Knowledge-Based Systems 000 (2018) 1–13

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Differentially private data publishing via optimal univariate microaggregation and record perturbation

Jordi Soria-Comas, Josep Domingo-Ferrer*

Department of Computer Science and Mathematics, UNESCO Chair in Data Privacy, CYBERCAT-Center for Cybersecurity Research of Catalonia, Universitat Rovira i Virgili, Av. Països Catalans 26, Tarragona 43007, Catalonia

ARTICLE INFO

Article history:
Received 22 June 2017
Revised 20 April 2018
Accepted 21 April 2018
Available online xxx

Keywords:
Differential privacy
Microaggregation
Anonymization
Statistical disclosure control
Privacy

ABSTRACT

We present an approach to generate differentially private data sets that consists in adding noise to a microaggregated version of the original data set. While this idea has already been pursued in the literature to reduce the sensitivity of attributes and hence the noise required to reach differential privacy, the novelty of our approach is that we focus on the microaggregated data set as our protection target (rather than aiming at protecting the original data set and viewing the microaggregated data set as a mere intermediate step). Interestingly, by starting from the microaggregated data set rather than the original data set, we achieve differential privacy for the individuals having contributed the original records while preserving substantially more utility. Compared with previous contributions using microaggregation as a prior step to reach differential privacy, the utility improvement comes from avoiding the need to use insensitive microaggregation. This claim is supported by theoretical and empirical utility comparisons between our approach and existing approaches. We analyze several microaggregation strategies: multivariate MDAV, individual-ranking MDAV, and optimal microaggregation. In particular, we reformulate optimal microaggregation to fit it to the generation of differentially private data sets.

1. Introduction

Microdata (that is, information at the individual level) are usually the most convenient type of data for secondary use. However, the risk of disclosure inherent to releasing such detailed information is significant. Traditionally, data were mostly handled by a reduced number of controllers (e.g. national statistical offices), who had collected them under strong pledges of privacy. In that scenario, reasonable assumptions about the knowledge available to intruders could be made and the methodology for disclosure risk limitation could be adjusted accordingly. Nowadays, the developments in information technology greatly facilitate the collection of all kinds of personal data by a variety of controllers. This bounty of information increasingly complicates making well-grounded assumptions about the background knowledge available to potential intruders [23].

Differential privacy [9] (DP) is a well-known privacy model that gives privacy guarantees without assuming anything on the intruder's background knowledge. In this sense, DP is well adapted to the current scenario with many data controllers. Unlike privacy models designed to protect sets of microdata (e.g. [12,14,19]), DP was originally introduced to limit the disclosure risk incurred when returning answers to queries on a database. Nevertheless, DP mechanisms to generate entire data sets were proposed soon after the inception of DP [13]; we will use the term DP data set to denote a data set output by a DP mechanism.

The dominant approach to generate DP microdata is based on the computation of DP histograms. Nevertheless, histogram-based approaches have severe limitations when the number of attributes grows: for fixed attribute granularities, the number of histogram bins grows exponentially with the number of attributes, which has a severe impact on both computational cost and accuracy. Alternatively, the DP data set can be generated using a record perturbation approach. The simplest way to do this is to collect DP answers to a set of queries that ask for each individual record in the original data set. However, the amount of noise required to attain DP for such queries is too large for the DP data set to stay useful. In [24], microaggregation is used as an intermediate step to reduce the sensitivity of the queries: rather than asking for the record of each individual, we arrange the records into clusters and query for a representative of each cluster (for example, the centroid record). The sensitivity of the latter queries is smaller because they depend on several individuals (those in the cluster). Related approaches were followed in [20,21,25]. However, these contributions are not

* Corresponding author.
E-mail addresses: jordi.soria@urv.cat (J. Soria-Comas), josep.domingo@urv.cat (J. Domingo-Ferrer).

https://doi.org/10.1016/j.knosys.2018.04.027
0950-7051/© 2018 Elsevier B.V. All rights reserved.

without limitations: [24] and [25] need special microaggregation algorithms that are less utility-preserving than the standard ones, whereas [20] and [21] can only deal with data sets containing a single attribute.

1.1. Contribution and plan of this paper

In this work we present a novel record-level perturbation-based methodology to generate DP data sets. Unlike existing perturbation-based approaches, we can use standard microaggregation algorithms and deal with multiple attributes, which leads to a significant improvement in the utility of the resulting DP data set. Our approach can work with any microaggregation algorithm, but we will choose a few well-known algorithms for the sake of evaluation. The use of standard microaggregation algorithms becomes possible because we switch the focus of DP from the original data set to the microaggregated data set. To make this change compatible with DP, we extend the definition of DP to data sets in which the usual assumption of a one-to-one mapping between records and individuals does not hold. This extension makes sense because the aim of DP is to protect individuals (rather than records).

In Section 2 we briefly introduce basic concepts about DP and recall the state of the art in DP data set generation via record masking. In Section 3 we extend the notion of DP to data sets where there is no one-to-one mapping between records and individuals. In Section 4 we describe our approach to generating DP data sets. In Section 5 we propose several microaggregation methods based on the MDAV heuristic for generating DP data sets. In Section 6 we introduce a new optimal microaggregation method for DP data set generation. In Section 7 we present theoretical analyses on the utility of the generated DP data sets. In Section 8 we experimentally evaluate the proposed microaggregation methods by comparing them with each other and with previous approaches. Finally, in Section 9 we summarize the conclusions and outline future research avenues.

2. Preliminaries

2.1. Background on differential privacy

Differential privacy [9] is popular among academics due to the strong privacy guarantees it offers. DP does not rely on assumptions about the background knowledge available to the intruders. Rather, disclosure risk limitation is tackled in a relative manner: the result of any analysis should be similar between data sets that differ in one record. As stated in [8], such a guarantee should encourage individuals to participate in a data set because the disclosure risk they incur when contributing their data is limited:

> Any given disclosure will be, within a multiplicative factor, just as likely whether or not the individual participates in the database. As a consequence, there is a nominally higher risk to the individual in participating, and only nominal gain to be had by concealing or misrepresenting one's data.

Differential privacy assumes the presence of a trusted party that: (i) holds the data set, (ii) receives the queries submitted by the data users, and (iii) responds to them in a privacy-aware manner. The notion of differential privacy is formalized according to the following definition:

Definition 1 (ϵ-differential privacy)**.** A randomized function κ gives ϵ-differential privacy (ϵ-DP) if, for all data sets D_1 and D_2 that differ in one record (a.k.a. neighbor data sets), and all $S \subset Range(\kappa)$, one has

$$\Pr(\kappa(D_1) \in S) \leq \exp(\epsilon) \Pr(\kappa(D_2) \in S). \qquad (1)$$

Given a query function f, the goal in differential privacy is to find a randomized function κ_f that satisfies ϵ-DP and approximates f as closely as possible. For the case of numerical queries, κ_f can be obtained via noise addition; that is, $\kappa_f(\cdot) = f(\cdot) + N$, where N is a random noise that has been properly adjusted to attain ϵ-DP. Adding Laplace-distributed noise whose scale has been adjusted to the global sensitivity of the query f is probably the most common method to derive κ_f (although other methods have been proposed [15,16,22]).

Definition 2 (L_1-sensitivity)**.** The L_1-sensitivity, Δ_f, of a function $f : \mathcal{D}^n \to \mathbb{R}^d$ is the maximum variation of f between data sets that differ in one record:

$$\Delta_f = \max_{d(D,D')=1} \left\| f(D) - f(D') \right\|_1.$$

Proposition 1. *Let $f : \mathcal{D}^n \to \mathbb{R}^d$ be a query function. The mechanism $\kappa_f(D) = f(D) + (N_1, \ldots, N_d)$, where N_i are drawn i.i.d. from a Laplace$(0, \Delta_f/\epsilon)$ distribution, is ϵ-DP.*

2.2. State of the art

In [20,21,24,25], DP data sets are generated via record masking. In these works, microaggregation is employed to reduce the sensitivity of the queries used to generate the DP data sets: rather than querying for each original record, representatives of the microaggregation clusters are queried. Since a cluster representative is an aggregation of the records in the cluster, it is less sensitive to changes than any single record. The amount of sensitivity reduction depends on how such representative values are computed.

Specifically, in [24,25], multivariate microaggregation is used to partition the original data set into clusters of k or more records. The DP data set is derived by querying the centroid (arithmetic average) of each cluster. Since multivariate microaggregation with minimal cluster size k over all the attributes ensures k-anonymity, one can regard this approach as combining k-anonymity and DP. However, a special type of microaggregation called *insensitive microaggregation* is required. The main reason is that, in standard microaggregation algorithms, changing one value in one record may yield a completely unrelated set of clusters, which would not reduce sensitivity; although it is certainly unlikely that changing one record value changes all clusters, to guarantee DP one needs to consider the worst case. In insensitive microaggregation, one requires changes in a single record to produce a set of clusters that differ (at most) in one record. In this way, the sensitivity of a centroid divides the sensitivity of the original records by the corresponding cluster size. The downside of insensitive microaggregation is that it yields worse within-cluster homogeneity than standard microaggregation and, hence, higher information loss. Furthermore, the minimum cluster size k grows with the data set size, as we show next. Let n be the number of records in a data set. To obtain a DP version of it, n/k centroids are released, each one computed on a cluster of cardinality k and having sensitivity Δ/k, where Δ is the sensitivity of the original records (strictly speaking, Δ is the sensitivity of a query Q_i that returns the i-th original record, for $i = 1$ to the number of records). Hence, the sensitivity of the whole data set to be released is $n/k \times \Delta/k$. Thus, for numerical data sets, Laplace noise with scale parameter $(n/k \times \Delta/k)/\epsilon$ must be added to each centroid to obtain an ϵ-DP output. For this approach to reduce the amount of noise required to attain ϵ-DP, the inequality $n/k \times \Delta/k \leq \Delta$ must hold. Equivalently, one needs $k \geqslant \sqrt{n}$.

To circumvent the previous problems of [24,25], an alternative approach was proposed in [20,21] based on individual-ranking (IR) microaggregation. Since IR microaggregates only one attribute at a time, it achieves insensitivity while avoiding the information loss penalty of multivariate insensitive microaggregation. On the other