Information Fusion 30 (2016) 1–14

Contents lists available at ScienceDirect

Information Fusion

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

Utility-preserving differentially private data releases via individual ranking microaggregation

David Sánchez*, Josep Domingo-Ferrer, Sergio Martínez, Jordi Soria-Comas

UNESCO Chair in Data Privacy, Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain

ARTICLE INFO

Article history:
Received 7 February 2015
Revised 7 October 2015
Accepted 2 November 2015
Available online 6 November 2015

Keywords:
Privacy-preserving data publishing
Differential privacy
k-Anonymity
Microaggregation
Data utility

ABSTRACT

Being able to release and exploit open data gathered in information systems is crucial for researchers, enterprises and the overall society. Yet, these data must be anonymized before release to protect the privacy of the subjects to whom the records relate. Differential privacy is a privacy model for anonymization that offers more robust privacy guarantees than previous models, such as *k*-anonymity and its extensions. However, it is often disregarded that the utility of differentially private outputs is quite limited, either because of the amount of noise that needs to be added to obtain them or because utility is only preserved for a restricted type and/or a limited number of queries. On the contrary, *k*-anonymity-like data releases make no assumptions on the uses of the protected data and, thus, do not restrict the number and type of doable analyses. Recently, some authors have proposed mechanisms to offer general-purpose differentially private data releases. This paper extends such works with a specific focus on the preservation of the utility of the protected data. Our proposal builds on microaggregation-based anonymization, which is more flexible and utility-preserving than alternative anonymization methods used in the literature, in order to reduce the amount of noise needed to satisfy differential privacy. In this way, we improve the utility of differentially private data releases. Moreover, the noise reduction we achieve does not depend on the size of the data set, but just on the number of attributes to be protected, which is a more desirable behavior for large data sets. The utility benefits brought by our proposal are empirically evaluated and compared with related works for several data sets and metrics.

1. Introduction

Releasing and exploiting open data is crucial to boost progress of knowledge, economy and society. Indeed, the availability of such data facilitates research and allows better marketing, better planning and better social services. However, data publication often faces privacy threats due to the confidentiality of the information that is released for secondary use. To tackle this problem, a plethora of methods aimed at data anonymization have been proposed within the field of statistical disclosure control [1]. Such methods distort input data in different ways (*e.g.* noise addition, removal, sampling, data generalization, etc.) so that the probability of re-identifying individuals and, thus, disclosing their confidential information is brought below a tolerable threshold. Even though those methods have been shown to improve privacy protection while preserving a reasonable level of analytical utility (the main motivation of data publishing), they offer no formal privacy guarantees.

In contrast, privacy models proposed in recent years within the computer science community [2,3] seek to attain a predefined notion of privacy, thus offering *a priori* privacy guarantees. These guarantees are interesting because they ensure a minimum level of privacy regardless of the type of transformation performed on input data. Among such models, *k*-anonymity and the more recent ε-differential privacy have received a lot of attention.

k-Anonymity [4,5] seeks to make each record in the input data set indistinguishable from, at least, $k-1$ other records, so that the probability of re-identification of individuals is, at most, $1/k$. Different anonymization methods have been proposed to achieve that goal, such as removal of outlying records, generalization of values to a common abstraction [5–8] or multivariate microaggregation [9,10]. The latter method partitions a data set into groups at least k similar records and replaces the records in each group by a prototypical record (*e.g.* the centroid record, that is, the average record). Whatever the computational procedure, *k*-anonymity focuses on masking quasi-identifier attributes; these are attributes (*e.g.*, Age, Gender, Zipcode and Race) that are assumed to enable re-identifying the

* Corresponding author. Tel.: +34 977559657.
E-mail addresses: david.sanchez@urv.cat (D. Sánchez), josep.domingo@urv.cat (J. Domingo-Ferrer), sergio.martinezl@urv.cat (S. Martínez), jordi.soria@urv.cat (J. Soria-Comas).

http://dx.doi.org/10.1016/j.inffus.2015.11.002
1566-2535/© 2015 Elsevier B.V. All rights reserved.

respondent of a record because they are linkable to analogous attributes available in external identified data sources (like electoral rolls, phone books, etc.). k-Anonymity does not mask confidential attributes (*e.g.*, salary, health condition, political preferences, etc.) unless they are also quasi-identifiers. While k-anonymity has been shown to provide reasonably useful anonymized results, especially for small k, it is also vulnerable to attacks based on the possible lack of diversity of the non-anonymized confidential attributes or on additional background knowledge available to the attacker [11–14].

Unlike k-anonymity, the more recent ε-differential privacy [15] model does not make any assumptions on which attributes are quasi-identifiers, that is, on the background knowledge available to potential attackers seeking to re-identify the respondent of a record. ε-Differential privacy guarantees that the anonymized output is insensitive (up to a factor dependent on ε) to the modification, deletion or addition of any single input record in the original data set. In this way, the privacy of any individual is not compromised by the publication of the anonymized output, which is a much more robust guarantee than the one offered by k-anonymity. Differential privacy is attained by adding an amount of noise to the original outputs that is proportional to the *sensitivity* of such outputs to modifications of particular individual records in the original data set. This sensitivity does not depend on the specific values of attributes in specific records, but on the domains of those attributes. Basing the sensitivity and hence the added noise on attribute domains rather than attribute values satisfies the privacy guarantee but may yield severely distorted anonymized outputs, whose utility is very limited. Because of this, ε-differential privacy was originally proposed for the *interactive* scenario, in which the outputs are the answers to interactive queries rather than the data set itself. When applying ε-differential privacy to this scenario, the anonymizer returns noise-added answers to interactive queries. In this way, the accuracy/utility of the response to a query depends on the sensitivity of the query, which is usually less than the sensitivity of the data set attributes. However, the interactive setting of ε-differential privacy limits the number and type of queries that can be performed. The proposed extensions of ε-differential privacy to the *non-interactive* setting (generation of entire anonymized data sets) overcome the limitation on the number of queries, but not on the type of queries for which some utility is guaranteed (see Section 2.2 below).

1.1. Contribution and plan of this paper

In previous works [16,17], we showed that the noise required to fulfill differential privacy in the non-interactive setting can be reduced by using a special type of microaggregation-based k-anonymity on the input data set. The rationale is that the microaggregation performed to achieve k-anonymity helps reducing the sensitivity of the input versus modifications of individual records. As a result, data utility preservation can be improved (in terms of less data distortion) without renouncing the strong privacy guarantee of differential privacy. With such a mechanism, the sensitivity reduction depends on the number of k-anonymized groups to be released; in turn, this number is a function of the value of k and the cardinality of the data set. The larger the group size k, the less sensitive are the group centroids resulting from the microaggregation; on the other hand, the smaller the data set, the smaller the number of different group centroids in the microaggregated data set. Thus, as the group size increases or the data set size decreases, the sensitivity decreases and less noise needs to be added to reach differential privacy. Hence, the resulting differentially private data have higher utility. In the two abovementioned works, we empirically showed that the noise reduction more than compensates the information loss introduced by microaggregation.

In line with [16,17], in this paper we investigate other transformations of the original data aimed at reducing their sensitivity. The proposal in this paper is based on individual ranking microaggregation, a kind of microaggregation that is more flexible and utility-preserving than the one used in [16,17]. In contrast to the previous works, the reduction of sensitivity achieved by the method presented in this paper does not depend on the size of the data set, but just on the number of attributes to be protected. This is more desirable for large data sets or in scenarios in which only the confidential attributes should be protected. In fact, experiments carried on two reference data sets show a significant improvement of data utility (in terms of relative error and preservation of attribute distributions) with respect to the previous work. Moreover, the microaggregation mechanism used in this paper is simpler and more scalable, which facilitates implementation and practical deployment.

The rest of this paper is organized as follows. Section 2 reviews background on microaggregation, ε-differential privacy and ε-differentially private data publishing. Section 3 proposes the new method to generate ε-differentially private data sets that uses a special type of microaggregation to reduce the amount of required noise. Implementation details are given for data sets with numerical and categorical attributes. Section 4 reports an empirical comparison of the proposed method and previous proposals, based on two reference data sets. The final section gathers some conclusions.

2. Background and related work

2.1. Background on microaggregation

Microaggregation [9,18] is a family of anonymization algorithms that works in two stages:

- First, the data set is clustered in such a way that: i) each cluster contains at least k elements; ii) elements within a cluster are as similar as possible.
- Second, each element within each cluster is replaced by a representative of the cluster, typically the centroid value/tuple.

Depending on whether they deal with one or several attributes at a time, microaggregation methods can be classified into univariate and multivariate:

- Univariate methods deal with multi-attribute data sets by microaggregating one attribute at a time. Input records are sorted by the first attribute, then groups of successive k values of the first attribute are created and all values within that group are replaced by the group representative (*e.g.* centroid). The same procedure is repeated for the rest of attributes. Notice that all attribute values of each record are moved together when sorting records by a particular attribute; hence, the relation between the attribute values within each record is preserved. This approach is known as *individual ranking* [18,19] and, since it microaggregates one attribute at a time, its output is not k-anonymous at the record level. Individual ranking just reduces the variability of attributes, thereby providing some anonymization. In [20] it was shown that individual ranking causes low information loss and, thus, its output better preserves analytical utility. However, the disclosure risk in the anonymized output remains unacceptably high [21].
- To deal with several attributes at a time, the trivial option is to map multi-attribute data sets to univariate data by projecting the former onto a single axis (*e.g.* using the sum of z-scores or the first principal component, see [19]) and then use univariate microaggregation on the univariate data. Another option avoiding the information loss due to single-axis projection is to use *multivariate microaggregation* able to deal with unprojected multi-attribute data [9,22]. If we define optimal microaggregation as finding a partition in groups of size at least k such that within-groups homogeneity is maximum, it turns out that, while optimal univariate microaggregation can be solved in polynomial