Acta Univ. Sapientiae, Informatica, **6,** 1 (2014) 33–55

DOI: 10.2478/ausi-2014-0017

Survey on privacy preserving data mining techniques in health care databases

Tamás Zoltán GÁL
U1 Research
2 INFOPARK Gábor Dénes
Budapest, Hungary
email: galt@u1research.org

Gábor KOVÁCS
U1 Research
2 INFOPARK Gábor Dénes
Budapest, Hungary
email: kovacsg@u1research.org

Zsolt T. KARDKOVÁCS
U1 Research
2 INFOPARK Gábor Dénes
Budapest, Hungary
email: kardkovacs@u1research.org

Abstract. In health care databases, there are tireless and antagonistic interests between data mining research and privacy preservation, the more you try to hide sensitive private information, the less valuable it is for analysis. In this paper, we give an outlook on data anonymization problems by case studies. We give a summary on the state-of-the-art health care data anonymization issues including legal environment and expectations, the most common attacking strategies on privacy, and the proposed metrics for evaluating usefulness and privacy preservation for anonymization. Finally, we summarize the strength and the shortcomings of different approaches and techniques from the literature based on these evaluations.

Computing Classification System 1998: H.1.1. H.2.8. K.4.1.
Mathematics Subject Classification 2010: 97-02, 97R50, 97R40, 62P10, 68P20
Key words and phrases: anonymization, privacy preservation, health care database, survey, data mining, sensitive data
DOI:10.2478/ausi-2014-0017

1 Introduction

Databases that contain useful information about people have always been in the focus of research. Researchers apply various methods to extract valuable information from data sets to understand people and make predictions for their future behaviour. While legal systems may vary over countries, democratic regulations protect privacy at the highest level, usually in their constutition. Specially, a specific type of personal data called sensitive data, e.g. ethnicity, religious affiliation, medical condition, can only be accessed, transfered or handled by entities explicitly stated in regulations, and with the consent of the data subject.

Health care databases have an especially strict regulation because of the large number of sensitive data contained. For instance, pharmaceutical research must work with accurate data, but that retains all sensitive patient data as well, hence researchers working with such databases stumble very early in the legal limitations. Records of health care databases hold sensitive information from which one may be able to reveal medical condition of a person. Medical conditions may relate to e.g. food consumption preferences, life expectancy, drug taking habits, and other personal strengths or weaknesses. In wrong hands, e.g. decisions over employments [27], or mortgages might depend on such information which would be very unethical to use, and it must be avoided at all costs. On the other hand, health care databases also serve as the basis for better health care services, drug developments, and cost efficiency which also are in the focus of public interests. Therefore before publishing any piece of information from the database, it has to go through an anonymization procedure to hide sensitive data. Hence researchers must be aware of the legal requirements, the methods applicable to meet these requirements and the level protection these techniques provide.

One may take into account that neither well-known personal identifiers like birth name, social security number nor sensitive data on medical statements on their own harm privacy; only making connections between these pieces of information. That is, the main task called data anonymization is to prevent from establishing of connection between individuals and their data. Data anonymization can be forced by physically limit the data access by means of security policies, deletion, data perturbations, or by guaranteeing that any piece of data could be connected to more than one individual using repetition, sampling, aggregations, etc. As a consequence, data quality is reduced.

Reduced data quality for data analysis means somekind of loss in usefulness which directly affects the performance of data analyis. For example, data mining procedures as the most commonly used data analysis tools aim at discovering valid, previously unknown (novel or hidden), potentially useful, understandable (actionable) patterns, information, or relationships in statistically large databases [12]. Data mining tools highly dependent from data quality, i.e., poor data quality may result in invalid, useless, empty or non-comprehensible knowledge discovery. We propose a novel metric, the accuracy to be used for evaluating the usefulness of the anonymized data instead of information loss metrics.

The question arises how to enable the extraction of useful and beneficial information from health care databases while maximizing the protection of privacy. In this paper, we review different aspects of data mining related data anonymization and privacy preserving data mining, and we analyze the question from legalislative, privacy intruder, and data owner points of view. We also investigate what level of protection existing health care anonymization methods provide by comparing them to general techniques, and point out their limitations showing additional aspects to be covered when protecting health data.

The paper is organized as follows. Section 2 makes an overview of past events, which helps to give and overview on the importance and motivations of data anonymization. We discuss legal regulations and limitation regarding sensitive data management in Section 3. Section 4 defines key terms used in this article, and it presents a novel classification of the different approaches to data anonymization by analyzing possible attacking techniques and motivations. We make a short summary on the most analyzed data anonymization techniques and illustrate them on example databases in Section 5. Section 6 gives a brief introduction to theoretical indicators of privacy preservation and data utility metrics. We also evaluate how data anonymization techniques perform on our examples.

2 Leaking examples: threats and motivations

In theory, private data are the most protected, only those who have specific permission can access them. Nevertheless, in United States the 35% of the employers had a deep insight of health records of employees to make decisions about them according to [27] made in the 1990's [27]. There is also an urban legend [13] about a bank officer in Maryland who cross-referenced a list of patients with cancer against a list of people with recallable mortgages [27]. But health care practioners have many another ways to harm patients' privacy even after several regulations on the topic [6].

The problem is not new, the first census in the United States faced similar problems. For example, for electoral registry four typical data is given: sex, date of birth, postal code, ethnicity. In Cambridge, MA in 1997 there were 54.804 voters, and 12% of whom was uniquely identified by date of birth, 29% by date of birth and sex, 69% by date of birth and not full postal code, and 97% by date of birth and full length postal code [27]. Similar results are observed in other countries, as well [18, 10].

But this is not the only way to get sensitive data. In 2005, Netflix published a believed-to-be de-identified movie preference database of 480.000 customers. Database contained information on pseudo-anonymized ids, movie ids and titles, dates, and preferences. For those who had more than eight evaluations for movies there was a 99% match against publicly available IMDB (Internet Movie Database) evaluations [25]. Login names and real names are overlapping hence a huge number of customers could have been directly identified using these information. In 2006, AOL published another database on pseudo-anonymized web search queries. [4] found out that web search queries tend to contain the surfers' name, social security number, or other private information about them. In this article, they retrieved a picture of a 62-year-old web surfer believed to be anonym.

Obviously, storing, transferring, and handling private data is strictly forbidden in general. Nevertheless, it is unavoidable in many cases specially for public services. This implies that unencrypted or decrypted personal data can be accessed through hacking or other intrusion techniques.

For example, researchers fights for new discoveries as the basis of their promotions and livelihood. They are sometimes careless on handling properly personal data and it is not infrequent that data is not deleted after a research is completed. Can personal data be provided or be available for research? If ethical norms can be passed for some reasons there is no strict boundary which leads to gradual destruction and degradation of privacy norms [7].

Sometimes, if data subjects give specific permission to an organization to handle their private data could lead to privacy issues. For instance, customers allow a supermarket to handle their data on custom behaviors. Data mining on customer baskets, transactions are commonly used for data mining in order to reveal customer preferences, and to increase revenue with proper marketing communication and logistical strategies. Collected data may be shared legally with suppliers to help product development. In this case, data sharing may lead to revenue maximization on the supplier side only if data indicate clear preference over the competitive products. In addition, shared data based marketing strategy could increase the profit in all but the data sharer company [30].

People in democratic countries have a strong need for privacy which must be protected at all costs. Researchers and other data consumers need a uniformly available dataset for innovation, and obviously a clear, ethical way to acquire those valueble information. What motivates data publishers in the use data anonymization? Data sharing is essential in the information era for product development, innovation and research, i.e. information and databases are just another types of raw materials. Firstly, data sharers wants to comply with regulations in order to re-sell databases. Secondly, they want to avoid responsibility for non-intensional loss in privacy. Thirdly, they try to keep the balance between revenues of data sharing and the risk of misuse. Finally, there are also anti-trust worries, that is why data sharing among competitors is regulated.

These examples clearly illustrate that there is a strong need for data publishing, for protecting privacy and data providers in a hand.

3 Privacy, public interests and legal regulations

In order to understand, what may harm privacy, we need to defined the notion privacy itself. In democratic countries, people have rights to be left alone, their actions and data shall be handled confidentially unless other, higher or public interests do not require otherwise. From this point of view, every action against an individual's specific will, or which disturbs private life including unreasonable publicity harm privacy.

Nevertheless, boundaries of the term ,,public interest" may vary in countries and it is hard to say where begins or ends private and social life. That is why interpretation of the right to privacy is not straightforward. It is a fact that birthday is a private event while wedding, hence marital status, is not because there are public interests about this information, e.g. in case of conflicts of interest, post incompatibility, bigamy. The right to health, and the right to research are often constitutional rights as well and as such they are part of the public interests, which may interfere with and limit the right to privacy in a reasonable and a proportional way. Health information are sensitive data, i.e., data which must be protected for privacy and they can only be handled according to the the regulations of records. That is, the right to health may require detailed analyis of health data while the right to privacy explicitly forbids the access of data out of medical condition related health care services.

Legislative solution for this problem is data anonymization. What does data anonymization mean? In United States, the HIPAA (Health Insurance Portability and Accountability Act), in the European Community the EC/95/46 directive regulates data publishing, hence data anonymization policies. According to EC/95/46 directive, data is anonymized if data subject is no longer identifiable and retained in a form in which identification of the data subject is no longer possible[1]. HIPAA's approach is slightly different: anonymized data „does not identify an individual and if the covered entity has no reasonable basis to believe it can be used to identify an individual"[2]. The Hungarian jurisdiction defines anonymization as „a technical procedure that ensures connections between data and data subjects are no longer possible"[3]. Definition 1 extends legal definitions by combining their different approaches to the problem. The definition means informally that any Φ transformation is an anonymization function for a given database Δ, if the probability of finding a Ψ inverse transformation, which may use background knowledge not present in the original database, is close to zero.

Definition 1 (Data Anonymization) *Let $\Phi : \mathcal{DB} \to \mathcal{DB}$ be a database transformation function, and $\Delta \in \mathcal{DB}$ be a database. We say the transformed database $\Phi(\Delta)$ is anonymized if*

$$\forall\Psi\forall\Delta_1 \ldots \forall\Delta_N \quad \Pr(\Psi(\Phi(\Delta), \Delta_1, \ldots, \Delta_N) \not\subseteq \Delta) \approx 0, \tag{1}$$

where $\Psi : \mathcal{DB}^{N+1} \to \mathcal{DB}$ ($N = 0, \ldots, \infty$) is an arbitrary function, and Δ_i, $i = 1, \ldots, N$ are database representations of all available background knowledge. If $\Phi(\Delta)$ is anonymized for any $\Delta \in \mathcal{DB}$ then we say Φ is a data anonymization function.

Legal regulations explicitly state what is allowed and what is forbidden, but with the exception of HIPAA's Safe Harbour method they do not define how to achieve anonymity. Safe Harbour is about to pseudonymize or completely remove the following values in databases: name; date different from years including death, birth, discharge dates, etc., ages above 89 years; fax numbers; social security numbers; medical record numbers; health plan beneficiary numbers; account numbers; certificate/license numbers; vehicle identifiers and serial numbers, including license plate numbers; device identifiers and serial numbers; URLs; IP addresses; biometric identifiers; full-face or comparable photos or images; and any other unique identifying number, codes. In

[1] Article 26 of EC/95/46

[2] Section 164.514(a) of HIPAA

[3] Act of 2003/III. 1.§.(2) and Act of 1995/CXIX. 2.§.(1)

other words, Safe Harbour requires to eliminate direct and indirect identifiers, and has no advice on unintentional or partial, i.e. non-unique identifiers. In Section 5, we determine the level of anonymity Safe Harbour provides, and show its relation to our definition.

4 Approaches to data anonymization

Section 3 points out that data anonymization is a challenge for research, business and even for regulatory activities. Anonymization is a congitive process, practitioners must understand what could lead to the identification of an individual besides the obvious; direct data access either phyisically or logically, carelessness, etc.

Each person has some natural identifiers, i.e. data which characterizes an individual; name, social security number, passport number, cellular phone number, vehicle plate number, biometric identifiers etc. Some of them may not identify a person uniquely, but in proper contexts they shall be assumed to be unique identifiers. In this paper, we use the term direct identifiers for these kinds of data. There are a set of natural identifiers called indirect identifiers which together provide a unique identification, e.g. birth date, mother's name, address. One must notice that personal data and data which enable identification of a human being are not necessarily different things. For example, thoughts, forms of expression, activities, friends, medical case history, etc., may also identify people as well; we call them unintentional identifiers.

In practice, direct and indirect identifiers are replaced with one-way hash functions, i.e., functions that cannot allow original data to be restored since they have no inverse. Such a non-reversible value replacement of direct identifiers is called de-identification method. De-identification not necessarily assumes the complete removal of all but direct identifiers. If a de-identification method maps every identical direct/indirect identifier into an identical (but non-reversible) value then it is called pseudo-anonymization or pseudonymization for short. Later on this paper, we use the term re-identification for a procedure or method which processes one or more datasets to determine the identify of data subjects.

To protect privacy one must understand the potential threats, i.e. possible re-identification strategies:

- Direct re-identification. Data themselves without any further action reveal data subject identity.
- Re-identification through linking. Sometimes, data set is believed to be

de-identified while using publicly available or legally accessible databases enables re-identification of data subjects. For example, Netflix prize award data set contained pseudonymized user ids and movie ratings. Netflix ratings were easily correlated to IMDB ratings [25] where user ids are often personal names; re-identification was made possible through linking the preferences.

- Publishing anonymization algorithms, or settings for predictive algorithms. Publishing always makes ways to de-construct or to invert applied functions using direct re-mapping, guessing, etc. If we are looking for an information on an individual, one may deduce the future medical condition using medical predictive function based on their observed symptoms.
- Re-identification through extremities. Outlier values, rare or very unusual behavior are specific by definition to a very limited number of individuals, which may lead to re-identification. For example, if inhabitants of a small town suffer from the same disease, it is easy to infer the medical condition of an individual from that town.
- Background knowledge based re-identification. Sometimes, not structured, not stored, or single fact or knowledge known or accessed by a limit number of individuals is applied to retrieve someone's identity, e.g. custom habits of neighbors when and how they leave, or activities and photos published on a social network portal. Background knowledge is one of the most probable attacking strategy in our social network era.
- Re-identification through event sequencing. Frequencies or the ordering of data items also may uniquely identify certain individuals. A company based on a sick-leave registry may easily reveal employee medical condition if published health care database contains only dates and medical conditions.
- Information misuse. We are talking about information misuse whenever published database makes alternative, possibly harmful use possible. See, for example, the problem of sharing customer transactions in Section 2.

Anonymization techniques against these attacks get more complex in the order above. A question may arise: is it possible to eliminate all threats by a data anonymization algorithm while data still have enough value for data analysis? The following strategies have been applied for anonymization in the literature [30, 1, 31]:

1. Access limitations.

   (a) Limitation of data access. The most common procedure to limit the number of queries and to be run over a controlled environment through proper authentication and information hiding [24, 23].
   (b) Ciphering algorithms. Change of data values in a way that makes impossible to retrieve original values.

2. Obfuscation. These algorithms cut or aggregate parts of the database in order to avoid re-identification.
   (a) Dynamic sampling. Limited number of data elements are published, which meet the functional anonymization criteria.
   (b) Aggregation oriented anonymization. Enforcing data aggregations and micro-aggregation (aggregation over a subset of data elements) to achieve functional anonymization [29].

3. Functional anonymization. It aims at reducing the confidence about a piece of information related to a specific individual. It is widely discussed in the literature, there are several approaches to achieve this goal, e.g. by adding random noise [3, 2], using random data permutation [11], or by redundancy oriented data perturbations [29, 22, 14].

The first strategy is a very protective but not data publishing friendly solution. The second one is efficient for data publishing, however, they reduce dramatically data quality, which implies very limited use for data mining. Functional anonymization is the most discussed solution in the literature and it reveals the depth of data anonymization problems, as well. We make a brief overview on these techniques in Section 5.

5 Functional data anonynimization techniques

Let us consider Table 1 as a database to illustrate anonymization and related problems. The relation itself currently contains one sensitive information, the list of the salaries. Additionally, it can be directly connected with individuals as the ID attribute is present. Publishing such database could be strongly resisted with respect to data protection. Although this database is not

The easiest de-identification is omitting ID column; the result can be seen in Table 2. Note that, selecting any two of the *Date of birth*, *Sex* and *Postal code* attributes can identify the set of individuals in that relation. In general, these attribute pairs are not enough for unique identification, however,

ID	Date of birth	Sex	Postal code	Salary
Annie	21-01-76	Female	1107	40 000
Bill	24-03-76	Male	1107	45 000
Cecile	27-02-76	Female	1117	50 000
Dennis	21-01-76	Female	1117	55 000
Elise	24-03-76	Male	1127	60 000
Fred	27-02-76	Male	1127	65 000

Table 1: Contents of the `basictable` relation

in different countries they can identify a very large percent of the whole population [29, 18, 10], i.e. they quasi identify people. In hungarian a health care database, an individual who lives in a town with less than 50000 inhabitants can be identified with this triple with 94.2% probability. With more than 50000 inhabitants this value falls to 40.4%, which is still an unacceptably high value.

Definition 2 ([29] Quasi-identifier) *Given a set of individuals* I *and relation* $r(R)$ *on the* $R(A_1, \ldots, A_n)$ *schema, and let* $f_c : t[Q_r] \rightarrow r(R)$ *and let* $f_d : r(R) \rightarrow I'$*, where* $I' \subseteq I$*,* $t \in r(R)$ *and* $Q_r \subseteq \{A_1, \ldots, A_n\}$*.* Q_r *is a quasi-identifier in relation* $r(R)$*, if* $\exists p_i \in I$ *such that* $\exists t_i \in r[R]$ *for which* $f_d(f_c(t_i)) = p_i$*.*

Sweeney's definition means informally that any tuple $t[Q_r]$ of a relation r is a quasi-identifier in that relation, if the subset of attributes Q_r is unique for some individuals p_i.

Date of birth	Sex	Postal code	Salary
21-01-76	Female	1107	40 000
24-03-76	Male	1107	45 000
27-02-76	Female	1117	50 000
21-01-76	Female	1117	55 000
24-03-76	Male	1127	60 000
27-02-76	Male	1127	65 000

Table 2: Contents of the `de-identified` relation

By using background knowledge on quasi-identifiers and by having information about individuals from public sources, researchers can join records on quasi-identifiers to the sensitive data items. A solution for this problem is to make data values ambiguous. Either one can delete some of the quasi-identifiers and/or sensitive data values as shown in Table 3 or one can add noise to data values as shown in Table 4.

Date of birth	**Sex**	**Postal code**	**Salary**
*	*	1107	40 000
24-03-76	Male	1107	45 000
27-02-76	Female	1117	50 000
*	*	*	*
24-03-76	Male	1127	60 000
27-02-76	Male	1127	65 000

Table 3: Contents of the `deleteddata` relation

If one deletes some values then it can be overwritten with any (other) value to inhibit data linking through external sources. This means, that if one finds a possible data re-connection through quasi-identifier values, one cannot be certain that quasi-identifier values or data linking restore any part of the original database. On the other hand, a clear disadvantage of this approach is that the usability for analysis is degrading fast with the number of data perturbations.

Date of birth	**Sex**	**Postal code**	**Salary**
21-01-76	Female	1107	50 000
24-03-76	Male	1107	35 000
27-02-76	Female	1117	50 000
21-01-76	Female	1117	55 000
24-03-76	Male	1127	65 000
27-02-76	Male	1127	60 000

Table 4: Contents of the `noisytable` relation

Another way is to add noise to sensitive data. In this case, the attacker cannot be certain about the real data value in any but all particular record. This solution inhibits exploring sensitive data thus linking external data sources provide no further information. As a special case noise can be added by using microaggregation on data to lower the possibility of re-identification as shown in Table 5. Nevertheless, noise specially aggregations significantly reduce data quality.

Date of birth	**Sex**	**Postal code**	**Salary**
--76	Female	11**	40 000
--76	Male	11**	45 000
--76	Female	11**	50 000
--76	Female	11**	55 000
--76	Male	11**	60 000
--76	Male	11**	65 000

Table 5: Contents of the `aggregated` relation

The concept of k-anonymity limits the applicability of attack using external relationships. Instead of identifying individuals, for any key or keylike attribute there must be at least k with the same quasi-identifier in the database. This is usually achieved by generalizations, for instance deleting some numbers from an IP address or a postal code.

Definition 3 ([29] k-anonymity) *Given a relation $r(R)$ over the schema $R(A_1, \dots, A_n)$, and Q_r is a quasi-identifier in $r(R)$, then $r(R)$ is k-anonym if for any Q_r $t[Q_r]$ value occurs at least k times in $r(R)$.*

The relation in Table 5 is 3-anonym. In this case, the most concrete knowledge of a record necessarily involves the uncertainity that for that record there are at least 3 other candidates. However, in general it is quite difficult to determine about a relation whether it is k-anonym.

The computational complexity has been shown to be at least of the order of $O(2^{|Q_r|})$ [5, 28] independently from whether the model allows deletion or not. If deletion, local rewrite and multidimensional partitioning is allowed, then finding the minimal k-anonym is NP-hard [19, 20]. Generally, it has $O(n^{2k})$ complexity, but an $O(n \log n)$ approximation has also been proposed with certain restrictions, assuming multidimensional clustering [18].

[22] has shown that k-anonymity is not sufficient hence re-identification is also possible through sensitive data linking as well. In other words, not only the entropy of quasi-identifiers, but the entropy of sensitive values should exceed a particular threshold. The database shown in Table 6 is 2-anonym. Note that,

there are no further constraints on sensitive data values, which are identical for different birth dates. As a consequence, one can easily reveal sensitive data in Table 6 database by knowing only birth dates.

Date of birth	Sex	Postal code	Salary
21-01-76	*	11**	40 000
24-03-76	*	11**	45 000
27-02-76	*	11**	50 000
21-01-76	*	11**	40 000
24-03-76	*	11**	45 000
27-02-76	*	11**	50 000

Table 6: Contents of the `diversity` relation

Definition 4 ([22] l-diversity) *Given relation* $r(R)$ *over the schema* $R(A_1, \ldots, A_n)$*, the relation* $r(R)$ *is* l*-diversive, if for any attribute* A_i *to be protected at least* l *different* A_i *values are assigned to any particular* $t[R \setminus \{A_i\}]$ *value. Formally* l*-diversity exists, if*

$$-\sum_{\nu \in t[A_i]} p(t,\nu) \log p(t,\nu) \geq \log l, \tag{2}$$

where

$$p(t,\nu) = \frac{|t'[R \setminus \{A_i\}] = t[R \setminus \{A_i\}] \wedge t'[A_i] = \nu|}{|t'[R \setminus \{A_i\}] = t[R \setminus \{A_i\}]|}, \tag{3}$$

where p *is the apriori probability of* ν *value.*

It's important to see that l-diversity implies k-anonymity using $k = l$. According to Definiton 4 the 2-anonym relation shown in Table 7 is 2-diversive at the same time. Computational complexity of l-diversity is greater than the computation complexity of k-anonymity as additional attributes have to be handled.

In addition, re-identification threats are still alive on an l-diversive microaggregated Table [21]. Consider the database represented in Table 8, which is 2-diversive. The exact sensitive data cannot be retrieved by linking quasi-identifiers, however, the difference between data values within the same quasi-identifier determined cluster is marginal. [21] therefore recommends to extend k-anonymity criterion with a diversity related constraint. If distance between

Date of birth	**Sex**	**Postal code**	**Salary**
--76	*	1107	40 000
--76	*	1107	45 000
--76	*	1117	50 000
--76	*	1117	40 000
--76	*	1127	45 000
--76	*	1127	50 000

Table 7: Contents of the `diversity2` relation

sensitive attribute values with the same quasi-identifier values and values of the entire relation are very different then the sensitive attribute values can be estimated with statistical probing. For simplicity, let us say that sensitive data values that share the same quasi-identifier values are in the same equivalence class. That is, there can be found several distributions of different equivalence classes.

Date of birth	**Sex**	**Postal code**	**Salary**
21-01-76	*	11**	40 000
24-03-76	*	11**	45 000
27-02-76	*	11**	50 000
21-01-76	*	11**	40 001
24-03-76	*	11**	45 001
27-02-76	*	11**	50 001

Table 8: Contents of the `tclosed` relation

The concept of `t`-closure investigates whether there is a `t` threshold, which is not exceeded by a distance measure.

Definition 5 ([21] `t`-closure) *An equivalence class is `t`-closed, if the distance between the distribution of the sensitive data within that class and the distribution of the entire relation within that class does not exceed a `t` threshold. The relation is `t`-closed, if any equivalence class contained is `t`-closed.*

Note that the definition does not define the distance function to be used; that is, it can be applied for various data types including textual, categorical, etc.

It is practical to boost uncertainity with permutation of sensitive data instead of modifying them. It has been already mentioned before that aggregation significantly degrades usability, based on Table 7 one can deduce almost nothing from the data. A permutation approach is to put sensitive data with the same quasi-identifier into hash buckets, then iteratively replacing the current value with one from the bucket with the highest cardinality [32]. Another approach [33] proposes the ordering of sensitive data and selecting the candidate for permutation from an e-wide interval with at least k cardinality.

Permutation provides the same level of privacy preservation as generalization, however, aggregate values are accurate in this case. For instance, the salary of individuals who work in a certain field, or were born in a certain year is a valid and usable value. On the other hand, permutation changes dramatically sensitive data and thus their hidden patterns, which leads to a completely different or alternative result after an analysis. However, if there is only one sensitive data attribute, permutation within the same equivalence class does not pose this problem.

Date of birth	**Sex**	**Postal code**	**Salary**
21-01-76	Female	1107	45 000
24-03-76	Male	1107	40 000
27-02-76	Female	1117	50 000
21-01-76	Female	1117	55 000
24-03-76	Male	1127	65 000
27-02-76	Male	1127	60 000

Table 9: Contents of the `permuted` relation

After reviewing theoretical anonymization techniques, we can see that the Safe Harbour method of HIPAA does provide k-anonymity according to its pseudonymization procedure as it eliminates all possible quasi-identifiers, however it does not modify data records themselves hence it can neither provide l-diversity, nor give protection against t-closure based probing. This means that additional measures have to be completed even after the Safe Harbour method to achive the level of anonymity required by the legal environment.

6 Metrics for privacy and data utility

Ethical data mining aims to meet legal requirements to meet privacy and eliminate stereotype conclusions. At the same time, any change of data decreases the efficiency of data mining. Is privacy measureable?

Let $\mathrm{PP}_i : \mathcal{DB} \to [0,1] \in \mathbb{R}$ be a membership function that maps to every database a measure which proportional to privacy preservation, where 1 stands for complete privacy protection, and 0 for no privacy. Assume that $\mathrm{PP}_i(\emptyset) = 1$.

Let $\mathrm{RV}_i : \mathcal{DB} \times \mathcal{A} \to [0,1] \in \mathbb{R}$ denote the relative usefulness (fitness, accuracy, etc.) measure of applying data mining model on a database. The value 0 indicates that the data cannot be used for data mining purposes, and 1 stands for model application is the best one can achieve on the database. Assume that $\mathrm{RV}_i(\emptyset) = 1$. Let κ, ν denote the acceptable threshold for privacy preservation measure, and relative usefulness, respectively. Anonymization as an optimization problem for a database $d \in \mathcal{DB}$, and for a data mining model a can be formulated as follows:

$$\exists ?\Phi : \quad \kappa \leq \mathrm{PP}_i(\Phi(d)) \ \wedge \ \nu(a) \leq \mathrm{RV}_i(\Phi(d), a), \tag{4}$$

where $\Phi : \mathcal{DB} \to \mathcal{DB}$ is data anonymization method. We call an anonymization method Φ optimal for a database d and a data mining model a denoted by $\hat{\Phi}_{d,a}$, if and only if

$$1 = \mathrm{PP}_i(\hat{\Phi}_{d,a}) \ \wedge \ 1 = \mathrm{RV}_i(\hat{\Phi}_{d,a}(d), a), \tag{5}$$

For research studies the following questions arise:

- Is there an optimal data anonymization for a given database and a given data mining model?
- Is there an optimal data anonymization for any given database and data mining model tuple?
- Is there an optimal data anonymization for a given database independently for all data mining model?
- Is there an optimal data anonymization for all database and all data mining model?

While research focuses on the last two questions, the former two may suffice for industry applications.

6.1 Privacy measurements

Note that, this paper does not introduce how to calculate PP or RV. In the literature, there are several measures depending on the different aspects of the problem. The simplest indicator is the ratio of the number of the identifiable individuals and the number of the total individuals in the database [27].

In case of data perturbations, privacy can be described with an $H(A|B)$ conditional entropy variable, where A is the remaining protection after B has been published [2]. Hence, the probability of correct data is leaking out is

$$\Pr(A|B) = 1 - \frac{2^{H(A|B)}}{2^{H(A)}} = 1 - 2^{-I(A;B)}, \tag{6}$$

where $I(A;B) = H(A) - H(A|B)$ is the mutual information between variables A and B. Similarly, [16, 8] adapt Shannon entropy to describe privacy:

$$\Pr(A|B) = 2^{-\int f_{A,B}(a,b)\log_2 f_{A|B=b}(a)\,da\,db}. \tag{7}$$

The two metrics above are essentially indentical. [26] simplifies these equations so that privacy is defined with variance:

$$\Pr(A|B) = \frac{\mathrm{Var}(A-B)}{\mathrm{Var}(A)}. \tag{8}$$

The size of potential information leakage can be bounded using matrices. Let $[d_{ij}]$ be a Boolean matrix representing an initial database containing the apriori probabilities $p_0^{ij} = \Pr[d_{ij} = 1]$. Once an adversary asks Q queries to the anonymized database as above, and all other values of the database are provided, we can define the posterior probability p_Q^{ij} of d_{ij} taking the value 1. The change in belief is $\Delta = |c(p_Q^{ij}) - c(p_0^{ij})|$, where $c(x) = \log(x/(1-x))$ is a monotonically increasing function.

This model can be generalized by approximating $d_{ij} = 1$ with the Boole-function $f(d_{i1}, \ldots, d_{ik}) = 1$ with k arguments. In this case, the for any query Q and all function f Δ shall be minimal. It has been shown that $O(\sqrt{Q(n)}/\delta)$ changes of data is sufficient for protection [1, 9], where δ is the maximum change allowed, and $Q(n)$ is the number of queries executed on the database of size n.

The efficiency of privacy for a data set on i individuals with p_i public, s_i sensitive, u_i unpublished data that can potentially be used for identifying s_i can be interpreted as follows. Let C be a classifier on p_i and u_i data items, and let C_1 be a classifier built on the published data $t_i =< p_i, s_i >$ with a_1

accuracy. Data protection damaged if for any C_2 classifier with accuracy a_2 that has access to C with regard to the t_i data set $a_1 < a_2$ [15]. The same statement is defined in [17] by means of the distance function δ over probability distributions such that

$$\delta(\mathrm{Pr}_t(s_i = a), \mathrm{Pr}_t(s_i = a|\mathbb{A}(\mathcal{DB}))) < \kappa, \tag{9}$$

where $\kappa \in \mathbb{R}$ and $\mathrm{Pr}_t(s_i)$ is the apriori probability that the sensitive data item s_i has a specific value. If this statement holds, then the database $\mathbb{A}(\mathcal{DB})$ is anonym for the data item s_i.

6.2 Indicators of usability and loss of information

From the point of usability, three relevant aspects of data mining has to be considered: accuracy, completeness and consistency [8]. Accuracy is a measure of difference between the original and anonymized data items. Completeness is the amount of data omitted in the process. Consistency measures the maintainability of inner relationships.

Accuracy can be defined as the difference between the real and the modified information by calculating the loss from the frequency of relative errors, formally:

$$\Delta(r, r') = \frac{\sum\limits_{i=1}^{n} |f_r(i) - f_{r'}(i)|}{\sum\limits_{i=1}^{n} f_r(i)}, \tag{10}$$

where i is a data item and $f_r(i)$ is the frequency of occurence of that data item in the r relation. This formula is sufficiently general to cover not only data modification, but data omission and the generation of new records.

When the anonymization process uses microaggregations, then the loss can be modelled [28] alternatively with

$$\Delta(r, r') = \frac{\sum\limits_{A \in R} \sum\limits_{t \in r} \frac{h}{|DOM(h_A)|}}{|r| \cdot |A|}, \tag{11}$$

where relation r fits the schema $R(A_1, \ldots, A_n)$ and A is an attribute in that schema, $|A| = n$ is the number of the attributes in R, t is a record in r, $|r|$ is the size of the relation, and h means the measure of the microaggregation in the domain $DOM(h_A)$, where $DOM(h_A)$ is the domain of a possible hierarchy levels of an attribute A. Note that we can not assume the omission of data items in this case, the size of protected and anonynimized databases are essentially the same. Also note that this formula does not calculate the concrete measure of loss, it is only proportional to the possible microaggregation levels, and additionally determining the size of the hierarchy is very promiscuous.

This concept can be generalized by taking interval frequencies used in the tranformations into account:

$$\Delta(r, r') = \frac{\sum\limits_{A \in R} \sum\limits_{t \in r} \frac{f(t[A])-1}{g(A)-1}}{|r| \cdot |A|}, \tag{12}$$

where function f returns the number of different $t[A]$ values from relation r that can be mapped to the aggregated $t'[A]$ value from relation r', and g gives the size of the codomain of A attribute. Note that, neither equations above considers that changing two attributes properly the original value can be restored, and whether there is a loss of information after the aggregation.

An alternative definition for accuracy is based on distribution functions [2]. Given two functions, f and g, where the domain of f is the original database and the domain of g is the anonymized database, and the codomains of them are the same. In this case the loss of information can be described with the measure of mutual information:

$$I(f, g) = \frac{1}{2} E \left[\int |f(t) - g(t)| dx \right]. \tag{13}$$

	$H(A\|B)$	$Pr(A\|B)$	$\Delta(r, r')$
r' =Table 2	0.232	80.4%	0
r' =Table 3	0.289	79.7%	0.00133
r' =Table 4	0.309	82.0%	0.00102
r' =Table 5	0.500	76.4%	0.00412
r' =Table 6	1.057	76.9%	0.00566
r' =Table 7	1.057	76.9%	0.00560
r' =Table 8	0.528	76.0%	0.00257
r' =Table 9	0.232	80.4%	0

Table 10: Data protection and information loss numerically using $A =$ salary, $B = \{\text{date of birth}, \text{sex}, \text{postal code}\}$, and $r =$ Table 2

Table 10 summarizes the different metrics for sample databases presented in this paper. It is easy to see that some of the quality metrics do not take into account quality changes using non-aggregating form of data replacements. For example, permutation method seems to keep data quality which is only true for calculating aggregating on equivalence classes. Moreover, it cannot capture small semantic distance between different values, that is why t-closure seems to have lower privacy preservation capability than k-anonymity. A comparable, valid measurement for privacy or for data utility is still missing from the literature.

7 Conclusions

This paper presents in very brief the depth, importance, and issues of data anonymization. Data anonymization is one of the most imminent problems which directly affects basic rights like the right to privacy, health, or research, the innovation policy of organizations, and last but not least, the future computing environments. Deconstruction and data linking are always possible in our information era, but inhibiting such threats by data perturbation does not help to exploit the values stored deeply inside the databases, specially health care databases.

We demonstrated by examples and equations that maintaining data quality for data mining and preserving privacy has some limitations; there are too many possibilities to reveal private information using public databases or background knowledge. We showed that legislation covers just parts of the problem by stating what but how to do. Therefore, preserving privacy data mining on health databases presents a new challenge for information engineers, hence the possible number of linkable data is dramatically increasing, e.g. using photos, personal data, activities, hobbies published in social networks.

We presented a well-known set of functional anonymization methods, which aims at perturbing data only the least possible. We demonstrated that these methods reduce data quality significantly. In order to measure such an information loss and privacy we briefly outlined the most discussed evaluation metrics. This paper proved that those metrics cannot capture semantical differences, which makes different anonymization methods incomparable. Open questions still remain for the future: do there exist an optimal data anonymization method for all databases and data mining models, how to compare different data anonymization methods, and how to measure the possible information leakage of a published real-life database assuming excessive usage of public databases and background knowledge.

Acknowledgements

Publishing of this work and the research project were funded by the European Union and co-financed by the European Social Fund under the name of „MEDICSPHERE – Complex, multipurpose, ICT technology driven medical, economic, and educational use of clinical data" and grant number TÁMOP-4.2.2.A-11/1/KONV-2012-0009.

References

[1] C. C. Aggarwal, P. S. Yu, An introduction to privacy-preserving data mining. in: *Privacy-Preserving Data Mining*, (Eds.: C. C. Aggarwal and P. S. Yu) chapter 1, pp. 1–9. Springer-Verlag, 2008. ⇒40, 49

[2] D. Agrawal, C. C. Aggarwal, On the design and quantification of privacy preserving data mining algorithms. *Proc. 20th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 247–255. ACM, 2002. ⇒41, 49, 51

[3] R. Agrawal, R. Srikant, Privacy-preserving data mining. *ACM Sigmod Record* **29,** 2 (2000) 439–450. ⇒41

[4] M. Barbaro, T. J. Zeller, S. Hansell, A face is exposed for aol searcher no. 4417749. *The New York Times*, 9 Aug. 1, 2006. ⇒36

[5] R. J. Bayardo, R. Agrawal, Data privacy through optimal k-anonymization. *Proc. 21st International Conference on Data Engineering*, ICDE '05, pp. 217–228, Washington, DC, USA, 2005. IEEE Computer Society. ⇒44

[6] D. Benatar, Indiscretion and other threats to confidentiality. *South African J. Bioethics and Law*, **3,** 2 (2010) 59–62. ⇒36

[7] J. J. Berman, Confidentiality issues for medical data miners. *Artificial Intelligence in Medicine* **26,** 1 (2001) 25–36. ⇒36

[8] E. Bertino, D. Lin, W. Jiang, *Privacy Preserving Data Mining*, chapter A survey of quantification of privacy preserving data mining algorithms, pp. 183–205. Springer, 2008. ⇒49, 50

[9] C. Dwork, K. Nissim, Privacy-preserving datamining on vertically partitioned databases. in: *Advances in Cryptology – CRYPTO 2004* pp. 134–138. Springer, 2004. ⇒49

[10] K. El Emam, D. Buckerdige, A. Neisa, E. Jonker, A. Verma, The re-identification risk of canadians from longitudinal demographics. *BMC Medical Informatics and Decision Making* **11,** 46 (2011) 1–12. ⇒36, 42

[11] A. V. Evfimievski, Randomization in privacy preserving data mining. *ACM SIGKDD Explorations Newsletter* **4,** 2 (2002) 43–48. ⇒41

[12] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases. *AI Magazine* **17,** 3 (1996)37–54. ⇒35

[13] R. Gellman, The story of the banker, state commission, health records, and the called loans: An urban legend?, `http://bobgellman.com/rg-docs/rg-bankerstory.pdf` 2011. ⇒35

[14] A. Gionis, A. Mazza, T. Tassa, k-anonymization revisited. *Proc. IEEE 24th Int. Conf. Data Engineering ICDE 2008*, pp. 744–753, 2008. ⇒41

[15] M. Kantarcıoğlu, J. Jin, C. Clifton, When do data mining results violate privacy? *Proc. 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA*, pp. 599–604, 2004. ⇒50

[16] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, On the privacy preserving properties of random data perturbation techniques. *Third IEEE International Conference on Data Mining, 2003. ICDM 2003*, pp. 99–106. IEEE, 2003. ⇒49

[17] D. Kifer, Attacks on privacy and deFinetti's theorem. *Proc. 2009 ACM SIGMOD International Conference on Management of Data*, pp. 127–138. ACM, 2009. ⇒ 50

[18] M. R. Koot, G. van't Noordende, C. de Laat, A study on the re-identifiability of dutch citizens. in: *Workshop on 3rd Hot Topics in Privacy Enhancing Technologies, HotPETs 2010*, 2010. ⇒36, 42, 44

[19] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Incognito: Efficient full-domain k-anonymity. *Proc. 2005 ACM SIGMOD International Conference on Management of Data*, pp. 49–60. ACM, 2005. ⇒44

[20] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Mondrian multidimensional k-anonymity. *Proc. 22nd International Conference on Data Engineering, 2006. ICDE'06.*, pp. 25–25. IEEE, 2006. ⇒44

[21] J.-L. Lin, J. Y.-C. Liu, Privacy preserving itemset mining through fake transactions. *SAC*, pp. 375–379, 2007. ⇒45, 46

[22] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data* **1,** 1(2007) 3–3. ⇒41, 44, 45

[23] G. Miklau, D. Suciu, A formal analysis of information disclosure in data exchange. *Proc. 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, pp. 575–586, New York, NY, USA, 2004. ACM. ⇒41

[24] M. Miller, J. Seberry, Relative compromise of statistical databases. *Austral. Computer J.* **21,** 2 (1989) 56–61. ⇒41

[25] A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets. *Proc. 2008 IEEE Symposium on Security and Privacy*, SP '08, pp. 111–125, Washington, DC, USA, 2008. IEEE Computer Society. ⇒36, 40

[26] S. R. M. Oliveira, O. R. Zaïane, A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration. *Computers & Security* **26,** 1 (2007) 81–93. ⇒49

[27] L. Sweeney, Datafly: A system for providing anonymity in medical data. *Proc IFIP TC11 WG11.3 Eleventh International Conference on Database Securty XI: Status and Prospects*, pp. 356–381, 1997. ⇒34, 35, 36, 49

[28] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression. *Intern. J. Uncertainty, Fuzziness and Knowledge-Based Systems* **10,** 5 (2002) 571–588. ⇒44, 50

[29] L. Sweeney, k-anonymity: A model for protecting privacy. *Intern. J. Uncertainty, Fuzziness and Knowledge-Based Systems* **10,** 5 (2002) 557–570. ⇒41, 42, 44

[30] J. Vaidya, Y. Zhu, C. W. Clifton, Privacy preserving data mining. *Advances in Information Security* **19** (2006) 1–121. ⇒37, 40

[31] K. Wahlstrom, J. F. Roddick, R. Sarre, V. Estivill-Castro, D. de Vries, *Encyclopedia of Data Warehousing and Mining*, volume 2, chapter Legal and technical issues of privacy preservation in data mining, pp. 1158–1163. IGI Publishing, 2nd edition, 2008. ⇒40

[32] X. Xiao, Y. Tao, Anatomy: Simple and effective privacy preservation. *Proc. 32nd International Conference on Very Large Data Bases (VLDB)*, pp. 139–150. VLDB Endowment, 2006. ⇒47

[33] Q. Zhang, N. Koudas, D. Srivastava, T. Yu, Aggregate query answering on anonymized tables. *IEEE 23rd International Conference on Data Engineering, 2007. ICDE 2007*, pp. 116–125, Istanbul, Turkey, 2007. IEEE Xplore. ⇒47

Received: October 9, 2013• Revised: March 4, 2014