) is added to each copy of the data. A second definition distinguishes fingerprinting from watermarking by the source of the fingerprint: While in watermarking information is added, fingerprinting uses intrinsic properties to uniquely differentiate the copies. In both definitions, however, the uniqueness of the fingerprint is the key concept that enables a data owner to uniquely link a customer to a specific copy.

In past literature, research on watermarking and fingerprinting techniques was primarily focused on multimedia data such as images or video files (Langelaar et al. 2000; Li et al. 2005; Fotopoulos and Skodras 2003; Hartung and Kutter 1999; Wu et al. 2004; Seo et al. 2005). Marking of non-multimedia files such as database tables, is far less explored. In 2002 Sion et al. (2002) presented a watermarking technique for relational databases based on watermarking a numeric collection, which is robust against several attacks such as data resorting, subset selection and linear data changes. Gross-Amblard (2003) discussed the problem of watermarking databases, while preserving a set of parametric queries in a specified language. Al-Haj and Odeh (2008) introduced a watermarking concept for databases based on the insertion of binary image watermarks in non-numeric multi-word attributes of selected tuples. The idea of embedding a watermarked image into a database was put forth by Zhang et al. (2004). Liu et al. (2005) introduced a block oriented fingerprinting approach for relational databases. In Lafaye (2007) a security analysis concerning watermarking schemas was performed. They analyzed the concepts in terms of uncertainty on the location of watermarked parts of the database, i.e., the difficulty for an attacker to identify the watermark. Another concept of fingerprinting microdata was discussed by Willenborg and De Waal (1996) and Willenborg and Kardaun (1999). In both approaches, fingerprints are built from combinations of identifying variables in the records.

Collusion-resistance, the inability of collaborating data receivers to remove the mark by combining their versions, is an important feature of watermarking and fingerprinting techniques. In recent literature several collusion-resistant approaches were introduced. Su et al. (2002) presented a collusion-resistant watermark for video data, furthermore, Su et al. (2005) introduced an approach against linear frame

collusions against watermarks in videos based on the implementation of a statistically invisible video watermark. Collusion-resistant fingerprinting was the target of research during the last decade as well. Trappe et al. (2003) investigated the problem of collusion-attacks and introduced a tree-structured detection algorithm for identifying colluders of fingerprinted multimedia files. Celik et al. (2003) proposed a fingerprinting schema for multimedia files that renders collusions ineffective by reducing the result of a collusion to low quality and a related technique considering a combination of digital watermarks and collusion secure fingerprints for digital images in the medical sector was proposed in (Dittmann et al. 2000).

Anonymization of microdata

In 2002 Sweeney showed that even after removing attributes that uniquely identify persons, e.g., the social security number, from medical data, it is possible to identify 87% of all Americans based on combining so-called quasi identifiers (QIs) like birthdate, zip-code, sex and combinations of QIs with external data (Sweeney 2002b). Thus, to prevent linking, Sweeney introduced a new concept called *k-anonymity* which is a widely adopted anonymization technique in academia nowadays (Sweeney et al. 2002). Since nowadays critical business information, especially considering customer information or CDRs (call detail records), is typically stored in databases, mechanisms for fingerprinting such structured datasets are needed.

Definition All attributes in a data set that either themselves (e.g., name) or in combination (date of birth, sex) can be used to uniquely identify a person are called ***quasi identifiers*** (Dalenius 1986).

Currently many publications (e.g., El Emam et al. (2009)), remove identifiers like names that themselves identify a person from the set of QIs and from the published data altogether.

Definition A set of records obeys the ***k-anonymity*** criterion for a given k, when each record is indistinguishable from at least k-1 other records with respect to all QIs.

Thus, the data is partitioned into equivalency classes, where each class holds at least k elements. The level of anonymity can be raised by increasing k, in practice however, lower levels need to be used for providing significance for meaningful analysis. Several improvements have been devised for this concept, e.g., l-diversity (Machanavajjhala et al. 2007), our approach works with them as well, thus we describe it based on k-anonymity for reasons of simplicity.

Example Table 1 shows data containing two QIs that is obeying k-anonymity (k=2).

Table 1 Anonymized data

Birthday	Sex	Disease
1970	F	Chest-pain
1970	M	Short-breath
1970	F	Obesity
1970	M	Short-breath

Generalization patterns

The most prominent technique for achieving k-anonymity rests upon generalizations of QIs: The granularity of the QIs is reduced in all records until the criterion is fulfilled (Sweeney, 2002a). Naturally, different generalization strategies may be used. Figure 1 shows an example containing a set of possible generalization levels for two QIs.

Definition Let n be the number of QIs. For each QI i=1,…, n exactly one generalization level a_i is chosen. Then the tuple $\boldsymbol{a}:=(\boldsymbol{a_1},\ldots\boldsymbol{a_n})$ of generalization levels is called a ***generalization pattern*** of the data set.

Example The data shown in Table 1 is anonymized with the pattern (0,2) with respect to the generalization levels of Fig. 1.

Each pattern represents exactly one way of generalizing the records of the set with respect to the strategies chosen for each QI. Thus, it is possible to construct a lattice diagram showing all possible patterns (see Fig. 2), the edges show direct generalizations that are derived by generalizing one identifier by a single level.

In addition to generalization, *suppression* of records, i.e., deleting a (small) subset of records from the set, is frequently used for achieving k-anonymity. Being compatible with our approach, its effects are omitted for the remainder of this paper.

Fingerprinting with k-anonymity

In this section we describe our general approach based on our work in (Schrittwieser et. al. 2011a) and (Schrittwieser et. al. 2011b), we included this chapter to achieve a comprehensive view on the subject. Furthermore, many formal aspects were not included in the previous works. The outline of the approach is based on the concept of k-anonymity for reasons of simplicity, still it can be exchanged for enhancements like l-diversity or t-completeness.

When distributing structured data to consumers, sensitive information has to be obfuscated in order to prevent identification of individuals. Still, especially in the case of medical data, or data of financial value, the owner needs to be able to detect information leaks, especially when the data constitutes part of the business case of the company, as it is the case for

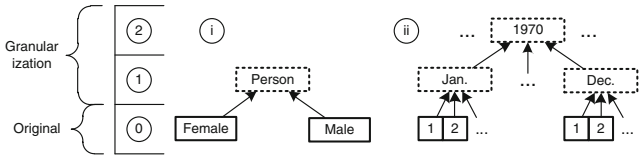

Fig. 1 Different levels of generalization

e.g., Facebook, or is used in order to apply anti-fraud mechanisms, financial clearing and network optimization utilizing external institutions and experts as is commonly done in the telecommunications industry. In our approach, intrinsic structural information of the data is used to form a fingerprint that cannot be removed without reducing its value. More precisely, the basic idea lies in the utilization of different anonymization strategies, every consumer receives her set in a slightly differently anonymized form (see Table 2): For example, consumer U_1 receives set one, where strategy (0,2) is used (see Fig. 1) and consumer U_2 receives set two with data in the form (1,1). In case the owner finds data like (M, 1970) in the wild, he is able to identify U_1 as source, since U_2 would not be able to provide this level of detail for QI "sex". Figure 3 illustrates the process of leak detection.

Data precision metrics (DPM) are used to find generalization strategies with approximately the same information loss to provide consumers with data of comparable value. To achieve this, all strategies that lead to k-anonymized sets with a given k are generated and the resulting information loss is determined with a predefined DPM. The strategies are clustered into equivalency classes according to their information loss and a tolerance t (see Fig. 4). If no class holds enough strategies, either t or k may be changed. It must be noted that the resulting classes are strongly depending on the chosen DPM which will largely depend on the specific use-case in order not to destroy the significance of the data.

We have devised an algorithm for fingerprint generation and leak detection in (Schrittwieser et. al. 2011b), which is based on El Emam's algorithm for calculating the optimal solution (see El Emam et al. (2009)).

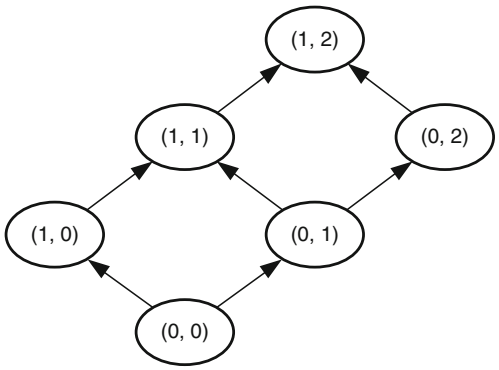

Fig. 2 Lattice diagram showing the generalization patterns

Collusion attacks

The combination of two or more data sets can be used to subvert both, the anonymization as well as the fingerprinting efforts (collusion attack), which was not taken into account in the original approach proposed in (Schrittwieser et. al. 2011a). Therefore, we analyze this threat and show which anonymization strategies must be used to mitigate the resulting effects. Chapter 4.1 is concerned with the analysis of collusions against k-anonymity, whereas Chapter 4.2 proposes a strategy for identifying leaks. The reduction of the anonymization level may not only result in reputation loss of the data owner, it might even lead to legal actions or reveal critical business information to competitors. On the other hand, the subversion of the fingerprinting property of the scheme might lead to severe financial loss for companies that base part of their business on the trading of information, by making it publically available without the possibility for the original data owner to hold the leaking party responsible for the disclosure.

As a prerequisite we assume that two different generalizations of the same record set can be identified and matched with each other, e.g., by using non-QIs like medical diagnosis or primary keys. This prerequisite constitutes a kind of worst-case scenario, still, in the case of large amounts of structured data, the ability to match records from differently anonymized sets must be taken into account to effectively protect sensitive data.

Example Table 2 shows original data together with two different anonymizations, the fields "sex" and "birthday" were identified as QIs. Both sets obey k-anonymity for k=2, still the records can be matched by using the field"disease" which was not identified as QI (and will not be generalized in order not to reduce the data value).

Analysis of collusion attacks against the k-anonymity criterion

As a starting point for our analysis we formalize the information gain inherent to knowing a specific pattern, thus defining its hull. Using this, we formalize the information gain through collusions and the possible consequences for anonymization.

Definition Let A be a set of generalization patterns. The ***hull*** of a pattern a_0 is the set of all patterns that can be constructed by lowering the granularity of QIs:

$$H(a_0) := \left\{ a_i \in A \;\middle|\; a_{i_j} \geq a_{0_j}, \forall j = 1, \ldots, n \right\},$$

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

Original data (0,0)				First set (0,2)			Second set (1,1)		
Name	Sex	Birthday	Disease	Sex	Birthday	Disease	Sex	Birthday	Disease
Bob	M	19.03.1970	Chest-pain	M	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	Cancer	F	1970	Cancer	P	04.1970	Cancer

where a_{i_j}, a_{0_j} denote the generalization levels the j-th QI of the patterns a_i and a_0 respectively. Furthermore, $a_{i_j} > a_{0_j}$ denotes that the generalization level of the j-th QI is higher in the pattern a_i than in pattern a_0, i.e., that the granularity of pattern a_0 is higher than of a_i with respect to the j-th QI. This slight change in notation compared to the previous chapters is needed since we will have to compare several patterns and wanted to avoid double indices before.

In other words, the hull of a pattern contains all generalizations that can be constructed by an attacker without additional information. Figure 5 shows an example lattice diagram based on two QIs together with the hull of a pattern.

Corollary If a given pattern a_0 obeys k-anonymity for a given k, then all elements of the hull $H(a_0)$ obey k-anonymity for at least the same k.

Proof The proof for this corollary is rather trivial: Following the definition of the hull it holds true that $\forall a_i \neq a_0 \in H(a_0) : \left(\exists j : a_{i_j} > a_{0_j}\right) \wedge \left(\nexists j : a_{i_j} < a_{0_j}\right)$, i.e., for all patterns in the hull $H(a_0)$ except for a_0 itself at least one QI has a higher generalization level, while none has a lower one (this is derived directly from the definition of the hull).

Following we want to analyze, what patterns can be constructed by applying collusions against two different known patterns $a=(a_i)_{i=1,\ldots,n}$ and $b=(b_i)_{i=1,\ldots,n}$. By using the definition of the hull, all patterns within the union of the two hulls can be constructed trivially (Fig. 6 shows an example with two QIs).

Furthermore, with the prerequisite given in the introduction of "Collusion attacks", the possibility of matching the same record in differently anonymized sets, it is possible to construct the pattern $c=(\min(a_1,b_1),\ldots,\min(a_n,b_n))$ and thus the hull $H(c) \supseteq H(a) \cup H(b)$ (Fig. 7 shows an example involving two QIs). We call c the *minimal generalization pattern* with respect to the patterns a and b. The minimal generalization pattern always exists because either, without loss of generality, $a \in H(b)$ trivially yielding $c=b$ (and vice versa for $b \in H(a)$), or not. In the latter case it is always possible to use the construction by calculating the minimal generalization levels with respect to each QI. Since we implicitly assume a finite number of QIs and the ">"-relation forms a well-ordering on the generalization levels for each QI, the minimum can always be calculated.

This can be further extended to the general case of a finite number of generalization patterns:

Definition Let $A=\{(a_i)_{i=1,\ldots,r}\}$ be a set of r generalization patterns with respect to n QIs. We define the hull $H(A)$ as $H(\overline{a})$, where $\overline{a}$ is the minimal generalization pattern with respect to all $a_i \in A$.

$$H(A) := H\left(\overline{a}\right) = H(\min(a_{1_1}, \ldots, a_r), \ldots, \min(a_{1_n}, \ldots, a_{r_n}))$$

Since this is just an extension of the two-dimensional to the general finite-dimensional case, the existence of the minimal generalization pattern can be derived easily by using complete induction.

Theorem Let A be a set of generalization patterns distributed to consumers, each obeying the k-anonymity-condition for k. Then the level of anonymity of the data cannot be reduced beyond k by any selection of colluding consumers if and only if the minimal generalization pattern of A at least obeys the same k-anonymity-condition.

Proof In case the minimal generalization pattern does not obey k-anonymity for k, a collusion of all consumers breaks

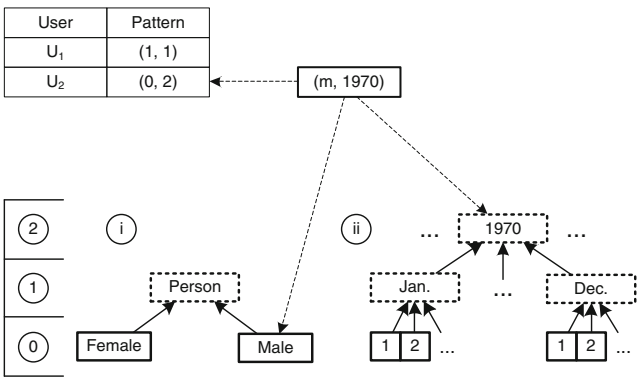

Fig. 3 Identifying the source of data leakage

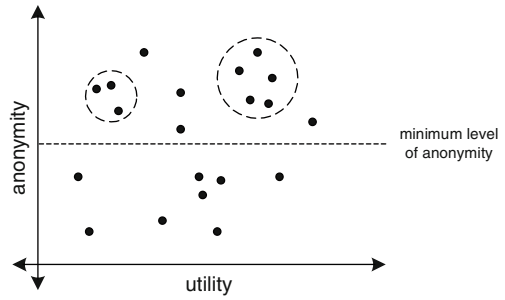

Fig. 4 Clustering anonymization strategies

the anonymization. On the other hand, the minimal generalization pattern contains the most granular generalization of each QI with respect to the information given to all consumers, i.e., it is not possible to generate another pattern from the patterns owned by the consumers that has a higher granulated QI. Thus, in case the minimal generalization pattern obeys *k*-anonymity, all other patterns derived by colluding the consumers data sets obey at least the same criterion.

Identification of colluding attackers

Following we discuss how to choose generalization patterns to identify the colluding attackers. For this, we need more prerequisites regarding our attacker model.

Prerequisites Like in the original fingerprinting approach, we assume that attackers always try to generate the best possible data set, i.e., they are not willing to reduce the granularity of a QI in case a higher one is available to them (this is also discussed in the original paper under "Limitations" and in "Evaluation" "Robustness" in this paper). One reason for this lies in the fact that every fingerprint (even without considering collusions) can be removed trivially by generalizing all QIs to the maximum generalization level (of course this includes losing virtually everything in terms of quality). Furthermore,

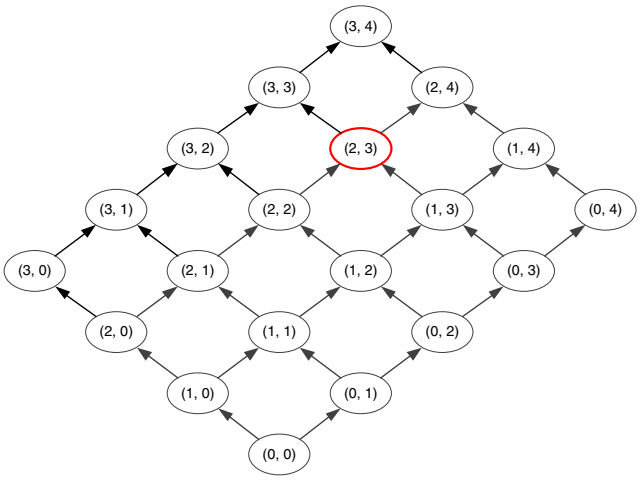

Fig. 5 The hull of a generalization pattern

we assume that consumers want to generate as good data as possible through collusions. As an additional prerequisite, we assume that attackers don't know the patterns distributed to non-colluding consumers. Please note that these prerequisites are not needed in "Analysis of collusion attacks against the k-anonymity criterion".

Thus, the following targets need to be achieved by such a solution to be **defined** as **resistant to collusion attacks** with respect to our prerequisites:

1. All colluding partners shall be identified
2. No innocent partner shall be suspected wrongfully

The first observation is that any pattern a_i within the hull of a pattern a_0 does not fulfill above requirements since a collusion would go undetected. Generalization of this statement leads us to the following theorem:

Theorem Let $A=\{a_i|i=1,\ldots,r\}$ be a set of patterns. Then A constitutes a set of patterns resistant to collusions, if and only if $a_i \notin H(A\backslash\{a_i\}), \forall i=1,\ldots,r$.

Proof (1) Let $a_0 \in A$ be a pattern for which the precondition holds true, i.e., $a_0 \notin H(A\backslash a_0) =: H(\overline{a_0})$, where $\overline{a_0}$ is the minimal generalization pattern of $A\backslash a_0$. Thus follows that a_0 has at least one QI a_{0_j} of finer granularity than $\overline{a_0}$, thus is not constructible solely using the patterns in $A\backslash a_0$. If a set anonymized with a pattern containing a_{0_j} with the granularity found in a_0, it can be guaranteed that the user holding pattern a_0 participated in the collusion. If generalized for all $i=1\ldots r$, the first implication follows trivially. (2) On the other hand, if all colluding partners shall be identifiable, it must not be possible to construct a pattern from the set containing the other patterns $A'=A\backslash a_0$, i.e., $a_0 \notin H(A') \Rightarrow \exists a_{0_j} : a_{0_j} < a_{i_j}, \forall i \neq 0$. Since this must hold true for all patterns a_i, we can generalize that $\forall l = 1\ldots r : \exists a_{l_j} : a_{l_j} < a_{i_j}, \forall i \neq l$ trivially leading to $a_i \notin H(A\backslash a_i), \forall i=1,\ldots,r$.

Example Figure 8 shows three patterns a_1, a_2, a_3. In case of collusion of the partners holding data generalized with a_1 and a_3, the partner having data of the form a_2 could be accused innocently. A collusion of partner a_2 and a_3 would not be detected, partner a_3 would still be deemed trustworthy.

Above theorem states that every pattern in A has to be more detailed than all the others with respect to at least one QI and the attacker model outlined at the start of this section.

Corollary The maximum number of independent patterns is bound by the number of QIs, i.e., the number of

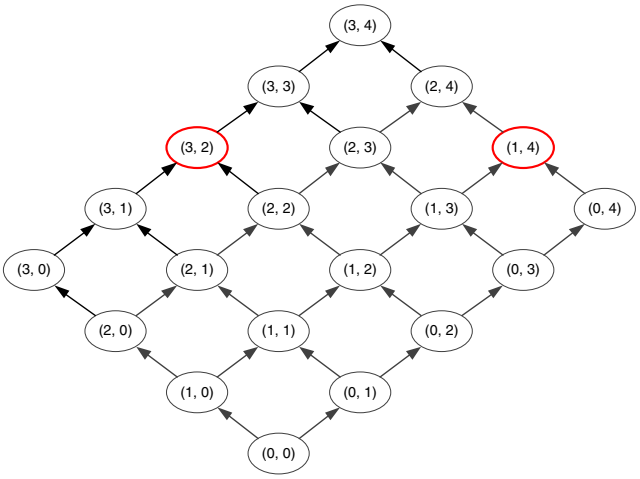

Fig. 6 The union of the hulls of two patterns

partners must be smaller or equal than the number of QIs.

It must be noted that detecting colluding attackers is completely independent from providing sets that obey *k*-anonymity. Thus in real-life scenarios the number of possible colluding consumers that can be detected while still obeying *k*-anonymity may be much smaller.

Algorithm for a privacy preserving generalization strategy

Based on the last section and the algorithm proposed in (Schrittwieser et. al. 2011b), we propose an algorithm for generating a privacy preserving solution. This solution is needed to achieve both features mandatory in a fingerprinting scheme targeting information business: An irreducible anonymity level while maintaining enough stability in order to identify colluding data customers.

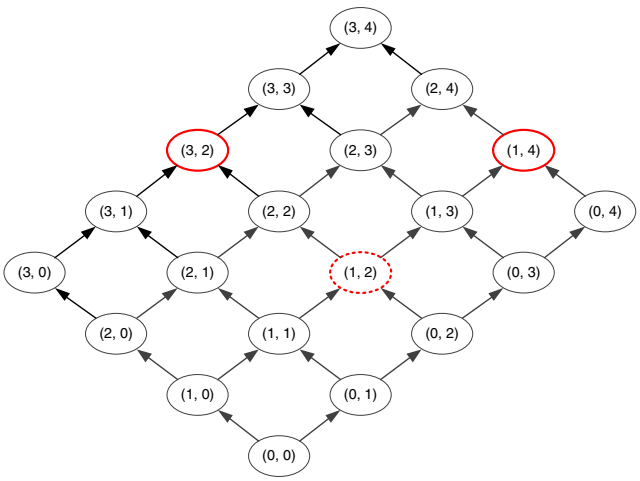

Fig. 7 The hull generated by colluding two patterns

Privacy preserving generalization strategies

Definition We define a set A consisting of r generalization patterns as *privacy preserving*, when the solution is conforming to the following properties:

1. Each pattern at least obeys the chosen anonymization level k.
2. The pattern generated by colluding all r patterns of A at least obeys k-anonymity for the same k.
3. In case of collusions, it is possible to identify the participants.
4. All patterns $a_i \in A$ have the same quality with respect to the chosen DPM and tolerance.

If several privacy preserving generalization strategies are found it is necessary to choose a "best" solution. Several measures can be used for defining this solution, e.g., *lowest average* data loss, *lowest difference in quality* between the patterns or *highest number of patterns*.

As discussed in "Analysis of collusion attacks against the k-anonymity criterion", a privacy preserving solution as defined above can only be found in case the number of consumers r is not greater than n, the number of QIs. Depending on the actual sets and the anonymization level demanded, r may even be significantly smaller than n to be able to provide a generalization strategy that obeys the criteria mentioned above. Thus, it may be useful to define a weaker definition for privacy preserving generalization strategies that can be obeyed more easily while still providing reasonable security. However, we believe that typical databases contain enough QIs for our approach to be practical.

The algorithm

The algorithm proposed in this section is based on El Emam's algorithm for calculating an optimal solution for the *k*-anonymity problem of single sets (i.e., finding the pattern with minimum information loss), combined with the results from "Collusion attacks" concerning resilience against collusions.

1. All side parameters are defined: The anonymization level k, minimum/maximum bounds for data loss l_{min} and l_{max}, the DPM and the tolerance t. Furthermore, the algorithm terminates in case $n<r$.
2. For each QI, a generalization strategy including the different levels of granularity is defined.
3. The lattice diagram holding all possible patterns is calculated.

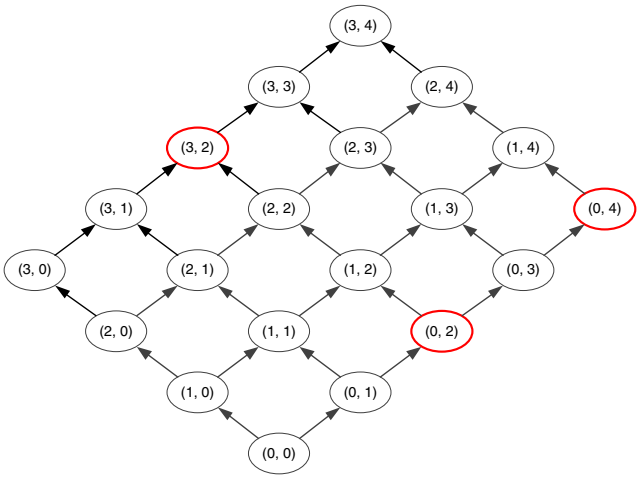

Fig. 8 Collusion example

4. A node at middle height is chosen and tested for *k*-anonymity.
   a. If not, all nodes below in the lattice diagram (including the current one) are ruled out as possible solutions.
   b. In case it does, all nodes above (including the current one) are marked as possible solutions.
5. Step four is repeated for the remaining subgraphs. If all nodes below were ruled out (4.a), one QI of the node used in step 4 is increased by one and used as input in step 4. In case of 4.b, the same approach is followed with one QI decreased. This step is repeated for all remaining subgraphs until all nodes in the original subgraph are evaluated.
6. In case no subgraph is left, another still unevaluated node at middle height is chosen and the algorithm proceeds with step four until all nodes are evaluated.
7. For each node marked as possible pattern, the DPM and the actual k is calculated. All patterns with a precision outside l_{min} and l_{max} are removed.
8. The remaining patterns are clustered into equivalency classes according to their quality and the tolerance t.
9. For each resulting class, the minimal generalization pattern is generated. Classes resulting in a pattern not obeying k-anonymity are removed from the possible solutions.
10. For each remaining class, the ability of detecting colluding attackers is tested according to "Identification of colluding attackers" and eventually removed from the set of possible solutions.
11. If the set of possible solution is empty, t, l_{min} and l_{max} are relaxed with respect to a predefined strategy. The algorithm then proceeds with step seven in case $t \leq l_{max}$ (in case $t \geq l_{max}$, t can obviously be set to l_{max} without losing any solutions. The algorithm then proceeds with step seven too).
12. If more than one class remains, the best solution is chosen (see Privacy preserving generalization strategies).
13. The data sets for the consumer are generated.

DPMs and maximum levels (Step 1) The algorithm should provide approximately the same level of precision to all consumers. Still, it may not be possible to find a generalization strategy with r patterns possessing exactly the same quality. Thus, the tolerance parameter t is defined as the maximal tolerated difference in quality inside an equivalency class, i.e., $t \geq (\max_i(M(a_i)) - \min_j(M(a_j))$, where M is the DPM used. The parameters l_{min} and l_{max} hold bounds for the minimal and maximal data precision of a valid solution.

Eliminating nodes (Step 4) Since we need the set of all possible solutions to construct the equivalency classes (this is a major deviation from the algorithm outlined in (El Emam et al. 2009)), the nodes holding less granular patterns cannot be removed from the set of possible solutions in case of 4.b.

Clustering the solutions (Step 8) Step eight clusters the patterns into classes by their quality and with respect to the tolerance t. Patterns may (and with a higher t will most likely be) inside several such equivalency classes, especially since all subsets of clusters are valid classes too. Classes holding less than r patterns are removed from the list of solutions.

Finding a privacy preserving solution (Steps 9 and 10) In these steps, the results from "Collusion attacks" are incorporated to find a privacy preserving solution as defined in "Privacy preserving generalization strategies". Step nine guarantees that colluding all patterns of the resulting generalization strategy does not result in a pattern that fails the k-anonymity criterion, while step ten guarantees the identification of each colluding consumer.

Relaxing the side parameters (Step 11) In case no solution for a privacy preserving generalization strategy with respect to the parameters t, l_{min} and l_{max} is found, the algorithm tries to relax these parameters and proceeds with step seven to find a privacy preserving generalization strategy with lower requirements. The strategy for relaxing these parameters must be defined by the user beforehand. Choosing suitable strategies depends very much on the actual data and requirements and is not part of this work.

Choosing the best solution (Step 12) Several strategies may be chosen for selecting the best privacy preserving strategy: "Privacy preserving generalization strategies" gives some ideas on reasonable choices for comparing the solutions.

In the course of the paper we always demonstrated our approach with respect to the common concept of k-anonymity. In case other, related, concepts like l-diversity or t-completeness are used, the check for compliance of the pattern in step 4 has to be adapted to resemble the new criteria.

Applicability of the algorithm

To demonstrate the applicability of the approach, we provide a small example. Due to the regulations regarding the paper length we will omit unnecessary intermediate results.

Step 1: Chose side parameters: The anonymization level k=2, l_{min}=0, l_{max}=4 and t=0 using the Samarati Metric (SM) (see (Samarati, 2001)). Furthermore we want to distribute the fingerprinted data to three consumers. The original data can be found in Table 3 (the attributes ID, birthday, ZIP-Code and sex, with generalization level 0, i.e., the columns with "0" in the second row).

Step 2: For the QIs birthday, ZIP-Code and sex the following generalization levels are defined:

- Birthdate: exact to: day, month, year, decade
- ZIP-Code: sub district (4 digits), district (3 digits), inner/outer city (2 digits), city (1 digit)
- Sex: female (F)/male (M), person (P)

See Table 3 for all generalizations.

Step 3: The lattice diagram holding all possible generalization patterns is calculated.

Step 4: Since birthday and ZIP-Code both possess 4 generalization levels $(0,\ldots,3)$ and sex possesses 2 levels $(0,1)$, the maximum height of the lattice diagram is 7 (3+3+1), the minimal 0 (0+0+0), thus all levels with a combined sum of generalization levels of 4 are at middle height (since we use the Samarati Metric for measuring the data precision, the definition of the height in the lattice diagram and the data loss is equal. This does not hold for other DPMs though). In our example seven nodes with height 4 exist.

A node at middle height is chosen at random and checked for compliance with k-anonymity for k=2. For example we choose the node a_0=(0,3,1) which doesn't obey the criterion (see Table 3). Thus all nodes (x,y,z) with $x\leq 0$, $y\leq 3$ and $z\leq 1$ are removed from the set of possible generalization patterns (see step 4.a in the algorithm).

Step 5: Step four is repeated for the remaining subgraphs, i.e., a node that can be derived by lowering the granularity of one QI in a_0 is constructed and used as input for step 4, e.g., (1,3,1). This node fulfils k-anonymity, thus all nodes (x,y,z) with $x\geq 1$, $y\geq 3$ and $z\geq 1$ are added to the list of possible generalization patterns. This is done for all remaining subgraphs.

Step 6: The steps four and five are repeated until all nodes at middle height are evaluated. Table 4 shows the set of possible generalization patterns before applying the bounds for the data precision.

Step 7: All patterns with data loss higher than 4 are removed, for all remaining nodes the actual k is calculated.

Step 8: The remaining generalization patterns are clustered with respect to the tolerance level t=0. Table 5 gives an overview on the resulting clusters (in order to keep this example well-arranged, Table 5 already contains the results from steps nine (the minimal generation pattern $\overline{a}$) and the clusters are grouped by their data loss and the number of patterns respectively).

Step 9: For each cluster, the minimal generalization pattern is constructed and tested whether it obeys k-anonymity. Due to the simple metric, small tolerance and small data set, the minimal generalization pattern for each cluster is the same in this example.

Step 10: All clusters that do not provide the ability to detect colluding attackers are removed from the sample, reducing the set of possible results to the pattern with ID 11 (highlighted in Table 5).

Steps 11 and 12: The side parameters$t$, l_{min} and l_{max} could be lowered to generate more solutions, which we omit in this example. The data is anonymized with the resulting patterns from the cluster with ID 11 and distributed (see Table 6).

Evaluation

Number of possible data consumers For our approach it is vital that every consumer receives the data anonymized with a unique, non-reusable pattern. In general, the number of possible fingerprints is heavily depending on side-parameters like the anonymization level, the minimal and maximal allowed data quality and clustering tolerance, as well as on the data itself. Furthermore, since the approach requires a privacy preserving generalization strategy, the number of possible patterns is reduced drastically and can be limited to $F_{opt}\leq n$. Still, e.g., considering call-detail-records (CDRs) in interconnection-billing or anti-fraud systems for telecommunication providers, the typical data set possesses much more QIs (more than 50 for a typical ICB-system) than possible consumers (usually less than 10), thus rendering this approach perfectly feasible.

Robustness of the fingerprints and the anonymization Another important evaluation criterion lies in the robustness of the

Table 3 All generalization levels for the chosen data set

	Birthday				ZIP-code				Sex	
ID	0	1	2	3	0	1	2	3	0	1
1	31.05.1970	05.1970	1970	70s	1042	104	10	1	F	P
2	25.04.1970	04.1970	1970	70s	1062	106	10	1	M	P
3	16.05.1970	05.1970	1970	70s	1041	104	10	1	F	P
4	08.04.1970	04.1970	1970	70s	1062	106	10	1	M	P

fingerprint against tampering and removal. Since the fingerprint is constituted by the intrinsic structure of the data, the only way to remove such fingerprints lies in changing the data structure, i.e., changing the underlying anonymization pattern. In order to hide, the consumer would need to find a pattern that is either distributed to another consumer (which is impossible when using a privacy preserving generalization strategy), or that at least lies in the hull of another distributed pattern. To achieve this, the attacker must (i) know the pattern used for another consumer and (ii) know the generalizations levels of the identifiers. At any rate, the value of the leaked data will be drastically reduced, since the attacker is only able to reduce the granularity of known data in order to hide the pattern. Concerning the robustness of the anonymization, the approach proposed in Chapter 4.1 ensures that the *k*-anonymity criterion cannot be broken by inferencing an arbitrary selection of the fingerprinted data sets. Still, attacks against the concept of *k*-anonymity itself, especially considering inference attacks involving external data, may lead to unwanted disclosure.

Example Let *A* and *B* be two consumers, where *A* gets data anonymized with (3,1) and *B* with (1,3). Assuming *A* knows the pattern for *B* and the generalization levels for each QI, the best unidentifiable pattern that *A* can generate would be (3,3).

Validity of the algorithm For the practical applicability of the approach an algorithm that is able to construct the fingerprint is needed. Thus we will show that the algorithm proposed in Chapter 5 terminates and that all possible privacy preserving solutions are generated:

Steps 1 to 3 in the algorithm set the side parameters and generate the generalization levels and the lattice diagram which terminates trivially in case of a finite number of QIs and possible generalizations. The steps 4 to 6 iterate through all nodes in order to decide if they obey *k*-anonymity. In every invocation of step 4, nodes are either marked as possible patterns or removed from the evaluation. In both cases, these nodes will not be chosen by steps 5 or 6 as input node for step 4, thus guaranteeing that each node is evaluated at most once. Thus this cycle terminates after a finite number of iterations. Step 7 terminates trivially, step 8 generates all possible clusters from the set of possible patterns which is again finite. Since the minimal generalization pattern always exists (see Chapter 4.1), step 9 and 10 terminate too.

Furthermore, the algorithm needs to generate all solutions. Lets assume one solution is missing, since the solutions are generated by clustering all patterns obeying *k*-anonymity into equivalency classes, it follows trivially that at least one pattern a_0 must be missing. Since the lattice diagram calculated in step 3 holds all possible patterns, a_0 must have been lost in the loops implemented in steps 4 to 6. This part is functionally equivalent to the algorithm proposed in (El Emam et al. 2009) and proven to be complete: For each middle node, either the lower subgraph (including the middle node) can be marked as possible pattern, or can be ruled out. For each middle node the algorithm cycles through the remaining subgraph and the process is repeated until no unevaluated nodes are left.

Conclusion

In this paper we introduced an algorithm for collusion-resistant anonymization and fingerprinting of microdata in one single step based on the very popular and widely used concept of *k*-anonymity. Both features are often demanded

Table 4 Possible generalization patterns

Pattern	SM	k	Pattern	SM	k	Pattern	SM	k	Pattern	SM	k
(1,1,0)	2	2	(1,3,0)	4	2	(2,2,0)	4	2	(3,1,0)	4	2
(1,1,1)	3	2	(1,3,1)	5	2	(2,2,1)	5	4	(3,1,1)	5	2
(1,2,0)	3	2	(2,1,0)	3	2	(2,3,0)	5	2	(3,2,0)	5	2
(1,2,1)	4	2	(2,1,1)	4	2	(2,3,1)	6	4	(3,2,1)	6	4
(3,3,0)	6	2	(3,3,1)	7	4						

when exchanging sensitive data and thus a key requirement in many business scenarios, e.g., the exchange of interconnection information between telecommunication providers or information trading. Based on our previous work, we evaluated the effects of colluding data consumers towards the stability of the anonymization and the robustness of the fingerprints and formalized them. This resulted in clear formal criteria on the characteristics of collusion-resistant fingerprinting strategies and allowed the selection of optimal strategies that guarantee a reasonable protection without needless reduction of data value. The algorithm we proposed based on these characteristics uses a strategic selection of different generalization patterns for achieving *k*-anonymity which provably cannot be combined to a set that is below a specified anonymity level. We further showed that our algorithm generates sets where colluding attackers can be identified with respect to a realistic attacker model. The approach was evaluated regarding its practical applicability based on the key factors robustness, validity and completeness, as well as the number of possible consumers. Due to the theoretical nature of the approach the actual applicability highly depends on the underlying data and the anonymization level. To indicate its practical applicability, we gave an example using a very small and limited data set while still being able to generate a reasonable solution. In case of targeting real life applications, e.g., in the telecommunications industry, we expect the approach to yield even better results due to data sets containing far more attributes.

We thus conclude that it is possible to construct fingerprints based on the intrinsic structure of an anonymization procedure that allow for the unique identification of leaks with respect to the outlined prerequisites. Furthermore we can guarantee that this distribution of several, differently anonymized sets based on the same original data does not pose a threat to the sensitive information contained therein. As the approach outlined in this work is not directly based on a specific anonymization strategy, it can be easily generalized to use other methods such as *l*-diversity (Machanavajjhala et al. 2007) or t-closeness (Li et al. 2005). We outlined this in the course of the paper, especially considering the applicability of our algorithm as given in "The Algorithm".

As already outlined in the introduction, we see many possible applications for this approach in situations that are omnipresent in modern electronic markets. Telecommunication and other network providers may involve several research institutes or companies to analyze (e.g., traffic based) user behavior and/or network utilization e.g., for optimization purposes. By using our algorithm any unwanted disclosure due to uncontrolled information exchange and collusion between these institutions can be identified. Another prime example for utilizing the approach outlined in this work relates to anti-fraud mechanisms based on independent data streams handling connection data, thus including sensitive user information. Furthermore it has to be kept in mind that even anonymized data constitutes a business asset for many companies targeting modern electronic markets. As modern society is increasingly getting aware of the sensitivity of private information, IRBs and ethical commissions require mechanisms for reliable protection. Whenever data is processed or stored and aggregate data suffices, it makes sense to only work with the aggregate data and to only transfer this data to collaborators. Our scheme further improves this process by adding identifying watermarks to the data that allow discovering collusion and correctly attributing the violations to the parties involved. We believe that the contributions in this paper will support scientists conducting research related to electronic markets by allowing them easier access to aggregate data to base their work on.

For our future work, we aim at researching the impact of different data precision metrics, especially regarding the scientific value of data sets. Furthermore we will explore different relaxed definitions for privacy preserving generalization strategies and implement the concept for the MySQL DBMS.

Table 5 The constructed clusters

ID	SM	# patterns	Patterns	$\overline{a}$	k for $\overline{a}$
1	3	3	(1,1,1),(1,2,0),(2,1,0)	(1,1,0)	2
2	4	5	(1,2,1),(1,3,0),(2,1,1),(2,2,0),(3,1,0)	(1,1,0)	2
3	4	4	(1,3,0),(2,1,1),(2,2,0),(3,1,0)	(1,1,0)	2
4	4	4	(1,2,1),(2,1,1),(2,2,0),(3,1,0)	(1,1,0)	2
5	4	4	(1,2,1),(1,3,0),(2,2,0),(3,1,0)	(1,1,0)	2
6	4	4	(1,2,1),(1,3,0),(2,1,1),(3,1,0)	(1,1,0)	2
7	4	4	(1,2,1),(1,3,0),(2,1,1),(2,2,0)	(1,1,0)	2
8	4	3	(1,2,1),(1,3,0),(2,1,1)	(1,1,0)	2
9	4	3	(1,2,1),(1,3,0),(2,2,0)	(1,1,0)	2
10	4	3	(1,2,1),(1,3,0),(3,1,0)	(1,1,0)	2
11	4	3	(1,2,1),(2,1,1),(2,2,0)	(1,1,0)	2
12	4	3	(1,2,1),(2,1,1),(3,1,0)	(1,1,0)	2
13	4	3	(1,2,1),(2,2,0),(3,1,0)	(1,1,0)	2
14	4	3	(1,3,0),(2,1,1),(2,2,0)	(1,1,0)	2
15	4	3	(1,3,0),(2,1,1),(3,1,0)	(1,1,0)	2
16	4	3	(1,3,0),(2,2,0),(3,1,0)	(1,1,0)	2
17	4	3	(2,1,1),(2,2,0),(3,1,0)	(1,1,0)	2

Table 6 The distributed data sets

Data set 1			Data set 2			Data set 3		
Birthday	ZIP-code	Sex	Birthday	ZIP-code	Sex	Birthday	ZIP-code	Sex
05.1970	10	P	1970	104	P	1970	10	F
04.1970	10	P	1970	106	P	1970	10	M
05.1970	10	P	1970	104	P	1970	10	F
04.1970	10	P	1970	106	P	1970	10	M

While this paper focuses on the concept of k-anonymity as anonymization strategy, we plan to apply our method on other techniques for privacy protection.

References

Al-Haj, A., & Odeh, A. (2008). Robust and blind watermarking of relational database systems. *Journal of Computer Science, 4*(12), 1024–1029.

Celik, M., Sharma, G., & Tekalp, A. (2003). *Collusion-resilient fingerprinting using random prewarping*. Proceedings of the International Conference on Image processing, ICIP 2003. vol. 1, pp. 1–509, IEEE.

Dalenius, T. (1986). Finding a needle in a haystack – or identifying anonymous census record. *Journal of Official Statistics, 2*(3), 329–336.

Dittmann, J., Schmitt, P., Saar, E., Ueberberg, J., & Schwenk, J. (2000). Combining digital watermarks and collusion secure fingerprints for digital images. *Journal of Electronic Imaging, 9*(4), 456–467.

El Emam, K., Dankar, F., Issa, R., Jonker, E., Amyot, D., Cogo, E., et al. (2009). A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association, 16*(5), 670–682.

Fotopoulos, V., & Skodras, A. (2003). Digital image watermarking: an overview. *EURASIP Newsletter, 14*(4).

Gross-Amblard, D. (2003). *Query-preserving watermarking of relational databases and xml documents*. SIGART Symposium on Principles of Database Systems.

Hartung, F., & Kutter, M. (1999). Multimedia watermarking techniques. *Proceedings of the IEEE, 87*(7), 1079–1107.

Lafaye, J. (2007). *An analysis of database watermarking security*. Symposium on Information assurance and security.

Langelaar, G., Setyawan, I., & Lagendijk, R. (2000). Watermarking digital image and video data. A state-of-the-art overview. *IEEE Signal Processing Magazine, 17*(5), 20–46.

Li, W., Yuan, Y., Li, X., Xue, X., & Lu, P. (2005). Overview of digital audio watermarking. *Tongxin Xuebao (Journal on Communications), 26*(2), 100–111.

Liu, S., Wang, S., Deng, R., & Shao, W. (2005). A block oriented fingerprinting scheme in relational database. *Information Security and Cryptologyy–ICISC 2004* (pp. 455-466). Berlin Heidelberg: Springer.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data, 1*, 3.

Samarati, P. (2001). Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering, 13*, 1010–1027.

Schrittwieser S., Kieseberg P., Echizen I., Wohlgemuth S., & Sonehara N. (2011a). *Using generalization patterns for fingerprinting sets of partially anonymized microdata in the course of disasters*, RISI 2011.

Schrittwieser S., Kieseberg P., Echizen I., Wohlgemuth S., Sonehara N., & Weippl E. (2011b). *An Algorithm for k-anonymity-based Fingerprinting*, IWDW 2011.

Seo, J., Jin, M., Lee, S., Jang, D., Lee, S., & Yoo, C. (2005). Audio fingerprinting based on normalized spectral subband centroids. *International Conference on Acoustics, Speech, and Signal Processing, 3*, 213–216.

Sion, R., Atallah, M., & Prabhakar, S. (2002). *Watermarking relational databases*.

Su, K., Kundur, D., & Hatzinakos, D. (2002). *A novel approach to collusion-resistant video watermarking*. Proceedings of SPIE (4675).

Su, K., Kundur, D., & Hatzinakos, D. (2005). Statistical invisibility for collusion-resistant digital video watermarking. *IEEE Transactions on Multimedia, 7*(1), 43–51.

Sweeney, L. (2002a). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 10*(5), 571–588.

Sweeney, L. (2002b). *Comments to the Department of Health and Human Services On "Standards of Privacy of Individually Identifiable Health Information."*

Sweeney, L., et al. (2002). k-anonymity: a model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems, 10*(5), 557–570.

Trappe, W., Wu, M., Wang, Z., & Liu, K. (2003). Anti-collusion fingerprinting for multimedia. *IEEE Transactions on Signal Processing, 51*(4), 1069–1087.

Willenborg, L., & De Waal, T. (1996). *Statistical disclosure control in practice*. New York: Springer Verlag.

Willenborg, L., & Kardaun, J. (1999). *Fingerprints in microdata sets.* Joint ECE/Eurostat Work Session on Statistical Data Confidentiality.

Wu, M., Trappe, W., Wang, Z., & Liu, K. (2004). Collusion-resistant fingerprinting for multimedia. *IEEE Signal Processing Magazine, 21*(2), 28–39.

Zhang, Z., Jin, X., Wang, J., & Li, D. (2004). Watermarking relational database using image. *International Conference on Machine Learning and Cybernetics, 3*, 1739–1744.