Journal of Biomedical Informatics 58 (2015) 37–48

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss

Florian Kohlmayer *[,1], Fabian Prasser [1], Klaus A. Kuhn

University of Technology Munich (TUM), Department of Medicine, Chair for Biomedical Informatics, 81675 München, Germany

ARTICLE INFO

Article history:
Received 15 June 2015
Revised 7 August 2015
Accepted 4 September 2015
Available online 15 September 2015

Keywords:
Security
Privacy
De-identification
Anonymization
Statistical disclosure control
Optimization

ABSTRACT

Objective: With the ARX data anonymization tool structured biomedical data can be de-identified using syntactic privacy models, such as k-anonymity. Data is transformed with two methods: (a) generalization of attribute values, followed by (b) suppression of data records. The former method results in data that is well suited for analyses by epidemiologists, while the latter method significantly reduces loss of information. Our tool uses an optimal anonymization algorithm that maximizes output utility according to a given measure. To achieve scalability, existing optimal anonymization algorithms exclude parts of the search space by predicting the outcome of data transformations regarding privacy and utility without explicitly applying them to the input dataset. These optimizations cannot be used if data is transformed with generalization and suppression. As optimal data utility and scalability are important for anonymizing biomedical data, we had to develop a novel method.

Methods: In this article, we first confirm experimentally that combining generalization with suppression significantly increases data utility. Next, we proof that, within this coding model, the outcome of data transformations regarding privacy and utility cannot be predicted. As a consequence, existing algorithms fail to deliver optimal data utility. We confirm this finding experimentally. The limitation of previous work can be overcome at the cost of increased computational complexity. However, scalability is important for anonymizing data with user feedback. Consequently, we identify properties of datasets that may be predicted in our context and propose a novel and efficient algorithm. Finally, we evaluate our solution with multiple datasets and privacy models.

Results: This work presents the first thorough investigation of which properties of datasets can be predicted when data is anonymized with generalization and suppression. Our novel approach adopts existing optimization strategies to our context and combines different search methods. The experiments show that our method is able to efficiently solve a broad spectrum of anonymization problems.

Conclusion: Our work shows that implementing syntactic privacy models is challenging and that existing algorithms are not well suited for anonymizing data with transformation models which are more complex than generalization alone. As such models have been recommended for use in the biomedical domain, our results are of general relevance for de-identifying structured biomedical data.

1. Introduction

Collaborative collection and sharing of sensitive personal data have become an important element of biomedical research. To protect patient privacy in complex research environments, a broad spectrum of safeguards must be implemented, including legal, contractual as well as technical methods. Data anonymization is a central building block in this context. It aims at sanitizing datasets in ways that prevent attackers from breaching the subjects' privacy. A number of incidents have shown that simply removing all directly identifying information (e.g. names) is not sufficient [1–3]. As a consequence, different definitions of privacy and techniques for sanitizing datasets have been proposed [4–7]. As sanitization inevitably leads to loss of information and thus a decrease in data utility, a balance has to be sought between privacy risks on one side and suitability for a specific use case on the other.

According to national laws, such as the US Health Insurance Portability and Accountability Act (HIPAA) [8], and international regulations, such as the European Directive on Data Protection [9], different methods may be used. In particular, the HIPAA Privacy Rule defines two basic methods for de-identifying datasets [10]. The first method requires the removal or the modification

* Corresponding author. Tel.: +49 89 41404345.
E-mail address: florian.kohlmayer@tum.de (F. Kohlmayer).
[1] Equal contributors.

http://dx.doi.org/10.1016/j.jbi.2015.09.007
1532-0464/

of a pre-defined set of attributes and attribute values. The second method, which is called *"expert determination"* requires that a professional *"determines that the risk is very small that the information could be used [...] to identify an individual"* [10]. For this purpose, methods of *statistical disclosure control* may be used.

In this work, we will focus on statistical disclosure control for structured data, which can be represented in a tabular form with each row corresponding to the data about one individual [6]. Specifically, we will describe methods implemented in ARX, a data anonymization tool that we have developed for the biomedical domain [11,12]. A typical use case is the de-identification of research data prior to sharing. To our knowledge, ARX offers the most comprehensive support of methods for anonymizing structured data to date. Its highlights include methods for risk analyses, risk-based anonymization, syntactic privacy models and methods for automated and manual analysis of data utility. Moreover, the tool implements an intuitive coding model, is highly scalable and provides a sophisticated graphical user interface with several wizards and visualizations that guide users through different aspects of the anonymization process.

1.1. Background

ARX implements methods that offer dynamic means for balancing privacy risks with data utility. Privacy requirements are expressed in the form of syntactic privacy criteria. Data is transformed with coding models, in particular generalization and suppression of attribute values, to ensure that they fulfill the specified privacy requirements. Risk models are used to estimate risks of re-identification, which are an inherent aspect of many privacy models. Finally, utility measures are used to estimate the suitability of the resulting datasets for specific usage scenarios. A balancing of privacy and utility is achieved through user feedback: by choosing different privacy models, risk estimates, transformation methods, and utility measures as well as by varying the parameters which regulate the different steps.

When anonymizing structured data, the general attack vector assumed is *linkage* of a sensitive dataset with an identified dataset (or similar background knowledge about individuals). The attributes that may be used for linkage are termed *quasi-identifiers* (or indirect identifiers, or keys). Such attributes are not identifiers per se but may in combination be used for linkage. Moreover, it is assumed that they cannot simply be removed from the dataset as they may be required for analyses and that they are likely to be available to an attacker. Furthermore, it is assumed that *directly identifying* information (such as names) has already been removed from the dataset. An example dataset with different types of attributes is shown in Table 1. The semantics of sensitive attributes will be explained in Section 2.1.

Datasets are often protected against *identity disclosure* (or *re-identification*), which means that an individual can be linked to a specific data entry [3]. This is a very important type of attack, as it has legal consequences for data owners according to many laws and regulations worldwide. Protection may be implemented with the *k-anonymity* privacy model [3]. A dataset is *k-anonymous* if, regarding the quasi-identifiers, each data item cannot be distinguished from at least $k-1$ other data items. This property can be used to define *equivalence classes* of indistinguishable entries [13,13a]. The output dataset from Table 1 fulfills 2-anonymity.

When data is anonymized, values of quasi-identifiers are transformed to ensure that the data fulfills privacy requirements. In ARX this data recoding is primarily performed with generalization hierarchies. Examples are shown in Fig. 1. Here, values of the attribute age are first transformed into age groups and then suppressed, while values of the attribute gender can only be suppressed. Diagnoses can be grouped by anatomy, nosology or etiology. Anatomy has been used in the example. Generalization hierarchies are well suited for categorical attributes. They can also be used for continuous attributes by performing categorization. In the example from Table 1, the attribute *age* is generalized to the first level of the according hierarchy.

To create anonymized datasets of high quality, ARX combines attribute generalization with the suppression of data records. This means that entries from equivalence classes that violate the privacy model (i.e. *outliers*) are automatically replaced with

Table 1
Example dataset and a privacy-preserving transformation. Age and *gender* are quasi-identifiers, *diagnosis* is a sensitive attribute. The attribute *age* has been generalized. The last two entries have been suppressed. The transformed dataset fulfills 2-anonymity regarding the quasi-identifiers and distinct-2-diversity regarding the sensitive attribute.

Quasi-identifying		Sensitive	Quasi-identifying		Sensitive
Age	Gender	Diagnosis	Age	Gender	Diagnosis
34	Male	Colon cancer	20–39	Male	Colon cancer
22	Female	Stroke	20–39	Female	Stroke
66	Male	Stroke	60–79	Male	Stroke
70	Male	Colon cancer	60–79	Male	Colon cancer
35	Female	Colon cancer	20–39	Female	Colon cancer
21	Male	Stroke	20–39	Male	Stroke
18	Female	Colon cancer	★	★	★
19	Female	Stroke	★	★	★

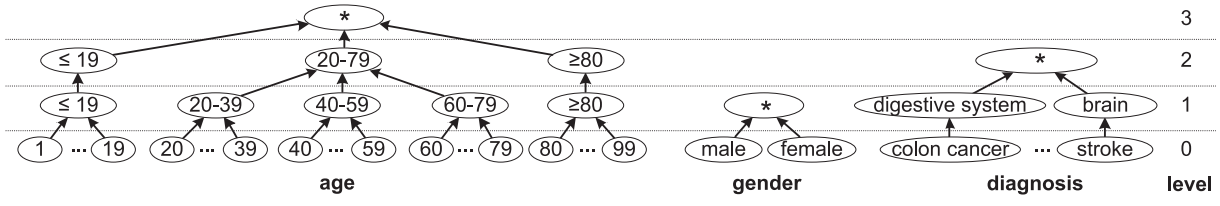

Fig. 1. *Generalization hierarchies for attributes age, gender and diagnosis.* Values of the attribute age are first transformed into age groups and then suppressed, while values of the attribute gender can only be suppressed. Diagnoses are grouped by anatomy.