How Much Is Enough? Choosing ϵ for Differential Privacy

Jaewoo Lee and Chris Clifton

Department of Computer Science, Purdue University,
West Lafayette, IN 47907
{jaewoo,clifton}@cs.purdue.edu
http://www.cs.purdue.edu

Abstract. Differential privacy is a recent notion, and while it is nice conceptually it has been difficult to apply in practice. The parameters of differential privacy have an intuitive theoretical interpretation, but the implications and impacts on the risk of disclosure in practice have not yet been studied, and choosing appropriate values for them is non-trivial. Although the privacy parameter ϵ in differential privacy is used to quantify the privacy risk posed by releasing statistics computed on sensitive data, ϵ is not an absolute measure of privacy but rather a relative measure. In effect, even for the same value of ϵ, the privacy guarantees enforced by differential privacy are different based on the domain of attribute in question and the query supported. We consider the probability of identifying any particular individual as being in the database, and demonstrate the challenge of setting the proper value of ϵ given the goal of protecting individuals in the database with some fixed probability.

Keywords: Differential Privacy, Privacy Parameter, ϵ.

1 Introduction

As volumes of personal data collected by many organizations increase, the problem of preserving privacy is increasingly important. The potential social benefits of analyzing such datasets drive many organizations to be interested in releasing statistical information about the data. In the field of privacy preserving data analysis, the main goal is to release statistical information about sample databases safely without compromising the privacy of any individuals who's records contribute to the database. These two conflicting objectives pose challenging trade-off between providing useful information about the population and protecting the privacy of any individuals.

Privacy laws typically protect *individually identifiable data*; data that cannot be linked to an individual is not considered a privacy risk. Unfortunately, what it means for data to be *individually identifiable* is not simple to define. Statistical summaries can reveal information about a single individual, particularly if an adversary knows information about other individuals. Differential Privacy [6] provides a strong guarantee of privacy even when the adversary has arbitrary

X. Lai, J. Zhou, and H. Li (Eds.): ISC 2011, LNCS 7001, pp. 325–340, 2011.

external knowledge. Basically, differential privacy hides the presence of an individual in the database from data users by making two output distributions, one with and the other without an individual, be computationally indistinguishable (for all individuals). To achieve this, differential privacy uses an output perturbation technique which adds random noise to the outputs. The magnitude of noise to add, which determines the degree of privacy, depends on the type of computation and it must be large enough to conceal the largest contribution that can be made to the output by one single individual. To be specific, let X be a database to release statistics about and f be a query function. ϵ-differentially private mechanism gives perturbed response $f(X)+Y$ instead of the true answer $f(X)$, where Y is the random noise.

While this seems a perfect solution, the amount of noise needed to achieve indistinguishability between two datasets generally eliminates any useful information. The actual definition is for ϵ-differential privacy (see Definition 1), where the ϵ factor is a difference between the probabilities of receiving the same outcome on two different databases. ϵ becomes a parameter on the degree of privacy provided. ϵ is a relative measure since it bounds the data user's information gain, instead of the absolute amount. Even for the same value of ϵ, the probability of identifying an individual enforced by differential privacy is different depending on the universe.

Unfortunately, ϵ does not easily relate to practically relevant measures of privacy. For example, assume a very simple problem where an adversary wants to determine the value of a binary attribute about an individual - simply "is the individual in the dataset" (such as a research dataset for diabetes, where simply revealing presence in the dataset places an individual at risk of discrimination.) What we would really like is a measure of the risk to an individual – what is the probability that an individual is in the dataset given release of statistical information about the data? If disclosure allows an adversary to calculate too high a probability that the individual is in the dataset, then that individual's privacy (in legal terms, which typically protect "individually identifiable data") is at risk. This is addressed for anonymization in [14], and the problem would seem a perfect match for differentially privacy. The challenge is that choosing an appropriate value of ϵ turns out to be quite challenging.

The problem is that protecting privacy requires knowing not only the data to be protected, but also the entire universe of individuals from which that data might be drawn. This is a known challenge with differential privacy, as calculating the sensitivity of a query is based on all *possible* databases differing by a single value. This may be an inherent problem with protecting privacy; δ-presence [14] faces the same issue (although an approximation based on univariate statistics is given in [13].) What makes this a particular problem for differential privacy is that not only do we need to know the entire universe to use a differentially private mechanism, it is also needed to determine an appropriate value of ϵ. In this paper, we will show that given a goal of controlling probabilistic disclosure of the presence of an individual, the proper of ϵ varies depending on individual values, even for individuals not in the dataset.

To see this, imagine the following (hypothetical) scenario. Purdue University has put together a "short list" of alumni as possible commencement speakers. A local newspaper is writing a story on the value Indiana taxpayers get from Purdue, and would like to know if these distinguished alumni are locals or world travelers. Purdue does not want to reveal the list (to avoid embarrassing those that are not selected, for example), but is willing to reveal the average distance people on the list have traveled from Purdue in their lifetimes. Using a differentially private mechanism to add noise to the resulting average will protect individuals - but how much noise is needed? Outliers in the data, such as Purdue's Apollo astronaut alumni (who have been nearly 400,000km from campus), result in a requirement for a significant amount of noise. More critically, we show that such outliers also change the appropriate value for ϵ. As a result, simply setting parameters to be used for differential privacy is an unsolved problem.

Although differential privacy has been extensively studied in many papers, to our best knowledge, no studies have been conducted toward the issues on the application of differential privacy in practice. In many papers, the value of privacy parameter ϵ is chosen arbitrarily or assumed to be given. This leaves an impression that ϵ can be freely chosen as needed but, in reality, decision on the value of ϵ should be made carefully with considerations of the domain and the acceptable ranges of risk of disclosure. In this paper, we illustrate why the choice of ϵ is important using the perspective of the risk of revealing presence and how an inappropriate value of ϵ can cause a privacy breach. We also show that a value of ϵ that is appropriate for a particular universe of values may lead to a breach with a different set of values.

2 Related Work

The concept of differential privacy was motivated by the impossibility of absolute protection [4] against adversaries with arbitrary external information [5]. In a differentially private mechanism, what a potential adversary can learn from interactions with the mechanism is limited (within a multiplicative factor) no matter what external information the adversary has. Essentially, what can be learned from a dataset with a particular individual also can be learned from a dataset without that individual [9,11]. This definition enables a privacy model that does not need to make assumptions on an adversary's external information, a key limitation of prior work on protecting privacy. A line of research on indistinguishability between two neighboring databases leads to emergence of differential privacy. [2,9,10]

The notion of differential privacy has received much theoretical attention in the privacy community and has been extensively studied in the literature [2,3, 10,8,1]; a recent survey on differential privacy is provided in [7]. However, most research on differential privacy has focused on exploring theoretical properties of the model. The main focus of study has been how to safely release database while preserving privacy for a particular function f. For example, [5] studies how to release count queries and [9] touches on more general query functions such

as histograms and linear algebraic functions. The concept of global sensitivity was introduced in [6] and it has been shown that releasing a database with noise proportional to the global sensitivity of the query functions achieves differential privacy. Nissim et al. [15] expanded the framework of differential privacy by introducing *smooth sensitivity*, which reduces the amount of noise added. It is motivated from the observation that, for many types of query functions, the local sensitivity is small while global (worst-case) sensitivity is extremely large. To decide the magnitude of noise, they use a smooth upper bound function S, which is an upper bound on local sensitivity.

There are a few implementations supporting differential privacy. PINQ [12] is an implementation of differential privacy that provides answers to SQL queries in a differentially private way. AIRAVAT [16] is another system that applies differential privacy mechanism for MapReduce computation in a cloud computing environment. Although their system has been built upon differential privacy framework, this doesn't mean that privacy is actually enforced by the system. It is still the responsibility of users who use the system to select the value of ϵ that prohibits any inferences on the dataset beyond what is allowed.

3 Differential Privacy

A database D is a collection of data elements drawn from the universe U. A row in a database corresponds to an individual whose privacy needs to be protected. Each data row consists of a set of attributes $A = A_1, A_2, ..., A_m$. The set of values each attribute can take, attribute domain, is denoted by $dom(A_i)$ where $1 \leq i \leq m$. A mechanism $\mathcal{M} : D \rightarrow \mathbb{R}^d$ is a randomized function that maps database D to a probability distribution over some range and returns a vector of randomly chosen real numbers within the range. A mechanism $\mathcal{M}$ is said to be ϵ-differentially private if adding or removing a single data item in a database only affects the probability of any outcome within a small multiplicative factor. The formal definition of an ϵ-differentially private mechanism is:

Definition 1 (ϵ-differentially private mechanism). *A randomized mechanism $\mathcal{M}$ is ϵ-differentially private if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{M})$*

$$Pr[\mathcal{M}(D_1) \in S] \leq \exp(\epsilon) \times Pr[\mathcal{M}(D_2) \in S]$$

This paper considers an interactive privacy mechanism and the same assumptions as in [6]. In an interactive model, users issue queries to the database and receive a noisy response where the magnitude of noise added to the response is determined based on the query function f. Sensitivity of a query function f represents the largest change in the output to the query function which can be made by a single data item.

Table 1. Example database X

Name	School year	Absence days
Chris	1	1
Kelly	2	2
Pat	3	3
Terry	4	10

Definition 2 (Global Sensitivity). *For the given query function $f : D \rightarrow \mathbb{R}^d$, the global sensitivity of f is*

$$\Delta f = \max_{x,y} |f(x) - f(y)|$$

for $\forall x, y$ differing in at most one element.

Let $Lap(\lambda)$ be the Laplace distribution whose density function is $h(x) = \frac{1}{2\lambda} \exp(-\frac{|X-\mu|}{\lambda})$ where μ is a mean and $\lambda(> 0)$ is a scale factor. Dwork et al. proved that, for the given query function f and a database X, a randomized mechanism $\mathcal{M}_f$ that returns $f(X)+Y$ as an answer where Y is drawn i.i.d from $Lap(\frac{\Delta f}{\epsilon})$, gives ϵ-differential privacy [9].

4 Example: Mean

In this section, we illustrate how the value of ϵ should be adjusted according to the change of domain (or universe) of the attribute in question to enforce the same level of protection. While essentially the same as the problem described in Section 1, we switch to a different motivation both to show the generality of this problem and to give realistic numbers that are easy to demonstrate.

Consider a database consisting of 4 students registered for a course, which includes each student's name, school year, and number of absence days in the previous semester. Let X denotes the database. In X, there is one student, Terry, who was placed on academic probation in the previous semester. Table 1 shows the example database. Note that Terry's number of absence days is relatively large compared to those of other students. Assume that the school wants to release data on students who have not been on academic probation to support academic success research. Let X' denote the database to be published, i.e., $X' = X - \{Terry\}$. Since knowing if a student has been on academic probation is clearly a privacy breach, the school allows faculty and staff (who may know the year in school and absence days of individual students, but not who has been placed on probation) to query X' only via an ϵ-differentially private mechanism. For the purpose of illustration, let us assume that the goal of the data provider is to hide the presence of data contributors (or, in this case, non-contributors) by keeping the adversary's probability of identifying their presence in the database less than 1/3. Throughout the example, we show that what values of ϵ needed to achieve that goal for queries on mean year and mean absence days, and

demonstrate that in spite of the similarity between the data columns (in fact identical for the data in X'), a different value of ϵ is needed depending on which column is queried.

4.1 Achieving Differential Privacy

We now describe an ϵ-differentially private mechanism κ for releasing the mean school year and absence days of students in X'. Informally, the goal of κ is to make the query responses from any databases that differ in only one element be indistinguishable within a factor of e^{ϵ}, so that the absence of Terry in X' (and thus probation status) isn't revealed. At the same time, the privacy of individuals who participated in the database X' needs to be protected as well. The sensitivity of the mean query function Δf is computed by measuring the maximum change in the query output caused by a single individual. Notice that calculating the sensitivity requires global knowledge on that domain since every possible attribute value that not only presently exists in X' but also could exist needs to be considered. For any possible data instance Y of size 3 and a tuple t of Y, Δf is determined as the maximum value among the results of the following computation:

$$\Delta f(X') = \max_{Y \subset X} |f(Y) - f(Y - t)| \, where \, |Y| = 3$$

From our example, it is calculated as follows:

$$\Delta f = \left| \frac{1+2+10}{3} - \frac{1+2}{2} \right| = \frac{17}{6}$$

For now, let's assume ϵ=2; we later show how this choice of ϵ discloses the information. The random noise drawn from the Laplace distribution with mean 0 and scale factor $\lambda = \frac{\Delta f}{\epsilon}$ is 1.1677. The query response γ is produced as $\gamma = \kappa_f(X') = f(X') + Lap(\frac{\Delta f}{\epsilon}) = 2 + 1.1677 = 3.1677$. Consider the following probabilities:

$$\frac{Pr[\kappa_f(X) > 3.1677]}{Pr[\kappa_f(X') > 3.1677]} = 3.2933 \leq e^2$$

Note that the above value computed from the cumulative density function of the Laplace distribution. The adversary cannot distinguish the response from queries against X and X' within the factor of e^2, so differential privacy is achieved. However, does this also mean that Terry's privacy is protected?

4.2 Adversary Model

In this example (as is the goal of differential privacy), we assume a very strong adversary who has complete knowledge on the universe, i.e., full access to all records in the universe U; thus each attribute value of all records in X is known to the adversary. In other words, the adversary can potentially access the records of every student in this school. The adversary knows everything about the universe except that which individual is missing in the database X' (i.e., who is on academic

probation). In addition to the complete knowledge about X, the adversary knows the fact that X' consists of students who have never been on academic probation. Assuming an adversary having complete knowledge about each individual in the database is not unrealistic because differential privacy is supposed to provide privacy given adversaries with arbitrary background knowledge.

In our model, the adversary has a database X consisting of n records, i.e., knowledge of the exact attribute values of each individual in X, and has an infinite computational power. Given a database X' with $n-1$ records sampled from X (i.e., $X' \subset X$ and $|X'| = |X| - 1$), the adversary's goal is to figure out absence of a victim individual in X' by using knowledge of X. This is identical to find out other individuals' presences in X'. With respect to our example, a privacy breach is to allow the adversary to guess absence/presence of an individual in X' correctly with high probability.

4.3 Attack Model

To determine membership in X', the adversary maintains a set of tuples $\langle \omega, \alpha, \beta \rangle$ for each possible combination ω of X', where α and β are the adversary's prior belief and posterior belief on $X' = \omega$ given a query response. Let Ψ denote the set of all possible combinations of X'. For simplicity, we assume α is a uniform prior, i.e., $\forall \omega \in \Psi, \alpha(\omega) = \binom{n}{n-1} = \frac{1}{n}$. We refer to each possible combination ω in Ψ as a possible world. The posterior belief β is defined in Definition 3.

Definition 3 (Posterior belief on $X' = \omega$). *Given the query function f and the query response $\gamma = \kappa_f(X')$, for each possible world ω, the adversary's posterior belief on ω is defined as:*

$$\beta(\omega) = P(X' = \omega | \gamma) = \frac{P(\kappa_f(\omega) = \gamma)}{P(\gamma)} = \frac{P(\kappa_f(\omega) = \gamma)}{\sum_{\psi \in \Psi} P(\kappa_f(\psi) = \gamma)}$$

where κ_f is an ϵ-differentially private mechanism for the query function f.

The posterior belief $\beta(\omega)$ represents the adversary's changed belief on each possible world that the underlying database being queried against is ω. To figure out which individuals are in the database, the adversary issues a query against X' and gets a noisy answer. After seeing the query response, the adversary computes the posterior belief for each possible world. Finally, the adversary selects one with the highest posterior belief as a "best guess". The confidence of the adversary's guess is calculated using Definition 4.

Definition 4 (Confidence level). *Given the best guess ω', the adversary's confidence in guessing the missing element is defined as*

$$conf(\omega') = \beta(\omega') - \alpha(\omega')$$

As the adversary's posterior belief on each possible world becomes large, the chances of disclosing any individual's presence in the database also become high, which makes disclosure of the statistics. This has an implication that the adversary's posterior belief on each possible world can be thought of as the risk of disclosure.

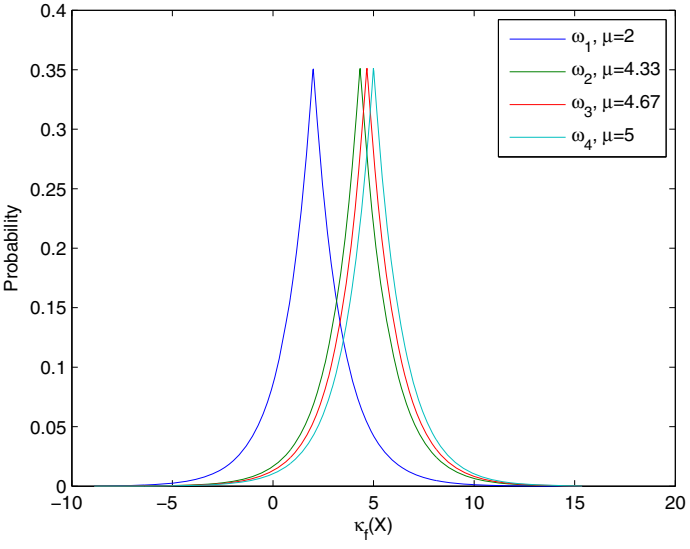


Fig. 1. Query response distributions

Definition 5 (Risk of disclosure Γ). *Given the set of possible worlds Ψ, the risk of disclosing presence/absence of any individual in the database Γ is defined as the adversary's maximum posterior belief.*

$$\Gamma = \max_{\omega \in \Psi} \beta(\omega)$$

4.4 Limitation of Differential Privacy

Basically, the underlying assumption that differential privacy is relying on is that, if two extreme query answers that can be produced from any dataset possible in the universe are indistinguishable, the presence or absence of any individual can be hidden. The difference between those two extreme answers is masked by random noise. However, there is a problem with this approach. Although differential privacy ensures that every possible database of the same size is indistinguishable within some factor ϵ, there always exists a distribution that is more likely than others given the query response. For example, in Figure 1, ω_1 is the most likely to be the true distribution among 4 possible worlds given the response γ=1. This allows the adversary to improve the belief of each possible world after seeing the response.

In our example, for illustration we assume $U = X$ and, without loss of generality, the response is 2.2013. Recall that the sensitivity Δf of mean query function for the domain of absence days is $\frac{17}{6}$. The adversary's posterior belief of ω_1 when $\epsilon = 2$ is:

Table 2. Posterior belief $\beta(\omega)$

Possible world(ω)	$\epsilon = 5$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.01$
{1, 2, 3}	0.9705	0.5519	0.4596	0.3477	0.2328	0.2482
{1, 2, 10}	0.0159	0.1859	0.2019	0.2305	0.2527	0.2503
{1, 3, 10}	0.0087	0.1463	0.1791	0.2171	0.2558	0.2506
{2, 3, 10}	0.0049	0.1159	0.1594	0.2048	0.2588	0.2509

Table 3. Posterior belief $\beta(\omega)$

Possible world(ω)	$\epsilon = 5$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.01$
{1, 2, 3}	0.9705	0.6825	0.4596	0.3477	0.2680	0.2518
{1, 2, 10}	0.0158	0.1315	0.2017	0.2303	0.2469	0.2497
{1, 3, 10}	0.0088	0.1039	0.1793	0.2172	0.2440	0.2494
{2, 3, 10}	0.0049	0.0821	0.1594	0.2048	0.2411	0.2491

$$\beta(\omega_1) = \frac{P(\kappa_f(\omega) = 2.2013)}{\sum_{i=1}^{4} P(\kappa_f(\omega_i) = 2.2013)} = \frac{0.3602}{0.3062 + 0.0784 + 0.0619 + 0.0489} = 0.6180$$

The adversary will come to the conclusion that $X' = \omega_1$. Even though the output, mean of absence days, is released via a differentially private mechanism, the adversary can still make a correct guess on who is absent from the list with high probability and confidence. Consider the probability when the attribute queried is school year. The sensitivity of mean query function for the school year domain is $\frac{5}{6}$. The adversary's posterior belief on ω_1 for this case is 0.3390. Although the same parameter and response values are used for both cases, the resulting adversary's probabilities are significantly different. Notice that the adversary's posterior belief, $\beta(\omega)$, is a random variable and Table 2 and Table 3 show two different instances of it. As shown in Table 2, an adversary's best guess would be $X' = \{\text{Chris}, \text{Kelly}, \text{Pat}\}$, which is correct, with the confidence of 0.7705=0.9705-0.25 when ϵ=5. When ϵ=0.5, the adversary still get it right but the confidence is only 0.0977(0.3477-0.25). When ϵ=0.01, the adversary fails to make the correct guess given the set of query responses.

5 Choice of ϵ

The previous section demonstrated that given sufficiently low ϵ, ϵ-differential privacy does limit an adversary's ability to identify an individual. However, as lowering ϵ reduces the utility of the answer, the question of the proper value of ϵ is still open. We now demonstrate how to choose ϵ to control the adversary's success at identification of an individual in this particular scenario, and demonstrate the difficulties that arise.

5.1 Upper Bound on Adversary's Posterior Belief

Let $X = \{x_1, x_2, \ldots, x_{n-1}, v\}$ and $X' = \{x_1, x_2, \ldots, x_{n-1}\}$. Without loss of generality, assume that elements are sorted in ascending order (i.e., $x_1 < x_2 < \ldots < x_{n-1} < v$). Two databases X and X' are identical except that only one element, v, is missing in X'. For illustration, we impose an ordering to the enumeration of possible worlds. Let ω_i denote the i^{th} possible world maintained by an adversary and x_i^k be the k^{th} smallest element of ω_i. For any $\omega_i, \omega_j \in \Psi, \omega_i$ is lower than ω_i if $\forall k s.t. 1 \leq k \leq n, x_i^k \leq x_j^k$. For example, $\omega_i = \{x_1, x_2, \ldots, x_{n-1}\}, \omega_2 = \{x_1, x_2, \ldots, x_{n-2}, v\}, \omega_3 = \{x_1, x_2, \ldots, x_{n-3}, x_{n-1}, v\}$, etc. To get an upper bound on the adversary's probability of a correct guess, we have to assume the worst case in which the correct answer seems to be most likely (i.e., $\gamma = f(\omega_i)$).[1] Given the query response γ, the adversary's posterior probability on $X' = \omega_i$ is $\beta(\omega_i) = \frac{P(\kappa_f(\omega_i)=\gamma)}{\sum_{k=1}^{n} P(\kappa_f(\omega_k)=\gamma)}$. If we divide numerator and denominator of $\beta(\omega_i)$ by $P(\kappa_f(\omega_i) = \gamma)$,

$$\beta(\omega_i) = \frac{1}{1 + \sum_{k=1,k\neq i}^{n} \frac{P(\kappa_f(\omega_k)=\gamma)}{P(\kappa_f(\omega_i)=\gamma)}} \tag{1}$$

$$= \frac{1}{1 + \sum_{k=1,k\neq i}^{n} \frac{e^{-\frac{|\gamma - f(\omega_k)|}{\lambda}}}{e^{-\frac{|\gamma - f(\omega_i)|}{\lambda}}}} \tag{2}$$

$$\leq \frac{1}{1 + \sum_{k=1,k\neq i}^{n} e^{-\frac{|f(\omega_i) - f(\omega_k)|}{\lambda}}} \tag{3}$$

$$\leq \frac{1}{1 + \sum_{k=1,k\neq i}^{n} e^{-\frac{\Delta v}{\lambda}}} \tag{4}$$

$$= \frac{1}{1 + (n-1)e^{-\frac{\epsilon \Delta v}{\Delta f}}} \tag{5}$$

where $\Delta v = \max_{1 \leqslant i,j \leqslant n} |f(\omega_i) - f(\omega_j)|$ and $i \neq j$.

In (4), the distance between $f(\omega_i)$ and $f(\omega_k)$ is approximated with Δv, the longest distance between $f(\omega_i)$ and $f(\omega_j)$ where $1 \leq i, j \leq n$. Recall that in our example, the missing value is the largest in the database, which means ω_1 is a true distribution. Under this condition, the distance between $f(\omega_i)(1 \leq i \leq n)$ and $f(\omega_1)$ has little difference, which makes (3) and (4) approximately the same. Therefore, the upper bound becomes tight. In order to make every possible world look equally likely, the following equation needs to be satisfied:

$$\frac{1}{1 + (n-1)e^{-\frac{\epsilon \Delta v}{\Delta f}}} = \frac{1}{n}$$

[1] This "all possible worlds" knowledge is the same information needed to calculate global sensitivity for differential privacy; extending differential privacy to work with more limited information is beyond the scope of this paper, and would face similar challenges to those addressed for generalization-based anonymization in moving from [14] to [13].

The value of ϵ which satisfies the equation is 0. Therefore, to make every possible dataset in the universe to look equally likely, the query results would be pure random noise, providing no utility.

5.2 Determining the Right Value of ϵ

We now show how the proper value of ϵ can be chosen given the goal of hiding any individual's presence (or absence) in the database. Assume that the privacy requirement for our example dataset is to limit the any individual's probability of being identified as present in the database to be no greater than $\frac{1}{3}$. We show two ways of selecting a good choice of ϵ that guarantees the probability of identifying any individual's presence is no greater than the maximum tolerable value δ. One is to use the upper bound presented in Section 5.1; the other is to search for the right value. We first consider how the upper bound on the adversary's probability can be utilized to enforce the requirement. Let ρ be the probability of being identified as present in the database. We need to find ϵ that satisfies the following inequality.

$$\frac{1}{1+(n-1)e^{-\frac{\epsilon \Delta v}{\Delta f}}} \leq \rho \tag{6}$$

Rearranging yields

$$\epsilon \leq \frac{\Delta f}{\Delta v} \ln \frac{(n-1)\rho}{1-\rho} \tag{7}$$

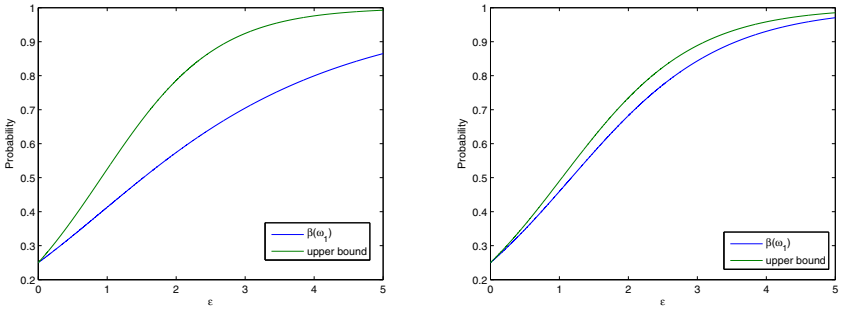
Note that the greater n and ρ are, the greater the minimum ϵ needed. Therefore, as the size of database to publish and the probability to bound get larger, less noise need be added.

In our example, the maximum distance between function values of every possible world, Δv, is $f(\omega_4) - f(\omega_1) = 5 - 2 = 3$. Thus, in order to enforce the adversary's probability to be no greater $\frac{1}{3}$,

$$\frac{1}{1+(n-1)e^{-\frac{\epsilon \Delta v}{\Delta f}}} \leq \frac{1}{3} \tag{8}$$

$$\epsilon \leq \frac{17}{18} \ln(\frac{3}{2}) \approx 0.3829 \tag{9}$$

Let's consider how this value changes when the attribute to release is the school year rather than absence days. In this case, the sensitivity Δf and the maximum distance between possible answers Δv need to be recalculated since those are the parameter values that are completely dependent on the universe. The recomputed values of Δf and Δv are $\frac{5}{6}$ and 1, respectively. Note that the mean value of school year information is less sensitive than that of absence days, which results in smaller values of both Δf and Δv. The right value of ϵ for the school year domain using the upper bound is $\epsilon \leq \frac{5}{6} \ln \frac{3}{2} \approx 0.3379$. However, even when $\epsilon = 0.5 (> 0.3379)$, the risk of disclosure is



(a) $dom(X') = \{1, 2, 3, 4\}$ (b) $dom(X') = \{1, 2, 3, 10\}$

Fig. 2. Upper bound on the risk of disclosure by varying domains

$$\max_i \beta(\omega_i) = \beta(\omega_1) \tag{10}$$

$$= \frac{1}{1 + e^{-\frac{1}{5}} + e^{-\frac{2}{5}} + e^{-\frac{3}{5}}} \tag{11}$$

$$\approx 0.3292 < \frac{1}{3}, \tag{12}$$

which is still lower than the maximum acceptable level of risk. This means that a more precise value of ϵ can be found. As shown in Figure 2, our upper bound on the risk of disclosure is not tight when the domain does not include outliers. In other words, when the domain of the attribute to be released has low sensitivity, the upper bound of Section 5.1 gives a value of ϵ that may be significantly greater than the actual value of ϵ needed to satisfy the privacy constraint (the actual impact on the amount of noise added follows from the differential privacy literature.) Although the value obtained using the upper bound ensures that any individual's risk of being identified as present in the database is no greater than δ, this might be overkill, especially when there is no value that significantly deviates from the mean of that distribution. In such case, we can perform binary search to determine the maximum ϵ that meets the requirement. Before performing the search, the range within which the value of ϵ will be searched needs to be determined. The minimum would be 0 which means no information can be learned while the maximum can be calculated using the upper bound above. Let ϵ_s and ϵ_f denote the beginning and end of the range to search, respectively. Firstly, compute the risk of disclosure when $\epsilon = \frac{\epsilon_s + \epsilon_f}{2}$. Next, if it is greater than δ, set ϵ_f to the current value of ϵ. Otherwise, set ϵ_s to ϵ. Repeat this search process until the maximum value of ϵ that satisfies the constraint is found.

6 Example: Median

We now show that the appropriate value of ϵ is dependent not only on the data and universe of values, but also on the query to be computed. This section

Table 4. Senstivity of median for the database X

Possible world(ω)	$LS_f(\omega)$
{1, 2, 3}	0.5
{1, 2, 10}	4
{1, 3, 10}	3.5
{2, 3, 10}	3.5

repeats the previous analysis on the example student database shown in Table 1, but where the query is instead *median*.

Let $f(X) = median(x_1, x_2, ..., x_n)$ where x_i are real numbers. The median of a finite list of numbers is defined to be the middle one when all the observations are arranged from lowest value to highest value. If a dataset has an even number of observations, there may be no single middle value; in this case we define the median to be the mean of the two middle numbers. Without loss of generality, assume $x_1 \leq \cdots \leq x_n$; this gives the following definition for median:

$$f(X) = \begin{cases} x_k & \text{for } n = 2k-1 \quad (13) \\ \dfrac{x_k + x_{k+1}}{2} & \text{for } n = 2k \quad (14) \end{cases}$$

where k is a positive integer. To calculate sensitivity, let X' be a database obtained by removing one element from X. If X has an odd number of elements (i.e., $n = 2k-1$), $f(X) = x_k$ and $f(X')$ could be $\frac{x_k+x_{k+1}}{2}$, $\frac{x_{k-1}+x_k}{2}$ or $\frac{x_{k-1}+x_{k+1}}{2}$. On the other hand, if X has an even number of elements (i.e., $n = 2k$), $f(X) = \frac{x_k+x_{k+1}}{2}$ and $f(X')$ is either x_k or x_{k+1}. Thus, the local sensitivity of median for X is

$$LS_f(X) = \begin{cases} \max\left(\frac{x_{k+1}-x_k}{2}, \frac{x_k-x_{k-1}}{2}, \left|\frac{x_{k+1}+x_{k-1}}{2} - x_k\right|\right) & \text{for } n = 2k-1 \\ \frac{x_{k+1}-x_k}{2} & \text{for } n = 2k \end{cases} \quad (15)$$

and the global sensitivity of median is

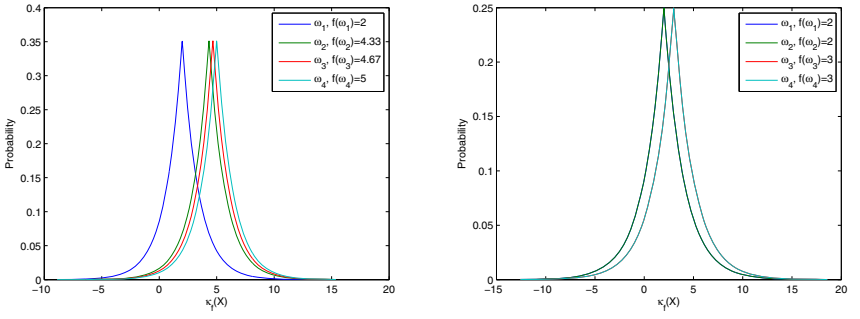
$$\Delta f(X) = \max_{\omega \in \Psi} LS_f(\omega)$$

Table 4 shows the sensitivity of median for the absence days attribute of our example database. As with Section 5.2, our target is that the adversary's probability estimate for the value of this attribute should be no greater than $\frac{1}{3}$. In our example, the maximum difference between medians of each possible world is 1 (i.e., $\Delta v = \max_{1 \leqslant i,j \leqslant 4} |f(\omega_i) - f(\omega_j)| = 1$). Applying the upper bound gives

$$\frac{1}{1 + (n-1)e^{-\frac{\epsilon \Delta v}{\Delta f}}} \leq \frac{1}{3} \quad (16)$$

$$\frac{1}{1 + 3e^{-\frac{\epsilon}{4}}} \leq \frac{1}{3} \quad (17)$$

$$\epsilon \leq 4 \ln(\frac{3}{2}) \approx 1.6219 \quad (18)$$



(a) $f = mean$ (b) $f = median$

Fig. 3. Distributions of each possible world for different type of queries

A more precise upper bound can be found by replacing Δv with exact values as follows:

$$\beta(\omega_i) = \frac{1}{1 + \sum_{k=1,k\neq i}^{n} e^{-\frac{|f(\omega_i)-f(\omega_k)|}{\lambda}}} \tag{19}$$

$$= \frac{1}{1 + e^{0} + e^{-\frac{1}{\lambda}} + e^{-\frac{1}{\lambda}}} \tag{20}$$

$$= \frac{1}{2 + 2e^{-\frac{1}{\lambda}}} \tag{21}$$

The inequality $\frac{1}{2+2e^{-\frac{\epsilon}{\Delta f}}} \leq \frac{1}{3}$ leads to $\epsilon \leq 4 \ln 2 \approx 2.776$.

Recall that the sensitivity of the mean function for the database X' is $\frac{17}{6} \approx 2.83$ and $\epsilon \leq 0.3829$ to limit the adversary's probability no greater than $\frac{1}{3}$. With the same universe, we can allow larger epsilon ($\epsilon \leq 2.776$) to enforce the same degree of privacy for the median whose sensitivity is larger than that of the mean. This is because the type of query affects the distributions of possible world. In Figure 3(a), given the response $\gamma < 3.16$, ω_1 is significantly more likely than others. On the other hand, in Figure 3(b), given any response value of γ, both ω_1 and ω_2 are always equally likely and difference of likelihood between each possible world is relatively small.

7 Conclusion

While the concept of differential privacy has received considerable attention in the literature, there has been little discussion of how to apply it in practice. Although ϵ is the privacy parameter for differential privacy, it does not directly correlate to a practical privacy standard. We have shown that given a practical standard, namely the risk of identifying an individual, it is possible to determine an appropriate value of ϵ. However, this requires knowing the queries to be

computed, the universe of data, and the subset of that universe to be queried. While this is not a disabling issue, as such knowledge (except the subset to be queried, presumably known to the data holder) is typically required to construct a differentially private mechanism anyway, it does raise additional research challenges. Succinctly, any discussion of a differentially private mechanism requires a discussion of how to set an appropriate ϵ for that mechanism, a challenge that may be as or more difficult than developing the mechanism itself.

Acknowledgments. Partial support for this work was provided by MURI award FA9550-08-1-0265 from the Air Force Office of Scientific Research.

References

1. Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K.: Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In: Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 273–282. ACM, New York (2007)
2. Blum, A., Dwork, C., McSherry, F., Nissim, K.: Practical privacy: the SuLQ framework. In: Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 128–138. ACM, New York (2005)
3. Blum, A., Ligett, K., Roth, A.: A learning theory approach to non-interactive database privacy. In: Proceedings of the 40th Annual ACM Symposium on Theory of Computing, pp. 609–618. ACM, New York (2008)
4. Dalenius, T.: Towards a methodology for statistical disclosure control. Statistik Tidskrift 15(429-444), 2–1 (1977)
5. Dinur, I., Nissim, K.: Revealing information while preserving privacy. In: Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 202–210. ACM, New York (2003)
6. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006, Part II. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)
7. Dwork, C.: Differential privacy: A survey of results. In: Agrawal, M., Du, D.-Z., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008)
8. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: Privacy via distributed noise generation. In: Vaudenay, S. (ed.) EUROCRYPT 2006. LNCS, vol. 4004, pp. 486–503. Springer, Heidelberg (2006)
9. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006)
10. Dwork, C., Nissim, K.: Privacy-preserving datamining on vertically partitioned databases. In: Franklin, M. (ed.) CRYPTO 2004. LNCS, vol. 3152, pp. 528–544. Springer, Heidelberg (2004)
11. Kasiviswanathan, S., Smith, A.: A note on differential privacy: Defining resistance to arbitrary side information. Arxiv preprint arXiv:0803.3946 (2008)
12. McSherry, F.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of the 35th SIGMOD International Conference on Management of Data, pp. 19–30. ACM, New York (2009)

13. Nergiz, M.E., Clifton, C.: δ-presence without complete world knowledge. IEEE Transactions on Knowledge and Data Engineering 22(6), 868–883 (2010), http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.125
14. Nergiz, M., Atzori, M., Clifton, C.: Hiding the presence of individuals from shared databases. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, Beijing, China, June 11-14, pp. 665–676 (2007), http://doi.acm.org/10.1145/1247480.1247554
15. Nissim, K., Raskhodnikova, S., Smith, A.: Smooth sensitivity and sampling in private data analysis. In: Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing, pp. 75–84. ACM, New York (2007)
16. Roy, I., Setty, S., Kilzer, A., Shmatikov, V., Witchel, E.: Airavat: Security and privacy for MapReduce. In: Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation, p. 20. USENIX Association (2010)