Contents lists available at ScienceDirect

Information Fusion

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

Full Length Article

Utility-preserving privacy protection of nominal data sets via semantic rank swapping

Mercedes Rodriguez-Garcia[a,*], Montserrat Batet[b], David Sánchez[a]

[a] *UNESCO Chair in Data Privacy, Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Catalonia, Spain*

[b] *Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Av. Carl Friedrich Gauss, 5, Parc Mediterrani de la Tecnologia, Castelldefels, Barcelona, Catalonia 08860, Spain*

ARTICLE INFO

Keywords:
Rank swapping
Nominal data
Semantics
Ontologies

ABSTRACT

Personal data are of great interest for research but, at the same time, they pose a serious privacy risk. Therefore, appropriate data protection measures should be undertaken by the data controller before making personal data available for secondary use. Also, such data protection should be done in a way that data are still useful for analysis. In the last years, a plethora of data protection mechanisms have been proposed. Among them, *rank swapping* is considered one of the best with respect to disclosure risk minimization and data utility preservation. Because rank swapping is based on sorting input data to swap values that are close to each other, in principle, it is a method restricted to numerical and ordinal categorical data. However, a significant amount of personal data currently compiled and used in data analysis are nominal, and their utility depends on the semantics they convey. To properly cope with this type of data, in this paper, we present rank swapping methods capable of protecting nominal data from a semantic perspective. Specifically, by exploiting ontologies, our methods are able to protect nominal data while properly preserving their semantics and, thus, their analytical utility. For that, we provide a suitable binary relation to semantically sort nominal data. Our proposal is capable of managing both independent individual attributes and non-independent multivariate data sets, being the latter especially relevant for data analysis. Empirical experiments carried on real clinical records and using a standard medical ontology show that our methods are able to preserve the semantic features of nominal data significantly better than standard permutation mechanisms.

1. Introduction

In the current era of big data and digital societies, information collection, storage and processing capabilities have meaningfully grown. Social networks, electronic records or web browsing generate huge volumes of information about individuals that are of great interest for public and private organizations. Collection and processing (e.g., data mining) of these data allows conducting a variety of surveys, improving decision-making in business or offering personalized services to enhance the online experience. However, the dissemination of personal data may compromise the individuals' privacy, which is considered a fundamental right, and it is supported by international treaties and constitutional laws, such as the Universal Declaration of Human Rights (1948).

In this scenario, governmental agencies and current legislations on data protection, such as the General Data Protection Regulation (GDPR) [1], emphasize the need of adequately protecting Personally Identifiable Information (PII) [2] to preserve individuals' privacy. PII includes not only identifying data, such as social security numbers, but also any non-identifying data that, in combination with other non-identifying data, can be used by attackers to re-identify individuals by linking them with external data sources, as shown several studies [3–5]. These non-identifying attributes that, in aggregate, can be used to unequivocally re-identify individuals are known as *quasi-identifier attributes* and they cause real privacy threats. Quasi-identifiers are currently employed by data brokers to compile and aggregate individuals' data and, from these, build user profiles that are later used or sold to third parties for commercial and business purposes [6].

To minimize the chance of re-identification, quasi-identifying attributes should be subjected to *anonymization*. In turn, data anonymization should be done in a way that the protected data still retain as much analytical utility as possible, so that conclusions or inferences extracted from the analysis of the anonymized data set are similar to those of the original data set. For that, different masking methods have

[*] Corresponding author.
E-mail address: mercedes.rodriguez@uca.es (M. Rodriguez-Garcia).

https://doi.org/10.1016/j.inffus.2018.02.008
Received 27 July 2017; Received in revised form 11 February 2018; Accepted 25 February 2018

been proposed within the disciplines of Statistical Disclosure Control (SDC) [7] and Privacy-Preserving Data Publishing (PPDP) [8]. Among them, perturbative masking methods are the most widespread, which include *noise addition, microaggregation* or *rank swapping*. These mechanisms generate a modified version of the original data by distorting or introducing ambiguity on the quasi-identifying attributes while preserving certain statistical features. As shown in several studies [9,10], *rank swapping* is considered one of the best perturbative mechanisms w.r.t. disclosure risk minimization and data utility preservation. This method, which is based on the idea of proximity swapping [11], ranks the values of each attribute in ascending order for later swapping each value with another one randomly chosen within a restricted size range. Thus, the higher the range size, the higher the ambiguity in the re-identification inferences and the lower the disclosure risk; but also, the lower the data utility, because swapped values would tend to be less similar. Concerning data utility, and on the contrary to other data protection mechanisms [7], rank swapping perfectly preserves univariate statistics, such as the mean, the variance and the frequency distribution, because the values in the protected attribute are the same as those in the original attribute but permuted. For this same reason, rank swapping also preserves other very useful features for data analysis, such as data granularity or outlying values.

In addition to the above advantages, a recent study [12] has shown that any anonymization method is functionally equivalent to a permutation plus a small amount of noise; this turns the spotlight on the permutation-based data transformation implemented by the rank swapping mechanism as the essential principle underlying any data anonymization.

Because rank swapping relies on the ability to sort attribute values, it has been designed to deal with numerical attributes (e.g., income) and ordinal categorical attributes, i.e., data that admit order relationships (e.g., color, where the different colors may be ranked on basis of their wave lengths) [7]. However, a significant amount of personal data that are currently gathered by data brokers for categorizing individuals (e.g. from social networks, electronic healthcare records or web browsing logs) and that should be subject of anonymization, are of nominal nature [6]. Unlike other data types, *nominal categorical* attributes (e.g., occupation, race, religion, etc.) are finite, discrete, textual and non-ordinal; thus, they do not admit order relationships. In this scenario, in principle, it is not possible to carry out the sorting operation needed to rank the values of the data set during the permutation process. Moreover, because nominal data utility is closely related to the preservation of data semantics [13-15], any data transformation performed to anonymize nominal data, such as ranking, should consider the *meaning* of the attribute values. So far, only microaggregation and noise addition methods have been adapted to work with nominal data from a semantic perspective [16-19]. To do so, they exploit the formal semantics modeled in ontologies, which are knowledge structures that formally describe the concepts of a domain and the semantic relationships between them.

In this paper we present rank swapping methods capable of protecting nominal data from a semantic perspective. Our objective is twofold: (i) to provide a binary relation capable of semantically sorting nominal data by exploiting the formal semantics modeled in ontologies, and (ii) to provide mechanisms to control the degree of permutation in order to enforce a certain level of protection while preserving, as much as possible, the semantic features, and thus, the analytical utility of the data. In particular, we propose semantically-grounded rank swapping solutions to perturb individual nominal attributes and multivariate nominal data sets. The latter is especially relevant because it is capable of protecting multivariate nominal data sets while reasonably preserving the correlation among attributes, which is of outmost importance for data analysis.

The rest of the paper is organized as follows. Section 2 discusses related works on permutation-based methods for data protection. Section 3 defines a suitable binary relation to semantically sort nominal data and presents our semantically-grounded rank swapping algorithms. Section 4 details the empirical experiments we carried out on real clinical records and by exploiting a standard medical ontology, and measures and compares the data utility preserved by our methods against several baselines. Section 5 contains the conclusions and depicts some lines of future research.

2. Related work

Various permutation-based methods have been proposed to protect data sets while preserving certain statistical features. The first one, named data swapping [20], is based on swapping the values of each attribute from a data set of t categorical attributes to yield a permuted data set whose t-order frequency counts, or t-order statistics, are the same as those of the original data set, i.e., a t-order equivalent data set. Since the t-order statistics are preserved, the inferences that derive from them are not altered. To distort a given attribute, the method builds all the equivalence classes for that attribute and randomly swaps the values within each class. Because of the way in which data swapping operates, this method is not suitable when most equivalence classes in the data set are composed of one or few records, which is the case of data sets with fine grained nominal attributes, since the swaps can hardly be carried out. On the other hand, if the data are released as microdata, it is necessary to add enough uncertainty on the true values of the individuals' data to reasonably protect their privacy. However, identifying a large number of swaps that preserve the t-order statistics is computationally impractical [21,22]. As a feasible approach for the release of microdata, Reiss proposes in [21] a variation of data swapping where the t-order frequency counts are approximately preserved. Firstly, the method computes the relevant frequency tables from the original data set, and then constructs a new data set consistent with these tables. To do this, the values of an attribute are randomly selected according to the probability distribution derived from the original frequency tables; because this may produce values not appearing in the original data set, this makes it a synthetic method, rather than a strict data swapping one.

Because the above methods do not limit the swapping range, very different values may be swapped, thereby increasing the loss of utility. In order to limit the scope of the swaps and, therefore, maintain each permuted value within a certain rank-distance from the original one, Greenberg, in an unpublished manuscript [23] described by Moore in [11], presents *rank swapping*, a method initially defined for ordinal categorical data and subsequently applied to numerical data [9]. This method restricts the swapping range by ranking the values of the attribute in ascending order and, then, by swapping each value with another unswapped one randomly chosen within a user-defined range *p, p* being a percent of the total number of records. In this way, the rank of two swapped values cannot differ by more than p% of the total number of records. Large values of p lead to greater permutations whereas smaller values of p incur in higher disclosure risk. In [9], rank swapping is pointed out as the best performer data protection mechanism in terms of disclosure risk minimization and data utility preservation.

Rank swapping has been designed to deal with numerical or ordinal categorical data. In both cases, total orders are available to build the value ranks in which the algorithms rely on. However, nominal data are not ordinal and, thus, lack natural total orders. For such data, rank swapping has been considered either non-applicable [7,24] or it has been suboptimally applied by defining artificial total orders (e.g., topological order of categorical labels for nominal attributes) [25] that, due to their lack of semantic coherence, may severely hamper the utility of the protected outcomes.

3. Semantic rank swapping methods

In this section, we propose semantically-grounded rank swapping