WP. 24
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (iv): Balancing data quality and data confidentiality

On differential privacy and data utility in SDC

Prepared by Jordi Sòria-Comas and Josep Domingo-Ferrer, Universitat Rovira i Virgili, Spain

On differential privacy and data utility in SDC

Jordi Sòria-Comas and Josep Domingo-Ferrer

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili.
UNESCO Chair in Data Privacy, {jordi.soria, josep.domingo}@urv.cat

Abstract: Differential privacy is a statistical disclosure control methodology for queryable databases. The disclosure risk limitation provided by differential privacy is based on limiting the contribution of any single record to the query result. In practice, differential privacy is achieved through output perturbation; the real query value is computed and masked before being released. Several approaches to differential privacy have been proposed, as well as several methods to calibrate the random noise. With the goal of achieving the best data quality, we analyze the different methods to achieve differential privacy for several types of query functions.

1 Introduction

Differential privacy [1] is a methodology to limit disclosure in queryable statistical databases; it guarantees that the response obtained by querying a database, before and after the contribution of any single individual, is statistically similar. If the information that can be extracted from a database before and after an individual's contribution is similar, then the risk of disclosure is limited. This is a quite generic approach, in the sense that it does not assume that some specific data combinations may lead to disclosure; instead, it limits the information that can be extracted from queries. In practice, differential privacy is achieved through output perturbation; the real value of the query response is computed and masked, by adding a random noise, before release.

The most common approach to differential privacy limits the knowledge gain between neighbor databases D and D' such that one can be obtained from the other by adding or removing a single record. Let us assume that, in addition to all records in D', D contains an additional record r with the information on individual I, that is $D = D' \cup \{r\}$. As D' contains no information on individual I, the level of privacy for I when querying D' is maximum. Since the knowledge gain between responses to queries to D and D' is limited (by the assumption of differential privacy), the disclosure risk when querying D is limited. The following definition of differential privacy can be found in [1].

Definition 1. A randomized function κ gives ε-differential privacy if, for all databases D, D' such that one can be obtained from the other by adding or removing a single record, and all $S \subset Range(\kappa)$

$$\mathrm{P}(\kappa(\mathrm{D}) \in \mathrm{S}) \leq \mathrm{e}^{\varepsilon} \times \mathrm{P}(\kappa(\mathrm{D}') \in \mathrm{S})$$

Note that the definition introduces a parameter ε. This parameter allows us to select the level of protection that we want to achieve. For example, by setting ε to 0.1, we are limiting the knowledge gain to about 11%. By increasing ε, the knowledge gain is increased and the protection level decreased; and conversely.

Another approach to differential privacy [4] limits the knowledge gain between neighbor databases D and D' such that one can be obtained from the other by modifying a single record. The idea behind this approach is to limit the knowledge gain between the database that contains the real data and a database that contains some fake data for individual *I*. This approach limits the comparison to databases with the same number of records *n*, and results in the following alternative definition.

Definition 2. A randomized function κ gives ε-differential privacy if, for all databases D, D' with cardinality n such that one can be obtained from the other by modifying a single record, and all $S \subset Range(\kappa)$

$$\mathrm{P}(\kappa(\mathrm{D}) \in \mathrm{S}) \leq \mathrm{e}^{\varepsilon} \times \mathrm{P}(\kappa(\mathrm{D}') \in \mathrm{S})$$

Criticisms to differential privacy have been raised [3], mainly regarding the applicability of the differential privacy methodology, the level of privacy protection achieved, and the data quality that can be expected for the differentially private responses.

2 Adjusting the noise

It has been mentioned above that differential privacy is an output perturbation methodology. When a user sends a query f against database D, the real response $f(D)$ is computed and masked by adding a random noise $\mathrm{N(D)}$ before being released, so the response is $\kappa_{\mathrm{f}}(\mathrm{D}) = \mathrm{f}(D) + \mathrm{N(D)}$. To achieve better data quality, the magnitude of the random noise must be as small as possible.

Several methods to calibrate the random noise have been proposed. We classify them in two categories: data-independent and data-dependent noises. Data-independent noises are those whose distribution is constant across databases, while data-dependent noises may have different distributions for different databases.

The most common method to achieve differential privacy is to use a data-independent noise following a Laplace distribution with zero mean and scale parameter that depends on the maximum change experienced by the query distribution between neighbor databases [2].

Using a data-independent noise is fine if the change in the query function between neighbor databases is constant. Otherwise, to achieve the desired level of protection between the pair of neighbor databases with the greatest change in the query function, we are overprotecting those databases having less variability of the query function value w.r.t. their neighbor databases. A mechanism to adjust the distribution of the random noise at each database to the variability of the query function between that database and its neighbors was proposed in [5]. The proposal is based on the concept of indistinguishability, which is similar to the concept of differential privacy, although more general. At first sight, adjusting the random noise to the variability of the query function at each database may seem an improvement over data-independent noises; however, the probability distributions eligible for data-dependent noises are not as good as those eligible for data-independent noises. As a result, using a data-dependent noise may sometimes lead to less data quality.

3 Data quality

The most typical example of differential privacy considers the query function to be the absolute frequency. The properties of the absolute frequency function make it a very suitable function for differential privacy. The change in the query function between databases that differ in one row (whether addition/removal or modification of records is performed) is constant. This means that, by using a data-independent noise, we are not overprotecting any database, but providing the exact level of protection required. However, other classes of query functions may not display such a good behavior.

An analysis of the data quality achievable by each of the methods used to obtain a differentially private response is required to determine the usability of differential privacy. Such an analysis must take into account all the possible sources of variability in the data quality. The factors to be taken into account are:

- The type of the query function. We have seen that the absolute frequency is well-suited for differential privacy. However, other types of query functions such as the maximum and the minimum require the introduction of a greater amount of noise to achieve the desired level of protection.

- The definition (approach) to differential privacy. The data quality achieved by a differential private mechanism may depend on the definition of differential privacy used, *i.e.* Definition 1 or Definition 2.
- The type of random noise. A given type of random noise may be better suited than another. This depends basically on the query function and on the definition of differential privacy. For example, when querying the whole database for the relative frequency of some property and using Definition 2, a data-independent noise provides better results (note that the variability between neighbor databases in this setting is constant). If instead of querying the whole database, we query only some of the records, the variability between neighbor databases is not constant anymore, and using a data-dependent noise may be better.
- The actual distribution of the noise. We have classified the methods to calibrate the random noise depending on the type of random noise: data-independent or data-dependent. However, within each of these two classes, many probability distributions may be eligible.

References

[1] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel,V. Sassone, and I. Wegener (eds.) *Automata, Languages and Programming*, LNCS 4052, pp. 1-12. Springer, 2006.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin (eds.) *Theory of Cryptography Conference-TCC 2006*, LNCS 3876, pp. 265-284, 2006. Springer, 2006.

[3] K. Muralidhar and R. Sarathy. Does differential privacy protect Terry Gross' privacy? In J. Domingo-Ferrer and E. Magkos (eds.) *Privacy in Statistical Databases-PSD 2010*, LNCS 6344, pp. 200-209, 2010.

[4] K. Nissim. Private data analysis via output perturbation. In A. K. Elmagarmid, C. C. Aggarwal, and P. S. Yu (eds.) *Privacy-Preserving Data Mining*, volume 34 of *The Kluwer International Series on Advances in Database Systems*, pp. 383-414. Springer US, 2008.

[5] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In D. S. Johnson and U. Feige (eds.) *Proc. of the 39th Annual ACM Symposium on Theory of Computing -STOC 2007*, pp. 75-84. ACM, 2007.