Information Sciences 250 (2013) 200–214

Contents lists available at ScienceDirect

Information Sciences

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

Optimal data-independent noise for differential privacy

Jordi Soria-Comas, Josep Domingo-Ferrer *

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

ARTICLE INFO

Article history:
Received 7 March 2012
Received in revised form 12 June 2013
Accepted 1 July 2013
Available online 12 July 2013

Keywords:
Data privacy
Differential privacy
Noise addition
Privacy-preserving data mining
Statistical disclosure control

ABSTRACT

ε-Differential privacy is a property that seeks to characterize privacy in data sets. It is formulated as a query-response method, and computationally achieved by output perturbation. Several noise-addition methods to implement such output perturbation have been proposed in the literature. We focus on data-independent noise, that is, noise whose distribution is constant across data sets. Our goal is to find the optimal data-independent noise distribution to achieve ε-differential privacy. We propose a general optimality criterion based on the concentration of the probability mass of the noise distribution around zero, and we show that any noise optimal under this criterion must be optimal under any other sensible criterion. We also show that the Laplace distribution, commonly used for noise in ε-differential privacy, is not optimal, and we build the optimal data-independent noise distribution. We compare the Laplace and the optimal data-independent noise distributions. For univariate query functions, both introduce a similar level of distortion; for multivariate query functions, optimal data-independent noise offers responses with substantially better data quality.

1. Introduction

ε-Differential privacy [6,5] is a statistical disclosure control methodology for queryable databases. A remarkable fact about ε-differential privacy is that, unlike other methods, it is not based on the understanding that some specific output may be disclosive. Instead it seeks to limit the knowledge gain that any database user may obtain from a response.

The initial formulation of differential privacy only in a query-response setting was justified by previous results [1,3,4] showing the imposibility of answering a large number of queries with a bounded error while preserving the utility of the data. This seemed to preclude using differential privacy for data set releases. In [2,9] it was shown that those previous results were overpessimistic, which opened the door to the generation of ε-differentially private data sets [13]. Nonetheless, the initial query-response formulation remains the basic use case for differential privacy, and the methods developed for such use case can also be leveraged to generate ε-differentially private data sets.

Computationally, ε-differential privacy is usually achieved by output perturbation; responses are computed on the real data and masked by adding a random noise. Other methods for attaining ε-differential privacy not based on directly adding noise to the real query response are, for instance, the exponential mechanism [14], and the sample and aggregate framework [16]. For a more complete overview of differential privacy and, in particular, of a variety of methods used to attain it, see [7,8,12].

Several methods to generate the required random noise have been proposed in the differential privacy literature. We classify them in two categories, according to whether the noise distribution takes the original data into account: data-independent noise and data-dependent noise. Methods based on adding data-independent noise conform the most basic approach. Laplace noise addition [6] belongs to this category. Methods based on adding data-dependent noise are more complex, but

* Corresponding author. Tel.: +34 977558270; fax: +34 977559710.
E-mail addresses: jordi.soria@urv.cat (J. Soria-Comas), josep.domingo@urv.cat (J. Domingo-Ferrer).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.07.004

usually they lead to less distortion being introduced. Calibration to smooth sensitivity [16] belongs to the data-dependent noise category. In this paper we focus on the data-independent noise approach, which is the most frequently used one (and the one that was first proposed).

To maximize the utility of the results provided by ε-differential privacy, the magnitude of the random noise should be as small as possible. Some criticisms have appeared to the data utility that results from using Laplace noise addition as the mechanism to obtain ε-differential privacy [15,17]. The question of the optimality of Laplace noise addition arises: is it possible to achieve ε-differential privacy with substantially more data utility using other noise distributions?

Our goal is to determine the optimal distribution to achieve ε-differential privacy with data-independent random noise. We will limit our discussion to absolutely continuous random noise distributions, as they provide the greatest level of generality. Similar results can also be obtained for discrete random noise; however, this type of noise is only applicable in very specific circumstances.

By using an optimal noise, the distortion required to achieve a certain level ε of differential privacy is minimized. This may lead to under-protection if the disclosure limitation offered by ε-differential privacy is measured by how much noise is added to the data (as in traditional noise addition for disclosure control, see [11]), rather than by the theoretical guarantee offered by differential privacy in terms of ε (see Definition 1). In what follows, we assume that a protection level ε is chosen such that the theoretical guarantee provides sufficient protection.

Before going into the details of the construction of the optimal data-independent random noise we briefly introduce some basic concepts about ε-differential privacy. The following formal definition of ε-differential privacy can be found in [5].

Definition 1. A randomized function κ gives ε-differential privacy if, for all data sets D and D' differing in at most one row (that is, one record), and all $S \subset Range(\kappa)$ measurable, it holds that

$$P(\kappa(D) \in S) \leqslant e^{\varepsilon} \times P(\kappa(D') \in S) \tag{1}$$

The interpretation of the above definition is as follows. Assume that we want to query the database with a function $f : \mathcal{D} \longrightarrow \mathbb{R}^d$ that maps each of the data sets to a value in $\mathbb{R}^d$. ε-Differential privacy returns a randomization κ_f of f such that the probability of obtaining a given response changes at most by a factor $\exp(\varepsilon)$ when adding or removing a record from the database.

The privacy guarantee provided by ε-differential privacy to an individual is that, no matter whether the record containing the individual's data is included in the data set, the responses returned for any query will be similar. Hence, the presence or absence of the individual's data are not easily noticed, which means privacy for the individual.

Definition 1 is stated in terms of data sets D and D' differing in at most one row. Data sets differing in one row, called neighbor data sets, can be obtained from one another in two ways: either by adding/removing one record (as assumed in [5]) or by modifying a single record (as assumed in [6]). Depending on the definition used, the magnitude of the required random noise may slightly change, but the methods used for noise calibration remain the same. For the sake of concreteness, in the sequel we will focus on addition and removal of records.

The randomization κ in Definition 1 can be seen as the addition of a random noise, whose distribution may depend on the data set D, to the real value of the query function $f(D)$:

$$\kappa(D) = f(D) + (\kappa(D) - f(D)) = f(D) + Y(D)$$

If the distribution of the random noise depends on the actual data set D, we say that noise is data-dependent. If the random noise distribution is constant across data sets, we say that noise is data-independent. As mentioned above, we focus on data-independent noise.

Data-independent noise for ε-differential privacy is usually implemented as proposed by Dwork et al. in [6]. These authors proposed to generate noise using a Laplace distribution whose scale parameter depends on the maximum variation of the query function between neighbor data sets. This variation is known as the L_1-sensitivity of the function, and it is formally introduced next.

Definition 2 (L_1-*sensitivity*). The L_1-sensitivity of a function $f : \mathcal{D} \to \mathbb{R}^d$ is defined as

$$\Delta f = \sup_{D,D'} \|f(D) - f(D')\|_1 = \sup_{D,D'} \sum_{i=1}^{d} |f_i(D) - f_i(D')|$$

where f_i is the ith component of f, for all D, D' such that one can be obtained from the other by adding or removing one record.

In order to reach ε-differential privacy, Laplace-distributed random noise with zero mean and $\Delta f/\varepsilon$ scale parameter is added to each component of f.

1.1. Contribution and plan of this paper

The randomized function κ that provides ε-differential privacy can be viewed as the addition of a random noise to the real value of the query function f. Hence, the quality of the resulting differentially-private data critically depends on the noise distribution. Taking this into account, the aim of this paper is to build the optimal data-independent noise distribution for ε-differential privacy.