Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A data recipient centered de-identification method to retain statistical attributes

Tamas S. Gal [a,*], Thomas C. Tucker [a], Aryya Gangopadhyay [b], Zhiyuan Chen [b]

[a] University of Kentucky, 2365 Harrodsburg Rd., Suite A230, Lexington, KY 40504, USA
[b] University of Maryland at Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA

ARTICLE INFO

Article history:
Received 28 August 2013
Accepted 3 January 2014
Available online xxxx

Keywords:
Privacy
Utility based privacy preserving data mining
Statistical analysis

ABSTRACT

Privacy has always been a great concern of patients and medical service providers. As a result of the recent advances in information technology and the government's push for the use of Electronic Health Record (EHR) systems, a large amount of medical data is collected and stored electronically. This data needs to be made available for analysis but at the same time patient privacy has to be protected through de-identification. Although biomedical researchers often describe their research plans when they request anonymized data, most existing anonymization methods do not use this information when de-identifying the data. As a result, the anonymized data may not be useful for the planned research project. This paper proposes a data recipient centered approach to tailor the de-identification method based on input from the recipient of the data. We demonstrate our approach through an anonymization project for biomedical researchers with specific goals to improve the utility of the anonymized data for statistical models used for their research project. The selected algorithm improves a privacy protection method called *Condensation* by Aggarwal et al. Our methods were tested and validated on real cancer surveillance data provided by the Kentucky Cancer Registry.

1. Introduction

The advances in Information Technology and the recent push from the federal government [1] made Electronic Health Records (EHR) systems widespread in the United States. Based on a survey by the American Medical Association (AMA), 42% of physicians use some kind of EHR system, and it is estimated, that by 2015 the coverage will grow to over 80% [2]. Electronically collected biomedical data needs to be made available for research but at the same time patient privacy must be protected. This is a major challenge for the Healthcare Data and Knowledge Management field that has technical, management and policy implications.

Various approaches have been proposed to address privacy issues regarding publicly released data. A popular solution is to mask the original values of the attribute that could be used to identify individuals. Perturbation based masking methods add random noise to the original data values [3–8]. Data swapping techniques exchange attribute values between different records [9,10]. Generalization methods replace original values with more general ones [11–13]. Suppression is a special format of generalization when the value of an attribute is removed from the record. These masking methods can be used by themselves or as parts of more complex anonymization schemas.

Microaggregation and *k*-anonymity are two grouping based de-identification approaches that gained considerable popularity in recent years [14–17]. The main idea behind them is to partition the data into groups of similar records and then mask the quasi identifier attributes at group level so the records within a group become indistinguishable. Multiple solutions have been proposed to used as partitioning and masking methods to optimize these anonymization methods [18,12,19,20].

The process of privacy preservation causes information loss, which can be considered as loss of *utility*. To produce useful output the data publisher has to balance the competing requirements of sufficient privacy protection and maximum possible utility. Table 1 shows an example of utility loss in privacy preservation [21]. {*Age, Insurance, Zip*} can be used to identify individuals in the dataset (quasi identifiers). *Diagnosis* is a sensitive attribute. *Screening* shows whether the individual is targeted for colon cancer screening or not. Suppose that, in order to protect the sensitive attribute (*Diagnosis*), 2-diversity is required, so the quasi identifiers need to be modified in such a way that based on the quasi identifiers {*Age, Insurance, Zip*} each individual in the dataset would be indistinguishable from at least one other person. Tables 1(A) and (B) are both valid 2-anonymizations of the original data (records sharing the same quasi identifiers have the same Group IDs). However, Table 1(A) provides more accurate results than Table 1(B) when answering the following queries:

* Corresponding author.
E-mail addresses: tamas.gal@uky.edu (T.S. Gal), tct@kcr.uky.edu (T.C. Tucker), gangopad@umbc.edu (A. Gangopadhyay), zhchen@umbc.edu (Z. Chen).

1532-0464/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2014.01.001

Table 1
Utility loss in privacy preservation.

ID	Age	Insurance	Zip	Diagnosis	Screening
Original data:					
1	54	No	40504	HIV	Y
2	55	No	40509	HEP-B	Y
3	60	HMO	40512	SM	N
4	60	HMO	40517	HEP-B	N
5	62	HMO	40524	HEP-B	N
6	62	PPO	40525	Prostate cancer	N

Group ID	ID	Age	Insurance	Zip	Diagnosis	Screening
De-identified data (A):						
1	1	[54–55]	No	4050X	HIV	Y
1	2	[54–55]	No	4050X	HEP-B	Y
2	3	60	HMO	4051X	SM	N
2	4	60	HMO	4051X	HEP-B	N
3	5	62	Private	4052X	HEP-B	N
3	6	62	Private	4052X	Prostate cancer	N
De-identified data (B):						
1	1	[54–60]	Any	405XX	HIV	Y
2	2	[55–62]	Any	405XX	HEP-B	Y
3	3	[60–62]	HMO	405XX	SM	N
1	4	[54–60]	Any	405XX	HEP-B	N
3	5	[60–62]	HMO	405XX	HEP-B	N
2	6	[55–62]	Any	405XX	Prostate cancer	N

Q1: *How many patients under age 59 are there in the data set?*
Q2: *Is an individual with Age = 55, Insurance = No, Zip = 40509 targeted for colon cancer screening?*

According to Table 1(A) the answer to Q1 is 2 and to Q2 is "Y". But according to Table 1(B), the answer to Q1 is an interval [0, 4], because 59 falls in the age range of record 1, 2, 4, and 6. The answer to Q2 is "Y" or "N" with 50% probability each.

Two conclusions can be drawn from this example:

- Different anonymization leads to different information loss. Tables 1(A) and (B) are on the same anonymization level but Table 1(A) provides better results. Therefore, utility loss should be minimized in privacy preserving.
- Data utility depends on the application. Q1 is an aggregate query, so the data is more useful if the values are more accurate. Q2 is a classification query so the utility of the data depends on how much the classification model is preserved in the de-identified data. Utility is the quality of the data for the intended use.

To decide whether one de-identification method preserves utility better than another, we need to measure utility of the de-identified data compared to the utility of the original data. In practical terms it means that we need to define a *distance measure* between the original data and the de-identified data based on utility. The content of this distance measure depends on the use of the data.

The followings are examples of utility measures used in the literature:

- *Query answering accuracy*: Answering queries such as *count*, *average* and *sum* is the most common use of published data. The quality of query answering depends on the distance of each original value from the corresponding value in the anonymized dataset. A quantitative measure was introduced by Xu et al., which uses the normalized interval size to measure the utility loss for numeric attributes and normalized number of descendants in the generalization hierarchy to measure the utility loss for categorical attributes [22,23].
- *Classification accuracy*: The published data is often used to train classifiers, therefore the data quality depends on how well the class structure is preserved in the anonymized data. Fung et al. propose a metric that measures entropy change during anonymization [24,25]. Ideally, the entropy of an equivalence class with respect to class label distribution should be minimized in the published data.
- *Distribution similarity*: Statistical distribution is an important characteristic of a dataset. A model which measures the difference between the distribution of the original and the anonymized data has been developed by Kifer et al. [26].
- *Discernibility measure:* Bayardo and Agrawal consider a discernibility measure as a utility measure as they try to minimize the equivalence class size while anonymizing the data [27]. The more records are in an equivalence class, the less specific information is preserved for those records.
- *Generalization measures* include *Generalization Height* [28], which measures the total number of generalization steps applied in the anonymization process. The idea behind this measure is that generalization causes information loss and the total number of generalization steps represents the total amount of loss. The *Loss Metric* penalizes the generalization made in that entry according to the size of the generalized subset [29,30]. *Ambiguity Metric* is the average size of the Cartesian products of all generalized entries in each record in the table [30].
- *Entropy based measures:* Gionis and Tassa introduced entropy as Mutual Information Utility Measure [31]. Private Mutual Information Utility Measure builds on the previously mentioned entropy measure and it quantifies the mutual information between the generalized public data and the private data [19].

The same de-identified dataset might be useful for one purpose but useless for another. When researchers request de-identified biomedical data, they already have a plan how they want to use it. Yet, these research plans are rarely utilized when choosing the de-identification method. We believe that de-identification methods should be tailored to the specific needs of the data recipient when possible and that this customization should reflect in utility measurements as well.

We present a de-identification framework to address the need for customized anonymization. Our approach investigates the requirements of the data recipient and selects a suitable de-identification method that is specific to the requirements. We evaluated our method by comparing it to three general purpose de-identification algorithms using utility measures that were specific to the data recipient's requirements.

Our experiments used real cancer surveillance data provided by the Kentucky Cancer Registry.

The rest of the paper is organized as follows: Section 2 gives a detailed review of related work. Section 3 explains the materials and methods used in our experiments. Section 4 describes our results. Section 5 discusses some of the issues that arose during our experiments and Section 6 concludes the paper and provides directions for future work.

2. Related work

Most medical providers follow the Safe Harbor standard [32] in the US when releasing data which removes 18 well defined identifiers from the dataset. Sweeney showed that removing obvious identifiers does not provide protection against privacy attacks [33]. As a solution, *k-anonymity* was proposed by Samarati and Sweeney [11]. *k*-anonymity divides the data attributes into *quasi identifiers*, *sensitive attributes* and *non-sensitive attributes* and creates equivalence classes by masking quasi identifier attributes in such a way that the quasi identifier attributes of any record would be identical to quasi identifier attributes of at least $k-1$ other records. Achieving optimal *k-anonymity* is NP-hard