Journal of Biomedical Informatics xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A flexible approach to distributed data anonymization

Florian Kohlmayer [a,*,1], Fabian Prasser [a,1], Claudia Eckert [b], Klaus A. Kuhn [a]

[a] Technische Universität München, University Medical Center (MRI), Ismaninger Strasse 22, 81675 München, Germany
[b] Technische Universität München, Department of Computer Science, Boltzmannstrasse 3, 85748 Garching bei München, Germany

ARTICLE INFO

Article history:
Received 27 August 2013
Available online xxxx

Keywords:
Personal data protection
Distribution
Privacy
Anonymization
Commutative encryption
Secure multi-party computing
SMC

ABSTRACT

Sensitive biomedical data is often collected from distributed sources, involving different information systems and different organizational units. Local autonomy and legal reasons lead to the need of privacy preserving integration concepts. In this article, we focus on anonymization, which plays an important role for the re-use of clinical data and for the sharing of research data. We present a flexible solution for anonymizing distributed data in the semi-honest model. Prior to the anonymization procedure, an encrypted global view of the dataset is constructed by means of a secure multi-party computing (SMC) protocol. This global representation can then be anonymized. Our approach is not limited to specific anonymization algorithms but provides pre- and postprocessing for a broad spectrum of algorithms and many privacy criteria. We present an extensive analytical and experimental evaluation and discuss which types of methods and criteria are supported. Our prototype demonstrates the approach by implementing k-anonymity, ℓ-diversity, t-closeness and δ-presence with a globally optimal de-identification method in horizontally and vertically distributed setups. The experiments show that our method provides highly competitive performance and offers a practical and flexible solution for anonymizing distributed biomedical datasets.

1. Introduction

Collaboration and data sharing have become core elements of biomedical research. Examples are international projects like the International Cancer Genome Consortium ICGC with its goal to "*make the data available to the entire research community*" [1], and BBMRI-LPC aiming "*to help scientists to have better access to large European studies on health*" [2]. Also, from the perspective of public funders, sharing of research data has become a request, and principles of sharing have been formulated [3,4]. Besides the international projects mentioned above, there are research projects on national, regional, and institutional levels, which collect, integrate, and share data.

The process of managing data from collection to analyses and also to sharing can be illustrated by different phases [5]. Research data is collected and managed, which may be accompanied by further processes, such as quality assurance. Sharing is initiated by allowing other researchers to get an overview over available data which fit their research objectives. Typically, access to core data is limited, and data access committees are involved before data use agreements (DUAs) are signed and data is released. Then, this data is integrated and used for new analyses.

There is a growing understanding of risks related to data sharing: disclosure of sensitive biomedical data may lead to harm for individuals, especially when different sources are available for linkage (for an overview see [6]). Basically, national laws and regulations, such as the HIPAA Privacy Rule [7], as well as international regulations, such as the European Directive on Data Protection [8], mandate stringent protection of personal data. In recent years, there has been extensive work on ethical, legal, social/societal issues (ELSI) of biomedical and genomic research and on data sharing, e.g., [9,10], which we will not further address in this article.

Anonymization is an important privacy measure when releasing and sharing sensitive datasets. As an important example, the HIPAA Privacy Rule has defined concrete measures to prevent re-identification. These include methods of statistical disclosure control. Basically, fuzziness is introduced to a degree which balances remaining semantics and usability against risk reduction. *K-anonymity* is a well known and understood privacy criterion, focusing on *quasi-identifiers*. These are attributes that are required for analyses but are associated with a high risk of reidentification. A dataset is k-anonymous if each data item cannot be distinguished from at least $k-1$ other data items regarding the quasi-identifiers [11]. Introducing k-anonymity is a measure against linkage attacks which may lead to identity disclosure when accessible data is combined with an attackers background knowledge [12].

Data is often collected from distributed sources, involving different types of data, different information systems, and different organizational units. Pseudonymity is another privacy measure of

* Corresponding author.
E-mail address: florian.kohlmayer@tum.de (F. Kohlmayer).
[1] These authors contributed equally to this work.

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.12.002

relevance, which leads to distribution. Here, directly identifying data is separated from medical data, and the links between identifiers and corresponding pseudonyms are secretly kept by a honest broker [13]. In general, data can be distributed vertically or horizontally. The former means that different sites hold different subsets of the attributes for a common set of individuals, so pseudonymity is a typical example. The latter means that different sites hold data with the same set of attributes for different individuals, for example, data for individuals in their region. Health services research is an example where integration of horizontally distributed data is needed, and disclosure has to be prevented.

In this article, we will focus on anonymization of datasets which are horizontally or vertically distributed. Existing approaches have focused on limited sets of privacy criteria, which in practice must often be combined with further criteria to prevent unintended disclosure of sensitive data. In most cases, specific algorithms were implemented which employ specific types of data transformations and search strategies. In contrast, we see are requirement for flexible solutions which allow the implementation of a broad spectrum of methods. Here, we agree with [14,15], that the suitability of methods depends on use cases. As efficient generic solutions do not exist, and as many approaches have unclear performance characteristics, we will also address performance questions. They are of relevance in situations which require near real-time updates, e.g. when the course of an infectious disease is analyzed over different areas.

1.1. Contributions

We will present a flexible and efficient approach to distributed data anonymization in the semi-honest model. It is based upon a secure multi-party computing (SMC) protocol, which constructs an encrypted global view out of horizontally or vertically distributed datasets. To this global view a broad spectrum of anonymization algorithms and privacy criteria can be applied. Thus, centralized versions of a large number of data anonymization algorithms are supported, and we will provide a detailed overview in the discussion. We will show the flexibility of our solution by anonymizing data with a broad spectrum of privacy criteria, including k-anonymity, ℓ-diversity, t-closeness and δ-presence, using a globally optimal data anonymization algorithm. Most related approaches in the distributed setting implement heuristic methods, as their coding models result in large search spaces. While it has been shown that these heuristics combined with, e.g., local recoding, can outperform optimal algorithms using single-dimensional global recoding in terms of data quality, we chose such an algorithm as these have said to be very well suited for the biomedical domain [14].

We present an extensive analytical and experimental evaluation of our solution and show that it offers highly competitive execution times. The performance of our approach can be accurately estimated with a model that only depends on basic data characteristics. Our protocol relaxes the guarantees of traditional secure multiparty computations by exchanging non-anonymized – but encrypted – subsets of the data. We present effective means to lower privacy risks and discuss a trade-off between privacy, data quality and efficiency. Together with estimates derived from our model, this can be utilized to tailor our method to project specific requirements.

2. Background

2.1. Centralized anonymization algorithms

A typical approach for anonymization is to introduce fuzziness. In this work, we focus on the most common transformation methods: generalization and suppression. For an overview of further techniques, such as perturbation or permutation, the interested reader is referred to [16].

Generalization is often implemented with *generalization hierarchies*. These are transformation rules that allow to iteratively generalize the values of an attribute. Tabular representations of example hierarchies for the categorical attribute *Gender* with two generalization levels and the discrete numerical attribute *ZIP* with six generalization levels are shown in Fig. 1. Generalization hierarchies are well suited for transforming categorical attributes and discrete numeric attributes. They can also be used for quasi-identifiers that are continuous numerical attributes. One solution is to formulate transformation rules as functions that dynamically create generalization hierarchies for the values of an attribute in a specific dataset. A more detailed discussion of how such quasi-identifiers can be handled with our method is given in Section 6.5. *Suppression* is a special kind of generalization, in which a data item is completely suppressed.

Many anonymization algorithms use the rules encoded in generalization hierarchies to transform a dataset. Depending on the type of transformations applied, this results in differently large search spaces. Some algorithms implement *local recoding*, while others implement *global recoding* [17]. The former means that different rules can be applied to equal data items, whereas the latter means that the same rule is applied. When *single-dimensional recoding* is implemented, the data items are values of an individual column, whereas *multi-dimensional recoding* means that data items are combinations of values from different columns, e.g., complete tuples [17]. Multidimensional global recoding means that the same rule is always applied to equal tuples. From the perspective of a single attribute, this results in local recoding of its values.

The method of generalization can be distinguished into *full-domain generalization* or *subtree generalization* [16]. The former means that the entire domain of a data item is transformed to a more general domain (i.e., level) of its generalization hierarchy [18]. The latter means that different generalization levels can be applied to different subsets of data items from the same domain.

Generalization-based techniques are sometimes distinguished by whether they are *hierarchy-based* or *partition-based* [17]. Partition-based algorithms are often used for continuous numerical attributes and require the existence of a total order on the data items. They generalize data items by partitioning them into ranges. Hierarchy-based approaches are often used for categorical and discrete numeric attributes and require the existence of generalization hierarchies.

We will now present a short overview of a broad spectrum of state-of-the-art anonymization algorithms. Apart from some restrictions, most of them are supported by our approach. A detailed discussion is presented in Section 6.2.

Optimal data anonymization algorithms often implement hierarchy-based global recoding with single-dimensional full-domain generalization to reduce the size of the search space. As this is a restrictive coding model that potentially results in much loss of information, suppression is added as a multi-dimensional global recoding technique. As a result, outliers are removed from the dataset as long as the total number of suppressed tuples remains

Lvl-0	Lvl-1
m	*
f	*

(a) Gender

Lvl-0	Lvl-1	Lvl-2	Lvl-3	Lvl-4	Lvl-5
80000	8000*	800**	80***	8****	*****
...	...	...	...	...	...
89999	8999*	899**	89***	8****	*****
90000	9000*	900**	90***	9****	*****
...	...	...	...	...	...
99999	9999*	999**	99***	9****	*****

(b) ZIP

Fig. 1. Exemplary tabular generalization hierarchies (Lvl = level, m = male, f = female).

Please cite this article in press as: Kohlmayer F et al. A flexible approach to distributed data anonymization. J Biomed Inform (2013), http://dx.doi.org/10.1016/j.jbi.2013.12.002