Individual Differential Privacy: A Utility-Preserving Formulation of Differential Privacy Guarantees

Jordi Soria-Comas, Josep Domingo-Ferrer, *Fellow, IEEE,* David Sánchez, and David Megías, *Member, IEEE*

***Abstract*—Differential privacy is a popular privacy model within the research community because of the strong privacy guarantee it offers, namely that the presence or absence of any individual in a data set does not significantly influence the results of analyses on the data set. However, enforcing this strict guarantee in practice significantly distorts data and/or limits data uses, thus diminishing the analytical utility of the differentially private results. In an attempt to address this shortcoming, several relaxations of differential privacy have been proposed that trade off privacy guarantees for improved data utility. In this work, we argue that the standard formalization of differential privacy is stricter than required by the intuitive privacy guarantee it seeks. In particular, the standard formalization requires indistinguishability of results between *any* pair of neighbor data sets, while indistinguishability between *the actual* data set and its neighbor data sets should be enough. This limits the data controller's ability to adjust the level of protection to the actual data, hence resulting in significant accuracy loss. In this respect, we propose *individual differential privacy*, an alternative differential privacy notion that offers *the same privacy guarantees as standard differential privacy to individuals* (even though not to groups of individuals). This new notion allows the data controller to adjust the distortion to the actual data set, which results in less distortion and more analytical accuracy. We propose several mechanisms to attain *individual differential privacy* and we compare the new notion against standard differential privacy in terms of the accuracy of the analytical results.**

***Index Terms*—Data privacy, Data utility, Differential privacy.**

I. Introduction

The development of information technologies has boosted the collection of data about individuals. Data stores containing personal data are a valuable resource to support decision making both in the public and the private sector. However, individuals have growing concerns on the use of their potentially sensitive data. In order to guarantee the fundamental right of individuals to privacy, appropriate data protection measures to limit disclosure risk must be implemented before making data available for analysis.

J. Soria-Comas, J. Domingo-Ferrer and D. Sánchez are with the UNESCO Chair in Data Privacy, Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia. E-mail {jordi.soria, josep.domingo, david.sanchez}@urv.cat

D. Megías is with the Estudis d'Informàtica, Multimèdia i Telecomunicació, Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Av. Carl Friedrich Gauss 5, E-08860 Castelldefels, Catalonia. E-mail dmegias@uoc.edu

Essentially, two main approaches exist in statistical disclosure control [1], [2] to limit the disclosure risk: (i) non-interactive protection, whereby a protected version of the original data set collected from the data subjects is generated and released, and (ii) interactive protection, whereby a user-queried data analysis is performed on the original data set, and then a protected version of the results is returned to the user. When the type of data analysis is not known at the time of data protection, the former approach is the only viable solution; however, for a fixed data analysis known beforehand, interactive protection should be preferred as it allows adjusting the level of protection to the analysis being performed, which makes it possible to maximize the accuracy of the results.

In this work, we focus on the interactive setting, which differential privacy [3] has substantially advanced. The main contribution of differential privacy is the strong privacy guarantee it provides: while it does not prevent disclosure, it guarantees that a disclosure is equally likely whether or not any particular individual contributes her data. This is diametrically opposed to the aim of privacy models focusing on the non-interactive setting [4], [5], [6], [7], which seek to provide absolute guarantees against specific types of intruders (with a specific knowledge on the data to be protected). The problem of the latter models is that real intruders may differ from the specific intruders considered by the models; in particular, real intruders may know more than assumed.

However, despite its popularity among researchers and the step forward it offers in terms of privacy guarantees, differential privacy is only being deployed to a limited extent in real-world applications. The basic reason is the poor accuracy/utility of differentially private results. Except for a number of well-behaved applications (queries that are stable to modification of one record), differential privacy has too large an impact on utility for it to be widely used in data analysis/mining (see related remarks in Section III). This is also confirmed by the amount of relaxations of differential privacy that have been proposed (see Section III), which seek to improve the accuracy of the results by trading it off against privacy guarantees.

In any case, the intuition behind differential privacy, that is, *"the presence or absence of an individual in a data set should not significantly modify the results of the analysis"*, is very adequate for disclosure risk limitation: if the data on an individual have a significant impact on the results of an analysis, most probably the privacy of this individual is at risk. Thus, intuitively, differential privacy ensures the data are adequately protected.

However, we argue here that the standard formalization of differential privacy is stronger (and, thus, leads to more data distortion and less useful results) than what the above data protection intuition requires. This can be seen from two different but related points of view:

- Differential privacy assumes the presence of a trusted party that holds the data set, receives queries submitted by the users and returns differentially private results for these queries. In the standard differential privacy formulation, the trusted party is not allowed to use its knowledge of the actual data set when computing the noise to be added to the query response in order to protect privacy.
- Differential privacy provides provable privacy guarantees to groups of data subjects (in addition to those that derive from the guarantees given to each data subject individually). These guarantees for groups go beyond the previously mentioned intuition behind differential privacy, which is about protecting single individuals.

While being more demanding than required by the intuitive notion of differential privacy may be meaningful, we argue in the following sections that enforcing an actually stronger privacy notion may have a significant negative impact on the accuracy/utility of the protected results. Thus, if data utility is a priority, the suitability of including additional guarantees (i.e., group-based differential privacy) should be carefully pondered. Besides, in statistical disclosure control, the goal is to prevent accurate inferences on single individuals, but accurate inferences on groups (statistical and aggregate computations) are usually viewed as legitimate analyses.

In this line, we propose here *individual differential privacy*, a novel formalization of differential privacy that is aligned with the intuitive differential privacy guarantees. Our approach limits the protection to the individuals in the data set (as suggested in the intuitive notion), rather than extending it to groups of individuals (as formalized in the standard definition of differential privacy). In practice, this means that the trusted party managing the data can make use of its knowledge of the actual data set at the time of query response and downwardly adjust the distortion to the actual data; as a consequence of the lower data distortion, the accuracy/utility of the protected results can be significantly improved in comparison with the standard formulation of differential privacy.

Nonetheless, we will also indicate how to extend individual differential privacy to group differential privacy, in case the data controller wants to offer privacy guarantees for groups while still leveraging his knowledge on the actual data set when computing the distortion to be added for protection.

The rest of this paper is organized as follows. Section II provides background on the standard formulation and the usual mechanisms to attain differential privacy. Section III reviews related work aimed at improving the accuracy/utility of differentially private results and highlights the differences with our approach. In Section IV, we present and formalize our *individual differential privacy* model and in Section V we propose several mechanisms to satisfy it that exploit its less strict formulation to reduce data distortion. Section VI analyzes the accuracy of individually differentially private results of several common queries and compares their accuracy against the one obtained using standard differential privacy. The final section presents the conclusions along with some future research lines.

II. Background on Differential Privacy

Differential privacy was originally proposed in [3] as a privacy model for the interactive setting, that is, to protect the results of queries to a database. In this setting, a differentially private sanitization mechanism sits between the user submitting queries and the database controller answering them. To preserve the privacy of individuals, the sanitization mechanism must guarantee that the contribution of each individual's data to a query result is limited (according to an ϵ parameter).

Definition 1 (ϵ-differential privacy)**.** A randomized function κ gives ϵ-differential privacy (or 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 $S \subset Range(\kappa)$, we have

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

Let D be the data set that is to be protected. Given a query f, the goal in DP is to find an approximation to f that satisfies the privacy requirements stated in Definition 1. Let us call κ_f such an approximation. The value of $\kappa_f(D)$ is then returned as the query result, instead of the actual value $f(D)$.

In the following, we refer to the differentially private condition in terms of indistinguishability: the distributions of the responses to the queries between data sets that differ in one record are similar.

An interesting property of DP, which the privacy models in the k-anonymity family (k-anonymity [4], l-diversity [5], t-closeness [6], etc.) do not have, is composition: composing several differentially private results still satisfies DP, although with a different ϵ parameter value. Several composition theorems have been stated. The most basic ones are:

Theorem 1 (Sequential composition)**.** *Let κ_1 be a randomized function giving ϵ_1-DP and κ_2 a randomized function giving ϵ_2-DP. Then, any deterministic function of (κ_1, κ_2) gives $(\epsilon_1 + \epsilon_2)$-DP.*

Theorem 2 (Parallel composition)**.** *Let κ_1 and κ_2 be randomized functions giving ϵ-DP. If κ_1 and κ_2 are applied to* disjoint *data sets or subsets of records, any deterministic function of (κ_1, κ_2) gives ϵ-DP.*

The computational mechanism to attain DP is often called a differentially private sanitizer. Differentially private mechanisms for numerical data can be seen as noise addition mechanisms that, rather than returning the actual query value $f(D)$, mask it by adding some noise.

The amount of noise that needs to be added depends on the variability of the query function between neighbor data sets. This noise may be: (i) independent of the data set, which requires adjusting the noise to the maximum variability between neighbor data sets, or (ii) dependent on the data set, which allows adjusting the level of noise to the variability of f in the actual data set.

A. Noise calibration to the global sensitivity

The global sensitivity measures the maximum variability of a function between neighbor data sets.

Definition 2 (l_1-sensitivity (a.k.a. global sensitivity))**.** The l_1-sensitivity of a function $f : \mathcal{D} \to \mathbb{R}^k$ is

$$\Delta f = \max_{\substack{x,y \in \mathcal{D} \\ d(x,y)=1}} \|f(x) - f(y)\|_1 ,$$

where $d(x, y)$ means that data sets x and y differ in one record.

In calibration to the global sensitivity, DP is attained by adding a noise to the query response that is proportional to the global sensitivity; that is, the maximum variability of the query response in the domain of the data. Several noise distributions are possible (e.g. the Laplace distribution [8], the optimal absolutely continuous distribution [9], and the discrete Laplace distribution [10]). For the sake of conciseness, we focus on the commonly used Laplace distribution.

Proposition 1. *Let f be a query function with values in $\mathbb{R}^k$. The mechanism $\kappa_f(X) = f(X) + (N_1, \ldots, N_k)$, where N_i are independent and identically distributed random noises drawn from a Laplace$(0, \epsilon/\Delta f)$ distribution, is ϵ-differentially private.*

B. Noise calibration to the smooth sensitivity

Even if the variability of the query function in the actual data set is small, quite often its variability in the data domain (i.e., global sensitivity) is large. In these cases, the use of a noise calibrated to the global sensitivity may seriously and unjustifiedly damage data, but a data-dependent noise may still permit accurate results. Given a data set D, the variability of a function between D and its neighbor data sets is known as the *local sensitivity.*

Definition 3 (Local sensitivity [11])**.** Let f be a function that is evaluated at data sets and returns values in $\mathbb{R}^k$. The local sensitivity of f at D is

$$LS_f(D) = \max_{y:d(y,D)=1} \|f(y) - f(D)\|_1 .$$

Obviously, the global sensitivity upper-bounds the local sensitivity. Moreover, the local sensitivity is usually small, except for especially ill-conditioned data sets. In this case, the gap between the global and the local sensitivity can be large. This is illustrated in the following example for a function that returns the median of a list of values.

Example 1. Consider a data set $X = \{x_1, \ldots, x_n\}$, where each record corresponds to a value in $\{0, 1\}$. For the sake of simplicity, we assume that the number of records is odd, say $n = 2m+1$, so that the median corresponds to a single record, the $m + 1$-th record. The global sensitivity of the median is 1, since we can consider the neighbor data sets:

$$\{0, \overset{m+1}{\ldots}, 0, 1, \overset{m}{\ldots}, 1\} \to median = 0,$$
$$\{0, \overset{m}{\ldots}, 0, 1, \overset{m+1}{\ldots}, 1\} \to median = 1.$$

The local sensitivity is, except for the two previous data sets, always 0. The reason is that, except for the previous data sets, changing the value of a record does not modify the median.

In most cases, releasing the value of a query with a magnitude of noise proportional to the local sensitivity (rather than the global sensitivity) would result in a significantly more accurate response. However, using the local sensitivity in mechanisms designed for global sensitivity does not yield DP.

Example 2. Consider the data sets: $\{0, \overset{m+2}{\ldots}, 0, 1, \overset{m-1}{\ldots}, 1\}$ and $\{0, \overset{m+1}{\ldots}, 0, 1, \overset{m}{\ldots}, 1\}$. In both cases the median is 0, but the local sensitivity differs; it is 0 in the first case and 1 in the second one:

	median	*local sensitivity*
$X = \{0, \overset{m+2}{\ldots}, 0, 1, \overset{m-1}{\ldots}, 1\}$	0	0
$X' = \{0, \overset{m+1}{\ldots}, 0, 1, \overset{m}{\ldots}, 1\}$	0	1

Given that the local sensitivity for X is 0, adding a noise proportional to the local sensitivity does not modify the median; thus, the probability of getting 1 is 0. For ϵ-DP to be satisfied, the probability of getting 1 for X' must also be 0. However, that is not the case, since the local sensitivity for X' is different from 0.

The previous example shows that the amount of noise used to protect a data set should not only be proportional to its local sensitivity but also take into account the local sensitivity of neighbor data sets. Considering the local sensitivity of neighbor data sets leads to the notion of *smooth sensitivity.*

Definition 4 (Smooth sensitivity)**.** For $\beta > 0$ and a data set $D \in \mathcal{D}$, the β-smooth sensitivity of f at D is defined as

$$S_{f,\beta}(D) = \max_{y \in \mathcal{D}} (LS_f(y) \exp(-\beta d(D, y))).$$

The greater the parameter β, the smaller the dependence of the smooth sensitivity on the local sensitivity of neighbor data sets. Thus, the amount of noise required to attain ϵ-DP must depend on other factors apart from the smooth sensitivity. In particular, a special kind of distributions, known as (α, β)-admissible distributions [11], must be used. The downside of these distributions is that they are heavy-tailed.

III. Related Work

Even though DP is appreciated for the strong privacy guarantees it offers, its practical deployment is hampered by the poor accuracy offered by differentially private mechanisms.

Works that discuss the accuracy limitations of differentially private results in different contexts include [12], [13], [14], [15], [16]. Although these accuracy issues have hindered the practical adoption of DP, they have also boosted further research to find fixes. This research has taken two main lines: (i) to come up with novel differentially private mechanisms that improve the accuracy of the results, and (ii) to propose relaxations of DP that require less data distortion and allow for more accurate results.

An outcome of the first line is the design of general mechanisms that improve accuracy with respect to the basic Laplace noise addition mechanism. In [11], the calibration of Laplace noise to the global sensitivity of the data is replaced by the calibration of suitable noises to the smooth sensitivity. On the other hand, the authors of [9] showed that the Laplace distribution is not optimal to attain DP based on calibration to the global sensitivity. They described and constructed the optimal absolutely continuous distributions: essentially, a distribution is optimal if the probability mass is as concentrated as possible around zero given the DP constraints. Other outcomes of the first line of research are mechanisms that are less sensitive. In [17], [18], [19], several methods based on microaggregation of records are proposed to generate differentially private data sets. In [20], the dependence between attributes is analyzed to reduce the dimensionality in the computation of differentially private histograms. Other works that try to improve the accuracy in histogram publication are [21], [22]. In [23], a differentially private alternative to the ID3 algorithm for learning decision trees is proposed.

The second line of research, focused on finding relaxations of DP, has been active since the inception of DP. In [24], the concept of δ-approximate ϵ-indistinguishability is presented (a.k.a. (ϵ, δ)-indistinguishability in [11]). This relaxation allows some additional margin δ to the requirements in DP. In [25], the notion of (ϵ, δ)-probabilistic differential privacy (a.k.a. (ϵ, δ)-pdp) is proposed. Rather than allowing some additional margin, it requires ϵ-DP to be satisfied with probability greater than $1 - \delta$. In other words, the probability that the adversary gains significant information about an individual is, at most, δ. Yet another relaxation is given in [15], who assume that confidential data become less sensitive over time, which allows relaxing privacy parameters for older data. In [26], a relaxation is presented that restricts the definition of neighbor data sets: they are no longer data sets differing in any record, but in a record within a certain subset. In [27], an alternative relaxation of DP, called (μ, τ)-concentrated differential privacy, is proposed. Similarly to (ϵ, δ)-pdp, concentrated differential privacy allows the ratio of probabilities to be arbitrarily large with a small probability that is determined by the parameters μ and τ.

In all the above relaxations, the accuracy gain is obtained by allowing the differentially private condition to be broken: the presence or absence of an individual may leak some information, although not too much or only with a small probability. The relaxation that we propose in the next section is different in the sense that the privacy guarantees for individuals envisioned in the original definition of DP are preserved.

IV. Individual differential privacy

The existing relaxations of DP violate strict DP because, under some circumstances, they permit the probability of responses to differ significantly between neighbor data sets. In this section, we take a different approach to improve data accuracy/utility. Rather than proposing another relaxation, we analyze the intuitive privacy guarantees that DP seeks to capture and we suggest an alternative definition that allows exactly attaining these privacy guarantees and nothing more.

A. Intuitive view of differential privacy

As introduced above, DP 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.

We described in Section II that the standard notion of DP is formalized in terms of indistinguishability of the response to queries between neighbor data sets. However, prior to this formalization, [3] provided an intuitive description of the privacy guarantees aimed at by DP:

> Any given disclosure will be, within a small multiplicative factor, just as likely whether or not the individual participates in the database. As a consequence, there is a nominally higher risk for the individual who participates, and only a nominal gain for the individual who conceals or misrepresents her data.

Thus, individuals should not be reluctant to participate in the data set. After all, the risk of disclosure is only very marginally increased by participation.

Providing relative (rather than absolute) privacy guarantees is the only sensible approach to deal with intruders having arbitrary side knowledge. Absolute privacy guarantees against arbitrary side knowledge are incompatible with releasing accurate statistics. This is illustrated in [3] with a simple example about the height of an individual I: if the intruder knows that the height of I is two inches above the average height of the population of a country, then releasing an accurate approximation of the average height of the population provides the intruder with an accurate estimate of I's height. Here it is important to notice that such disclosure happens even if I does not participate in the data set. Thus, the disclosure is not associated to I's participation in the data set, but to the accurate side knowledge. This illustrates that differential privacy cannot guarantee that disclosure will not happen; what it guarantees is that, should disclosure happen, it will not be due to the participation of any particular individual.

B. Formalization of individual differential privacy

DP is the result of formalizing the intuitive view of privacy described in Section IV-A into a rigorous mathematical definition. However, the fact that in this formalization (see Definition 1) results are required to be indistinguishable between *any* pair of neighbor data sets is more stringent than required by the intuitive view of privacy that is described above. Let us explain this in greater detail.

Consider an individual I who has to decide between participating in a data set or not. To neutralize any reluctance by I to disclose her private information, I is told that query answers based on the data set will not allow anyone to learn anything that was not learnable without I's presence; this is precisely the intuitive privacy guarantee DP offers. Although DP controls the access to a data set regardless of the way the data were collected (either through voluntary participation or not), thinking in terms of voluntary participation is clarifying: we can think that an individual accepts to participate if she regards privacy protection as sufficient.

To attain such privacy guarantees, DP requires the response to be indistinguishable between any pair of neighbor data sets. While such a requirement yields the target privacy guarantees indeed, it is an overkill because, in reality, queries are not answered on an arbitrary data set, but on the actual data set held by the trusted party. In other words, if D is the collected data set, the target privacy guarantees can be attained by just requiring indistinguishability of the responses between D and its neighbor data sets. Notice that, although the data set D is not known until all the individuals have made their decisions about participating/contributing to it, the presence of a trusted party that controls the access to D makes it possible to give the previously described indistinguishability guarantees. After all, the data set D is known to the trusted party at the time of query response.

According to the previous discussion, we propose the following privacy model, which we call ϵ-individual differential privacy.

Definition 5 (ϵ-individual differential privacy)**.** Given a data set D, a response mechanism $\kappa(\cdot)$ satisfies ϵ-individual differential privacy (or ϵ-iDP) if, for any data set D' that is a neighbor of D, and any $S \subset Range(\kappa)$ we have

$$\exp(-\epsilon) \Pr(\kappa(D') \in S) \leq \Pr(\kappa(D) \in S) \\ \leq \exp(\epsilon) \Pr(\kappa(D') \in S).$$

In line with Definition 1, we require the probability of any result to differ between neighbor data sets at most by a factor of $\exp(\epsilon)$. However, unlike in Definition 1, the role of the data sets D and D' is not exchangeable: D refers to the actual data set, and D' to a neighbor data set of D. The asymmetry between D and D' is relevant, because indistinguishability is achieved only between D and its neighbor data sets. As a side effect of this asymmetry, we need to explicitly enforce an upper bound ($\Pr(\kappa(D) \in S) \leq \exp(\epsilon) \Pr(\kappa(D') \in S)$) and a lower bound ($\exp(-\epsilon) \Pr(\kappa(D') \in S)) \leq \Pr(\kappa(D) \in S)$). This was not needed in Definition 1 because the upper bound could be obtained from the lower bound by exchanging the roles of D and D'.

To illustrate the implications of the differences between DP and iDP, consider a data set D whose records are either 0 or 1, and assume that we query the data set for the median. We consider two cases, $D_1 = \{0, 0, 0, 0, 1\}$ and $D_2 = \{0, 0, 0, 1, 1\}$, which are discussed next:

- In data set D_1, the median is 0. Moreover, the median remains 0 after modification of any single record. Since the median is unaffected by the contribution of any single individual, there is no way to learn anything about an individual from the median of D_1. Hence, according to iDP, the median of D_1 can be released without any masking. This is not the case according to DP, where the mere existence of a pair of neighbor data sets (not involving D_1) with different values for the median requires masking it (even if the median of D_1 discloses nothing about any individual).
- In data set D_2, the median is 0, but a change in a single record is enough for the median to become 1. Since a single individual may affect the result of the median, information about an individual might be learned from the value of the median. To limit the risk of disclosure, both DP and iDP require masking the outcome of the median.

In the previous example, we have considered boundary cases where the change of a single record can make it necessary or unnecessary to mask the query result. Let us now consider a different case, consisting of a data set D_3 with 99 binary records, of which 90 contain the value 1. Clearly, no single individual has any effect on the median of D_3. Thus, iDP does not require any masking for the answer to the median query. In contrast, under DP the possible existence of a data set D_4 containing forty-nine 0s and fifty 1s has to be considered, in which a single individual can change the median; since this data set must be indistinguishable from its neighbors given the value of its median, the answer to the median query must be masked even for D_3 (although D_3 and D_4 are far from neighbors).

Like in DP, two different notions of neighbor data sets D and D' are possible in iDP. We can either think of D' as a data set generated from D by modifying one record or as a data set generated from D by adding or removing one record. Although such a choice can make some difference in terms of query answers, both alternatives are similar as far as disclosure risk limitation is concerned. In the sequel, we assume that D' is generated from D by modifying one record.

Strictly speaking, *individual differential privacy* is a relaxation of DP where indistinguishability is only required between the actual data set D and its neighbor data sets. Thus, the following holds:

Proposition 2. *Any mechanism that satisfies ϵ-DP also* *satisfies ϵ-iDP for any actual data set D.*

However, *unlike other relaxations in the literature, individual differential privacy preserves the strict privacy guarantees that DP gives to individuals.* In this sense, rather than being regarded as a relaxation, individual differential privacy can be construed as a more precise formalization of the intuitive notion of privacy described in Section IV-A.

Even though we can satisfy ϵ-iDP by resorting to the mechanisms commonly used to satisfy ϵ-DP, doing so would squander the potential accuracy gains of ϵ-iDP. The reason is that, as discussed in Section II, being able to adjust the noise to the actual data set may significantly reduce its magnitude. In Section V, we design mechanisms that are specific to ϵ-iDP.

C. Disclosure limitation for groups of data subjects

If ϵ-iDP offers all the intuitive privacy guarantees of ϵ-differential privacy while allowing more accuracy, it is because ϵ-iDP offers *nothing more* than those intuitive guarantees. Let us discuss this issue here.

Even if DP seeks to protect privacy by limiting the impact of *each single* individual on a query response, Definition 1 also results in (limited) privacy guarantees for *groups* of individuals. For example, if data sets D_1 and D_3 differ in two individuals (two records), by considering the intermediate data set D_2 that differs in one individual from D_1 and in one individual from D_3, we have

$$\Pr(\kappa(D_1) \in S) \leq \exp(\epsilon)\Pr(\kappa(D_2) \in S),$$
$$\Pr(\kappa(D_2) \in S) \leq \exp(\epsilon)\Pr(\kappa(D_3) \in S),$$

which results in

$$\Pr(\kappa(D_1) \in S) \leq \exp(2\epsilon)\Pr(\kappa(D_3) \in S).$$

That is, ϵ-DP guarantees that the knowledge gain for groups of two individuals is limited by the factor $\exp(2\epsilon)$. More generally, the knowledge gain for a group of n individuals is limited by the factor $\exp(n\epsilon)$.

In contrast, with iDP, the fact that the response mechanism is adjusted to each concrete data set makes the concatenation of inequalities not possible. Specifically, for the data sets D_1, D_2 and D_3 above, iDP provides the upper bounds

$$\Pr(\kappa(D_1) \in S) \leq \exp(\epsilon)\Pr(\kappa(D_2) \in S),$$
$$\Pr(\kappa'(D_2) \in S) \leq \exp(\epsilon)\Pr(\kappa'(D_3) \in S),$$

where κ is a mechanism whose distortion is calibrated on D_1 and κ' is a mechanism whose distortion is calibrated on D_2. Note that the probability on the right-hand side of the first inequality does not match the probability on the left-hand side of the second inequality.

Therefore, with iDP we lose any *direct* privacy guarantees for groups of individuals. In other words, any disclosure risk limitation for a group that may subsist under ϵ-iDP is an indirect consequence of the disclosure risk limitation obtained by each individual in the group.

Direct group privacy guarantees not being part of the intuitive notion of DP (which is aimed at individuals), but rather a by-product of the formalization in Definition 1, we drop such group guarantees in order to improve data utility. However, the proposed iDP could be modified to retain privacy guarantees for groups of individuals, by requiring indistinguishability between the data set D and the data sets D' that differ from D in a group of records, as follows.

Definition 6 ($(\epsilon_1, \ldots, \epsilon_n)$-group differential privacy)**.** Given a data set D, a response mechanism κ satisfies $(\epsilon_1, \ldots, \epsilon_n)$-group differential privacy (denoted $(\epsilon_1, \ldots, \epsilon_n)$-gDP) if, for any data set D' at distance $i \in \{1, \ldots, n\}$ from D, and any $S \subset Range(\kappa)$ it holds

$$\exp(-\epsilon_i)\Pr(\kappa(D') \in S) \leq \Pr(\kappa(D) \in S)$$
$$\leq \exp(\epsilon_i)\Pr(\kappa(D') \in S).$$

Notice that, in general, the level of protection for groups should be smaller than for individuals. For example, in DP, the ϵ parameter grows linearly with the cardinality of the group and, thus, the level of indistinguishability decreases exponentially with the size of the group. As a result, even in standard DP, protection for groups of individuals is only noticeable for small values of ϵ (which on the other hand severely damage data utility) and for small groups. To attain an equivalent level of protection with gDP, it must be $\epsilon_i = i\epsilon$ for $i \in \mathbb{N}$.

D. Composition theorems

The composition theorems for DP described in Section II are also trivially satisfied by individual differential privacy.

Theorem 3 (Sequential composition)**.** *Let κ_1 be a mechanism giving ϵ_1-iDP and κ_2 a mechanism giving ϵ_2-iDP. Then, any deterministic function of (κ_1, κ_2) gives $\epsilon_1 + \epsilon_2$-iDP.*

Theorem 4 (Parallel composition)**.** *Let κ_1 and κ_2 be mechanisms giving ϵ-iDP. If κ_1 and κ_2 are applied to* disjoint *data sets or subsets of records, any deterministic function of (κ_1, κ_2) gives ϵ-iDP.*

The proofs of the theorems for iDP are not detailed because they are straightforward adaptations of the proofs of Theorems 1 and 2 for DP.

E. Privacy axioms

In [28], an axiomatization of the notions of privacy and utility is proposed. In particular, two desirable properties of privacy models are described: the axiom of transformation invariance and the privacy axiom of choice.

The axiom of transformation invariance says that post-processing of sanitized data must be safe as long as no sensitive information is incorporated into the post-processing.

Definition 7 (Axiom of transformation invariance)**.** Suppose we have a privacy definition, a privacy mechanism M

that satisfies this definition, and a randomized algorithm A whose input space is the output space of M and whose randomness is independent of both the data and the randomness in M. Then $M' = A \circ M$ must also be a privacy mechanism satisfying that privacy definition.

Proposition 3. *ϵ-iDP satisfies the axiom of transformation invariance.*

Proof. Let M be an ϵ-iDP mechanism, and let A be a randomized algorithm whose input space is the output space of M and whose randomness is independent of both the data and the randomness in M. According to [28], we can think of M and A as returning independent probability distributions. That is, $M : \mathcal{D} \to \mathrm{Dist}(E)$ assigns to each data set D a probability distribution over E such that, for each D' that differs in one record from D, we have $\exp(-\epsilon)\Pr(M(D') \in S) \leq \Pr(M(D) \in S) \leq \exp(\epsilon)\Pr(M(D') \in S)$. On the other hand, $A : E \to \mathrm{Dist}(F)$ assigns to each $e \in E$ a probability distribution over F.

To check that $\exp(-\epsilon)\Pr(A(M(D') \in S)) \leq \Pr(A(M(D)) \in S) \leq \exp(\epsilon)\Pr(A(M(D')) \in S)$, we decompose $\Pr(A(M(d)) \in S)$ as $\int_{x\in S}\Pr(M(d) = y)\Pr(A(y) = x)$ and we use that M satisfies ϵ-iDP. □

The privacy axiom of choice allows randomly choosing between ϵ-iDP mechanisms.

Definition 8 (Privacy axiom of choice)**.** Let M_1 and M_2 be privacy mechanisms that satisfy a certain privacy definition. For any $p \in [0, 1]$, let M_p be a randomized algorithm that, on input i, outputs $M_1(i)$ with probability p (independent of the data and the randomness in M_1 and M_2) and $M_2(i)$ with probability $1-p$. Then M_p is a privacy mechanism that satisfies the privacy definition.

Proposition 4. *ϵ-iDP satisfies the privacy axiom of choice.*

Proof. Let M_1 and M_2 be ϵ-iDP mechanisms, and let $p \in [0, 1]$. Let M_p be a mechanism that, on input i, outputs $M_1(i)$ with probability p and $M_2(i)$ with probability $1-p$.

We want to check that M_p is ϵ-iDP. That is, for any D' neighbor of D, we have $\exp(-\epsilon)\Pr(M_p(D') \in S) \leq \Pr(M_p(D) \in S) \leq \exp(\epsilon)\Pr(M_p(D') \in S)$. This is easily seen by expressing $\Pr(M_p(D) \in S)$ as $p \times \Pr(M_1(D) \in S) + (1-p) \times \Pr(M_2(D) \in S)$ and using that both M_1 and M_2 satisfy ϵ-iDP. □

V. ϵ-iDP for Numerical Queries

In DP, response mechanisms for numerical queries are often expressed in terms of noise addition (see Section II). In this line, given a numerical query f, this section presents an ϵ-iDP mechanism of the form $\kappa(x) = f(x) + N$, where N is a random noise.

For $\kappa(x) = f(x) + N$ to be ϵ-iDP, we have to make sure that the result of the mechanism is indistinguishable between D and its neighbor data sets. As we show later, this can be done by adjusting the noise to the *local sensitivity* $LS_f(D)$ (see Definition 3). The ability to calibrate the noise to the local sensitivity rather than to the global or smooth sensitivities is a significant advantage to preserve data utility: first, many functions have small local sensitivity but large global sensitivity, as noted right after Definition 3 above; second, calibration to the local sensitivity also improves the accuracy of query responses with respect to calibration to the smooth sensitivity, not only because the local sensitivity is smaller than the smooth sensitivity, but also because calibration to the local sensitivity allows using exponentially decreasing noise distributions (rather than the heavy-tailed distributions required by calibration to the smooth sensitivity). Last but not least, calibration to the local sensitivity is simpler than calibration to the smooth sensitivity.

We next detail how to enforce iDP for both continuous and discrete numerical results.

A. Laplace mechanism

The $Laplace(\mu, b)$ distribution (a.k.a the double exponential distribution) is an absolutely continuous distribution whose density function is

$$L_{\mu,b}(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right),$$

where μ is the location parameter, and b is the scale parameter. We use $L_b(x)$ when the mean parameter is zero.

In case of using the Laplace distribution to generate the random noise, iDP can be rewritten as follows.

Proposition 5. *Let f be a query function that takes values in $\mathbb{R}^k$. The mechanism $\kappa(x) = f(x) + (N_1, \ldots, N_k)$, where N_i are independent identically distributed $Laplace(0, LS_f(D)/\epsilon)$ random noises, gives ϵ-iDP.*

Proof. If N is a vector of independent $Laplace(0, \beta)$ distributed variables, we know that for all $s \in \mathbb{R}^k$

$$\exp\left(-\frac{\|z-z'\|_1}{\beta}\right) \leq \frac{\Pr(z+N=s)}{\Pr(z'+N=s)} \leq \exp\left(\frac{\|z-z'\|_1}{\beta}\right).$$

Let D' be a data set that differs from D in one record. By taking $z = f(D)$, $z' = f(D')$, and $\beta = LS_f(D)/\epsilon$ we get

$$\exp(-\epsilon) \leq \frac{\Pr(f(D)+N=s)}{\Pr(f(D')+N=s)} \leq \exp(\epsilon).$$

□

B. Discrete Laplace mechanism

The previous mechanism (based on adding random noise with values in $\mathbb{R}$) is also capable of providing DP to query functions with values in $\mathbb{Z}$. However, for such query functions the use of a noise distribution with support over $\mathbb{Z}$ is a better option. The discrete version of the Laplace distribution is defined as follows.

Definition 9 (Discrete Laplace distribution [10])**.** A random variable N follows the discrete Laplace distribution with parameter $\alpha \in (0,1)$, denoted by $DL(\alpha)$, if for all $k \in \mathbb{Z}$,

$$\Pr(N = i) = \frac{1-\alpha}{1+\alpha}\alpha^{|i|}.$$

The discrete Laplace distribution can also be used to attain ϵ-iDP. To this end, the parameter α must be adjusted to the desired privacy level ϵ and to the local sensitivity $LS_f(D)$ of the query.

Theorem 5 (The discrete Laplace mechanism)**.** *Let f be a function with values in $\mathbb{Z}^k$. The discrete mechanism $\kappa_D(x) = f(x) + N$, where $N = (N_1, \ldots, N_k)$ and $N_i \sim DL(\exp(-\epsilon/LS_f(D)))$ are independent random variables, gives ϵ-iDP.*

Proof. If N is a vector of independent $DL(\alpha)$ distributed variables, we know that for all $s \in \mathbb{Z}^k$

$$\alpha^{\|z-z'\|_1} \leq \frac{\Pr(z+N=s)}{\Pr(z'+N=s)} \leq \alpha^{-\|z-z'\|_1}.$$

Let D' be a data set that differs from D in one record. By taking $z = f(D)$, $z' = f(D')$ and $\alpha = \exp(-\epsilon/LS_f(D))$, we get

$$\exp(-\epsilon) \leq \frac{\Pr(f(D)+N=s)}{\Pr(f(D')+N=s)} \leq \exp(\epsilon).$$

□

VI. Evaluation

In this section, we compare the accuracy obtained with the standard DP definition and with iDP for basic noise addition mechanisms. In particular, we compare calibration to the global sensitivity (for DP), to the smooth sensitivity (for DP), and to the local sensitivity (for iDP). Noise addition mechanisms provide simple ways to make an output compliant with DP (or iDP) and are agnostic to the specific data uses; hence, using standard noise addition mechanisms is relevant to make a general comparison between DP and iDP.

Since the accuracy we obtain from standard DP calibrated to the smooth sensitivity and from iDP calibrated to the local sensitivity depends on the actual data set, a purely empirical evaluation carried out on some sample data sets would only provide a partial picture. Therefore, we give both an empirical analysis based on simulations and a theoretical analysis. We will see that, even in the worst case, iDP offers a significant improvement on the accuracy of the responses.

A. Median

The median is a measure of central tendency. Compared to the arithmetic mean, the median is quite stable, because it is relatively insensitive to outliers. Since statistical queries like the median represent properties of the data set rather than properties of a specific individual, we could expect DP to provide accurate results for such queries. However, the fact that standard DP guarantees must hold for *any* pair of neighbor data sets forces noise addition mechanisms to significantly degrade accuracy. Instead, iDP allows adjusting the amount of protection to each specific data set; thus, most of the times, iDP yields reasonably high accuracy, as we discuss below.

1) Differential privacy via calibration to the global sensitivity: Let the values $x_1, \cdots, x_n$, with $n = 2m+1$, be instances of an attribute with domain $Dom(A)$. Consider data sets D_1 and D_2 such that: in D_1 values $x_1, \ldots, x_m$ are equal to the minimum of $Dom(A)$, and values $x_{m+1}, \ldots, x_n$ are equal to the maximum of $Dom(A)$; and in D_2 values $x_1, \ldots, x_{m-1}$ are equal to the minimum of $Dom(A)$, and values $x_m, \ldots, x_n$ are equal to the maximum of $Dom(A)$. As a result, the median x_m in D_1 is the minimum of $Dom(A)$, and the median x_m in D_2 is the maximum of $Dom(A)$; thus, the global sensitivity for the median is

$$\Delta(\mathit{median}) = \max(\mathit{Dom}(A)) - \min(\mathit{Dom}(A)).$$

Having a global sensitivity equal to the domain size severely compromises the accuracy of differentially private estimations of the median via calibration to the global sensitivity. The situation becomes even worse when the domain of the attribute is not naturally bounded (e.g. incomes are unbounded, even if the income of most individuals falls within a given window); in such cases, it may not even be possible to compute the global sensitivity without artificially restricting the values to a fixed interval.

2) Differential privacy via calibration to the smooth sensitivity: Calibration to the smooth sensitivity tries to avoid the shortcomings of global sensitivity by adjusting the amount of noise to each data set. From [11], the smooth sensitivity for the median is

$$S_{median,\epsilon}(D) = \\ = \max_{k=0,\ldots,n}\{\exp(-k\epsilon) \max_{0\leq t\leq k+1}\{x_{m+t} - x_{m+t-k-1}\}\}.$$

Calibration to the smooth sensitivity is effective at reducing the amount of noise that is added to most data sets. However, it still has some drawbacks:

- Computing the smooth sensitivity can be complex. In particular, the formula for the smooth sensitivity of the median has time complexity $\mathcal{O}(n^2)$, which may not be feasible for large data sets.
- Similarly to the global sensitivity, the smooth sensitivity can only be computed when $Dom(A)$ is bounded.
- The mechanisms used to attain DP are more complex than with global sensitivity, and the eligible noise distributions are heavy-tailed (rather than the exponentially decreasing distributions used in calibration to the global sensitivity).

3) iDP (via calibration to the local sensitivity): iDP also allows the response mechanism to be independently adjusted to each data set. iDP can be reached via calibration to the local sensitivity (see Section V). For the median, the local sensitivity is:

$$LS_{median}(D) = \max\{x_m - x_{m-1}, x_{m+1} - x_m\}.$$

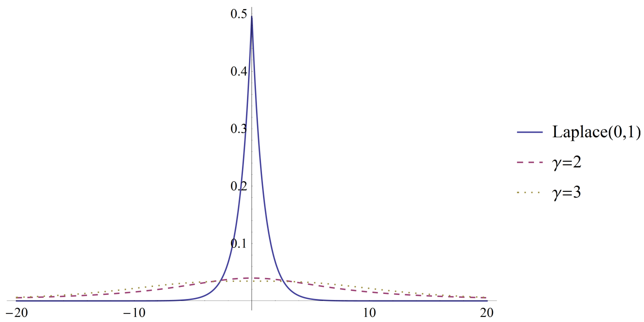


Figure 1. Comparison of the noise distributions required by ϵ-DP via calibration to the smooth sensitivity (for $\gamma = 2$ and $\gamma = 3$, where γ is the noise parameter) and ϵ-iDP via calibration to the local sensitivity (with a Laplace distribution), where the query is the median, $\epsilon = 1$ and sensitivities are 1.

iDP via calibration to the local sensitivity offers the following advantages compared to calibration to the smooth sensitivity:

- The computation of the local sensitivity is less complex.
- The local sensitivity is lower than the smooth sensitivity.
- The local sensitivity does not depend on the size of the domain. It only depends on x_{m-1}, x_m and x_{m+1}.
- The mechanism used to attain iDP is simpler and the noise distributions can be selected to be exponentially decreasing (e.g. the Laplace distribution).

4) Experimental comparison between DP with smooth sensitivity and iDP: After the previous theoretical comparison, we now turn to experimental evaluation of the accuracy for the median. Due to the poor accuracy provided by calibration to the global sensitivity, the comparison focuses on DP via calibration to the smooth sensitivity and iDP via calibration to the local sensitivity.

For iDP, we have used the mechanism proposed in Proposition 5. For standard DP, we have used a noise N having a density of the form

$$f_N(x) \propto \frac{1}{1 + |x|^\gamma},$$

where $\gamma > 1$. This noise distribution is $(\frac{\epsilon}{4\gamma}, \frac{\epsilon}{\gamma})$-admissible to achieve DP calibrated to the smooth sensitivity [11].

In the comparison, we take $\epsilon = 1$ and we assume that the local sensitivity equals the smooth sensitivity. Since i) noise in DP is proportional to the smooth sensitivity, ii) noise in iDP is proportional to the local sensitivity, and iii) the latter sensitivity is never greater than the former, both sensitivities being equal is the worst case for iDP. Specifically, we take both sensitivities equal to 1. Figure 1 shows the distribution of the noise in the response. Even in the most favorable case for the smooth sensitivity, the accuracy obtained with iDP is much better. The reason is that, to enforce ϵ-DP via calibration to the smooth sensitivity, the noise distribution must be heavy-tailed. Table I shows 95%-confidence intervals for these noise distributions.

Table I
95% CONFIDENCE INTERVALS OF THE NOISE DISTRIBUTIONS REQUIRED BY ϵ-DP VIA CALIBRATION TO THE SMOOTH SENSITIVITY (FOR DIFFERENT VALUES OF THE NOISE PARAMETER γ) AND ϵ-iDP VIA CALIBRATION TO THE LOCAL SENSITIVITY (WITH A LAPLACE DISTRIBUTION), WHERE THE QUERY IS THE MEDIAN, $\epsilon = 1$ AND SENSITIVITIES ARE 1.

Privacy model	ϵ-iDP	ϵ-DP $\gamma = 2$	ϵ-DP $\gamma = 3$
95% interval	[-3,3]	[-101.7,101.7]	[-34.2,34.2]

These results show that the use of calibration to the smooth sensitivity may be incompatible with an accurate and ϵ-differentially private approximation for the median. Although the smooth sensitivity is definitely better than the global sensitivity for noise calibration, at 95% confidence the amount of noise can be as large as 101 times the smooth sensitivity. Whether this is acceptable depends on the data set. Only if the smooth sensitivity is very small compared to the attribute domain, accuracy may still be acceptable. In contrast, the accuracy provided by iDP is much better: at 95% confidence, the noise is at most 3 times the local sensitivity.

Finally, we evaluate the accuracy of the median by computing, for a collection of data sets, the average absolute error incurred; we use the absolute error because the relative error may be misleading when the median is close to 0. We consider data sets with three different sizes (10, 100, and 1000) whose records are drawn from $U[0, 1]$ (the uniform distribution in $[0, 1]$), $N(0, 1)$ (the normal distribution with mean 0 and variance 1) and $Exp(1)$ (the exponential distribution with rate 1).

For the smooth sensitivity to be finite, the domain of the records must be bounded. Thus, to be able to compute the smooth sensitivity in the $N(0, 1)$ and the $Exp(1)$ cases, we have to restrict the domain of the data. This is done by bounding the domain to the range between the minimum and the maximum values in the data set. Figure 2 shows the average absolute error for each of the data sets. We include the results obtained for DP by adjusting the noise to the smooth sensitivity with $\gamma = 3$, and for iDP by adjusting a Laplace noise to the local sensitivity. We took $\gamma = 3$ as the best performing value after evaluating several other values of γ; specifically, smaller values led to too much probability mass in the tails, whereas greater values introduced greater correction factors to the noise distribution that offset the gain produced by having the probability mass closer to zero. By looking at the figure, we notice that, regardless of the underlying data distribution and the data set size, the results are significantly better for iDP.

B. Maximum

Rather than comparing the accuracy obtained with different mechanisms in a baseline scenario, this example focuses on the maximum value of a data set and shows how iDP is useful to avoid calibrating the noise to a domain-dependent (and hence too large) sensitivity.

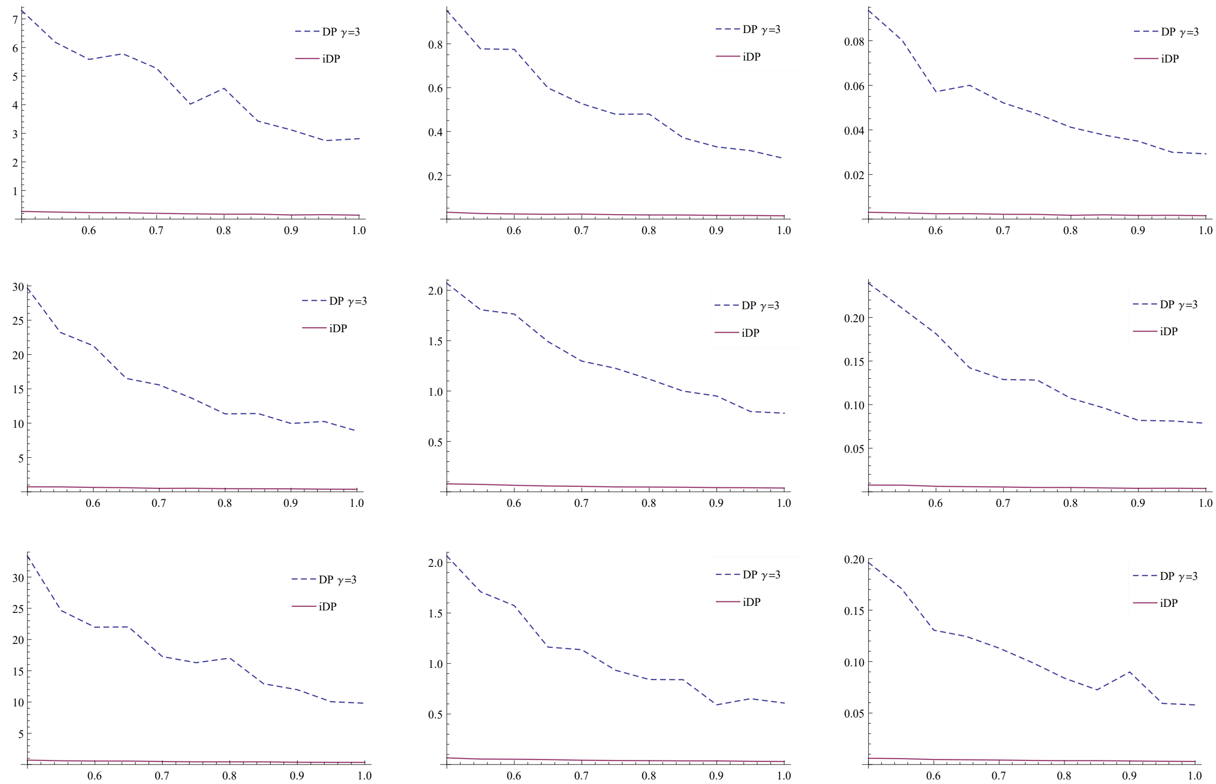


Figure 2. Absolute error in the median (y-axis) for ϵ in the range $[0.5, 1]$ (x-axis) and several data sets. Data sets have been drawn from a specific distribution: $U[0,1]$ (top row graphs), $N(0,1)$ (middle row graphs) and $Exp(1)$ (bottom row graphs). Data set sizes are 10 (left column graphs), 100 (middle column graphs), and 1000 (right column graphs). Especially in the left column graphs, it can be seen that strict DP yields completely off-range median values for all considered ϵ, as the median of a $U[0,1]$ must be within $[0,1]$, and, with very great probability, the median of a $N(0,1)$ is within $[-3,3]$ and the median of an $Exp(1)$ is within $[0,5]$. Clearly, iDP is much more utility-preserving.

1) Differential privacy via calibration to the global sensitivity: The global sensitivity of the maximum of a set of values equals the length of the domain. Let us consider data sets D_1 and D_2 such that: in D_1, all the values $x_1, \ldots, x_n$ are equal to the minimum of $Dom(A)$, and in D_2 the values $x_1, \ldots, x_{n-1}$ are equal to the minimum of $Dom(A)$, and the value x_n is equal to the maximum of $Dom(A)$. Thus, the global sensitivity for the maximum is

$$\Delta(maximum) = \max(Dom(A)) - \min(Dom(A)).$$

Like in the case of the median, having a global sensitivity equal to the domain size severely degrades the accuracy of differentially private estimations of the maximum via calibration to the global sensitivity. Also, if the domain is unbounded, the global sensitivity cannot even be computed.

2) Differential privacy via calibration to the smooth sensitivity: The smooth sensitivity for the maximum can be computed as

$$\begin{aligned}&S_{maximum,\epsilon}(D) = \\ &\max_{k=0,\ldots,n}\{\exp(-k\epsilon)(\max(Dom(A)) - x_{n-k}), \\ &\exp(-k\epsilon)(x_n - x_{n-k-1})\}.\end{aligned}$$

Calibration to the smooth sensitivity is effective to reduce the amount of noise that is added to most data sets. In fact, except for the worst-case data sets described when computing the global sensitivity, the smooth sensitivity is smaller. However, the smooth sensitivity does not avoid the dependence on the domain of the attribute.

3) iDP (via calibration to the local sensitivity): The local sensitivity of the maximum can be computed as

$$LS_{maximum}(D) = \max\{\max(Dom(A)) - x_n, x_n - x_{n-1}\}.$$

While iDP offers the expected advantages (the local sensitivity is smaller than the smooth sensitivity, and we can use exponentially decreasing noise distributions), the local sensitivity still depends on the domain of the attribute, which can be unbounded. However, iDP allows an easy workaround to solve this issue: rather than querying for the maximum value, we can query for the second maximum. Unless there is a single record that has a significantly greater value (as in the worst-case data set described above), querying for the second maximum should be a reasonably good approximation; otherwise, the approximation will be very inexact, but, anyway, the accuracy when querying for the maximum would also be poor. The advantage of querying for the second maximum is that the local sensitivity does not depend on the domain anymore:

$$LS_{2-\max}(D) = \max\{x_n - x_{n-1}, x_{n-1} - x_{n-2}\}.$$

It is important to note that this workaround to avoid having a sensitivity that depends on the domain attribute is not possible with calibration to the smooth sensitivity.

C. Range queries

Given a data set D and a range $\mathcal{R}$, a range query counts the number of records of D that are contained in $\mathcal{R}$:

$$f_{\mathcal{R}}(D) = |\{x \in D : x \in \mathcal{R}\}| .$$

Range queries behave well under DP: their sensitivity is 1. This low sensitivity is preserved in histogram queries (a collection of range queries over disjoint ranges).

The mechanisms to attain ϵ-iDP described in Sections V-A and V-B were based on calibrating the noise to the local sensitivity at D. As the local sensitivity of $f_{\mathcal{R}}$ at any D is 1, which is equal to the global sensitivity, the use of these mechanisms does not provide improved accuracy w.r.t. DP.

VII. Conclusions and future work

This work formalizes and discusses *individual differential privacy*, an alternative to the standard formulation of DP that reduces the noise to be added to the query results and, thus, better preserves their accuracy/utility. While, at first sight, individual differential privacy may look like another relaxation of DP, it exactly maintains the intuitive disclosure limitation guarantee of DP: the presence or absence of one individual in the data set must be unnoticeable from the query result. Improving the accuracy of the results is possible because individual differential privacy exploits the fact that the actual data set is known by the trusted data controller at the time of answering queries. By focusing only on indistinguishability between the actual data set and its neighbor data sets, the sensitivity of the query and, and hence, the magnitude of noise to be added significantly decrease.

We have also proposed several mechanisms to attain individual differential privacy. First, we have explained that any mechanism providing ϵ-DP also provides ϵ-individual differential privacy. However, direct use of ϵ-differentially private mechanisms fails to reap the potential accuracy improvements of individual differential privacy. Next, we have shown that, for numerical queries, ϵ-iDP can be attained by adjusting the noise to the local sensitivity, which results in substantial accuracy gains:

- Since the local sensitivity is normally significantly smaller than the global sensitivity employed by standard DP, iDP leads to substantially better accuracy.
- Even if noise is calibrated to the smooth sensitivity, iDP still offers much better accuracy and this for two reasons: (i) the local sensitivity is smaller than the smooth sensitivity (although the difference is not as important as with respect to the global sensitivity); (ii) exponentially decreasing random noise can be used with local sensitivity (in contrast with the heavy-tailed noises that are to be employed with calibration to the smooth sensitivity).

In addition to improved accuracy, iDP via local sensitivity is less dependent on the attribute domain (which may be large or even unbounded); what is more, in case the local sensitivity depends on the domain, workarounds can be found.

In our opinion, the significant accuracy/utility gains brought by ϵ-iDP, together with its strong privacy guarantee (the same intuitive privacy guarantee of ϵ-DP for individuals), pave the way to using ϵ-iDP where standard DP is not viable. As future work, we plan to study the performance of iDP for other common queries in the interactive scenario. We also plan to design mechanisms, other than noise addition, that may offer improved utility for specific tasks. Specifically, we aim at mechanisms for non-numerical discrete functions that leverage the advantages of iDP. Finally, we also plan to apply ϵ-iDP to non-interactive data releases. Although such releases provide more flexibility regarding data uses, they have been traditionally neglected by researchers because of the enormous distortion that making them differentially private according to the standard definition would entail [18].

Acknowledgments and disclaimer

We thank Frank McSherry and Daniel Kifer for their comments and suggestions on individual differential privacy. This work was supported by the European Commission (projects H2020-644024 “CLARUS” and H2020-700540 “CANVAS”), by the Spanish Government (projects TIN2014-57364-C2-1/2-R “SmartGlacis”, TIN2016-80250-R “Sec-MCloud” and TIN2015-70054-REDC) and by the Government of Catalonia under grant 2014 SGR 537. Josep Domingo-Ferrer is partially supported as an ICREA-Acadèmia researcher by the Government of Catalonia. The opinions expressed in this paper are the authors’ own and do not necessarily reflect the views of UNESCO. The first and second authors thank the Isaac Newton Institute for Mathematical Sciences at the University of Cambridge for support and hospitality during the “Data Linkage and Anonymisation” programme (funded by EPSRC Grant Number EP/K032208/1), when work on this paper was completed and discussed.

References

[1] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte-Nordholt, K. Spicer, and P. de Wolf, *Statistical disclosure control*. Wiley, 2012.

[2] J. Domingo-Ferrer, D. Sanchez, and J. Soria-Comas, *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-Based Inter-Model Connections*. Morgan & Claypool, 2016.

[3] C. Dwork, “Differential privacy,” in *Automata, Languages and Programming*, Springer, 2006, pp. 1–12.

[4] P. Samarati, “Protecting respondents’ identities in microdata release,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, Nov. 2001.

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “l-diversity: privacy beyond k-anonymity,” *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, Mar. 2007.

[6] N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: privacy beyond k-anonymity and l-diversity,” in *ICDE*, R. Chirkova, A. Dogac, M. T. Özsu, and T. K. Sellis, Eds. IEEE, 2007, pp. 106–115.

[7] J. Soria-Comas and J. Domingo-Ferrer, "Probabilistic k-anonymity through microaggregation and data swapping," in *FUZZ-IEEE 2012, IEEE International Conference on Fuzzy Systems*, Brisbane, Australia, IEEE, pp. 1-8.

[8] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006*, New York, NY, USA, March 4-7, 2006, pp. 265–284.

[9] J. Soria-Comas and J. Domingo-Ferrer, "Optimal data-independent noise for differential privacy." *Information Sciences*, vol. 250, pp. 200–214, 2013.

[10] S. Inusah and T. J. Kozubowski, "A discrete analogue of the Laplace distribution," *Journal of Statistical Planning and Inference*, vol. 136, no. 3, pp. 1090–1102, 2006.

[11] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, STOC '07*. New York, NY, USA: ACM, 2007, pp. 75–84.

[12] J. Bambauer, K. Muralidhar, and R. Sarathy, "Fool's gold: an illustrated critique of differential privacy," *Vanderbilt Journal of Entertainment & Technology Law*, vol. 16, no. 4, pp. 701–755, Summer 2014.

[13] R. Sarathy and K. Muralidhar, "Evaluating Laplace noise addition to satisfy differential privacy for numeric data," *Transactions on Data Privacy*, vol. 4, no. 1, pp. 1–17, Apr. 2011.

[14] A. Friedman and A. Schuster, "Data mining with differential privacy," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10*. New York, NY, USA: ACM, 2010, pp. 493–502.

[15] P. Mohan, A. Thakurta, E. Shi, D. Song, and D. Culler, "GUPT: privacy preserving data analysis made easy," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 349–360.

[16] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing," in *Proceedings of the 23rd USENIX Security Symposium*, San Diego CA, EUA, 2014, pp. 17–32.

[17] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Improving the utility of differentially private data releases via k-anonymity," in *12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, July 2013, pp. 372–379.

[18] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Enhancing data utility in differential privacy via microaggregation-based k-anonymity," *The VLDB Journal*, vol. 23, no. 5, pp. 771–794, Oct. 2014.

[19] D. Sánchez, J. Domingo-Ferrer, S. Martínez, and J. Soria-Comas, "Utility-preserving differentially private data releases via individual ranking microaggregation," *Information Fusion*, vol. 30, pp. 1 – 14, 2016.

[20] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: private data release via Bayesian networks," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD '14*. New York, NY, USA: ACM, 2014, pp. 1423–1434.

[21] J. Xu, Z. Zhang, X. Xiao, Y. Yang, and G. Yu, "Differentially private histogram publication," in *Proceedings of the 28th IEEE International Conference on Data Engineering, ICDE '12*. Washington, DC: IEEE Computer Society, 2012, pp. 32–43.

[22] Y. Xiao, L. Xiong, and C. Yuan, "Differentially private data release through multidimensional partitioning," in *Secure Data Management*, Springer, 2010, pp. 150–168.

[23] G. Jagannathan, K. Pillaipakkamnatt, and R. N. Wright, "A practical differentially private random decision tree classifier," *Transactions on Data Privacy*, vol. 5, no. 1, pp. 273–295, Apr. 2012.

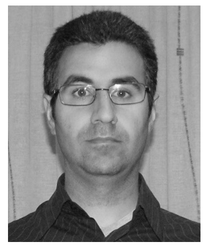
[24] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: privacy via distributed noise generation," in *Advances in Cryptology - EUROCRYPT 2006*, Springer, 2006, pp. 486–503.

[25] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: theory meets practice on the map," in *Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE 2008)*, 2008, pp. 277–286.

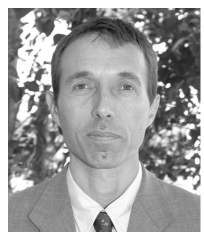
[26] A. Machanavajjhala and D. Kiefer, "Designing statistical privacy for your data," *Communications of the ACM*, vol. 58, no. 3, pp. 58–67, 2015.

[27] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," *CoRR*, vol. abs/1603.01887, 2016. [Online]. Available: http://arxiv.org/abs/1603.01887

[28] D. Kifer and B.-R. Lin, "Towards an axiomatization of statistical privacy and utility," in *Proceedings of the 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '10*. New York, NY, USA: ACM, 2010, pp. 147-158.



Jordi Soria-Comas is a postdoctoral researcher at Universitat Rovira i Virgili, Tarragona, Catalonia. He received an MSc in computer security (2011) and a PhD in computer science (2013) from Universitat Rovira i Virgili. He also holds an MSc in finance from the Autonomous University of Barcelona (2004) and a BSc in mathematics from the University of Barcelona (2003). His research interests are in data privacy and security.



Josep Domingo-Ferrer (Fellow, IEEE) is a distinguished professor of computer science and an ICREA-Acadèmia researcher at Universitat Rovira i Virgili, Tarragona, Catalonia, where he holds the UNESCO Chair in Data Privacy. He received the MSc and PhD degrees in computer science from the Autonomous University of Barcelona in 1988 and 1991, respectively. He also holds an MSc degree in mathematics. His research interests are in data privacy, data security and cryptographic protocols.

David Sánchez is an associate professor of computer science at Universitat Rovira i Virgili, Tarragona, Catalonia. He received a PhD in computer science (2008) from the Technical University of Catalonia. He also holds a BSc and an MSc in computer engineering from Universitat Rovira i Virgili. His research interests are in data semantics and data privacy.

David Megías (Member, IEEE) received a PhD in computer science from the Autonomous University of Barcelona (2000), where he was an assistant lecturer from 1994 to 2001. Since 2001, he is a tenured associate professor with Universitat Oberta de Catalunya (UOC). He currently serves as the director of the Internet Interdisciplinary Institute (IN3) of UOC. His current research interests include security and privacy in multimedia content distribution, steganography, steganalysis and information security and privacy.