Differential Privacy in Practice

Maryam Shoaran, Alex Thomo, and Jens Weber

University of Victoria, Victoria, Canada
{maryam,thomo}@cs.uvic.ca, jens@uvic.ca

Abstract. Differential privacy (DP) has attracted considerable attention as the method of choice for releasing aggregate query results making it hard to infer information about individual records in the database. The most common way to achieve DP is to add noise following Laplace distribution. In this paper, we study differential privacy from a utility point of view for single and multiple queries. We examine the relationship between the cumulative probability of noise and the privacy degree. Using this analysis and the notion of relative error, we show *when* for a given problem it is *reasonable* to employ a differentially private algorithm without losing a certain level of utility. For the case of multiple queries, we introduce a simple DP method called *Differential* (*DIFF*) that adds noise proportional to a query index used to express our preferences for having different noise scales for different queries. We also introduce an equation capturing when *DIFF* satisfies a user-given relative error threshold.

Keywords: Statistical Databases, Differetial Privacy, Utility.

1 Introduction

Publishing analysis results of massive data collections, while providing substantial potential for research and public advantage, brings up the matter of privacy. Protecting sensitive information about participants has become one of the fundamental problems in society. For example, consider databases of medical records. Public release of statistical information over such data is prone to disclosing sensitive details about the health of individuals.

In recent years, ϵ-differential privacy [5] (DP for short) has become one of the foremost methods to protect information contained in individual records. DP guarantees that the privacy of an individual or a group is highly unlikely to be breached by participating in the computation of the aggregate results. The most common approach to achieve DP is by adding random noise with Laplace distribution to query answers, where the scale of noise is calibrated by the sensitivity of queries (the maximum difference in query answers on two databases differing by one tuple).

In this paper, we analyze when it is "reasonable" to use DP. We quantify *reasonable* in terms of the *relative error*, which is the ratio of the noise to the true query answer. The notion of relative error is important. For example, adding a noise of 60 to a query answer of 30 obviously is not reasonable from a utility point of view.

W. Jonker and M. Petković (Eds.): SDM 2012, LNCS 7482, pp. 14–24, 2012.

We examine the relationship between the cumulative probability of noise and the privacy degree. For single queries, we analyze when it is reasonable to employ DP without losing a certain level of utility.

Turning to the case of multiple queries, when using the method of [5], the scale of noise is the same for all the queries. To see the problem with this approach, consider two count queries, the *small* and the *big*, with answers 30 and 30,000, respectively. Having the same amount of noise, suppose 30, added to the answers of each query with the same probability will cause different "harm" to each query. While *big* easily tolerates this amount of noise without significantly affecting the utility of the answer, adding this amount of noise to *small* makes the released answer quite useless.

In order to alleviate this problem, we introduce a simple method called *Differential* that adds noise guided by a "query index," which is a set of numbers used to express our preferences for the noise scales to each query. For instance, in the above example the query index can be $\{1, 1000\}$. We show that Differential satisfies DP.

Our contributions are as follows.

1. We analyze the relationship between the cumulative probability of noise, the sensitivity, and the privacy degree (Section 3).
2. We examine when employing DP does not harm the utility beyond a certain level. We perform the analysis for two categories, single query (Section 4) and multiple queries (Section 5).
3. We propose *Differential*, a mechanism that, for the case of multiple queries, achieves DP while adding noise guided by a query index. Also, we analyze when Differential satisfies a user-given relative error threshold. (Section 6).

Related Work. There are several works that aim at controlling noise produced by DP mechanisms ([1,8,7,2]). These works study different settings from ours. They focus on reducing absolute error (not relative) and have consistency constraints (e.g. marginals that add up to some specific number).

A method, similar to our Differential, is introduced by Xiao et al. [11]. Their method called *Proportional* computes for a set of queries a set of noise scales that are proportional to the magnitude of query answers on the databases they are applied to. Proportional is shown to not satisfy DP. Another approach that [11] introduces is called *iReduct*. The latter is an iterative algorithm using a sophisticated procedure to minimize relative errors with respect to a database.

On the other hand, here we are interested in minimizing the relative error based on a query index. These preference weights might reflect the proportion of magnitudes of query answers on some static database[1], but this is not necessary.

[1] For example if we are to privately release mortality counts for different diseases for a hospital serving a big city, we can set the index to reflect the proportions of mortality rates of diseases in the city or country where the hospital is located. Certainly, such an index will approximately match the disease mortality proportions in the hospital, but this is public information, not a privately sensitive aspect of the data in the hospital database.

In fact, we can have different weights even if the magnitudes of the queries are the same. In such a case we release a more accurate answer to some query at the expense of less accurate answers to other queries in the set.

2 Background

Let q be an aggregate query. For example, q can be a count or a sum query on database D. Such a query can also be considered as a function $q : \mathbf{D} \to \mathbb{R}$, where $\mathbf{D}$ is the set of all databases. Thus, we use the terms query and function interchangeably. We denote by $q(D)$ the true answer of q on database D.

The definition of *differential privacy* (DP) uses the notion of *neighboring databases*. Two databases D_1 and D_2 are called *neighbors* if one of them can be obtained from the other by adding or removing at most one record.

Definition 1. (Differential Privacy [5]) A randomized algorithm $\mathcal{M}$ satisfies ϵ-differential privacy (ϵ-DP) if and only if for any two neighboring databases D_1 and D_2, and for any subset $S \in Range(\mathcal{M})$,

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

Dwork et al. showed in [5] that ϵ-DP can be achieved by adding appropriately chosen random noise to the true query answer $q(D)$. Specifically, noise follows a Laplace distribution with probability density function

$$f(x) = \tfrac{1}{2\lambda} e^{-|x|/\lambda} \tag{1}$$

denoted as $Lap(\lambda)$, where λ is called *noise scale*.

Definition 2. (Sensitivity *[5]*) For $q : \mathbf{D} \to \mathbb{R}$, the sensitivity of q is

$$\Delta(q) = \max_{D_1, D_2} |q(D_1) - q(D_2)|$$

for all neighboring databases D_1 and D_2.

Dwork et al. prove that an algorithm that sets the noise scale to be $\lambda = \Delta(q)/\epsilon$ enjoys ϵ-DP. Namely, when a query q is posed to database D, the output of the randomized algorithm $\mathcal{M}$ will be

$$q(D) + Lap(\Delta(q)/\epsilon).$$

The parameter ϵ is the privacy degree and one can think of it typically as 0.01 or 0.1.

The ϵ-DP can also be obtained for any sequence of queries $q_1, q_2, \ldots, q_m$ on a single database by running the algorithm $\mathcal{M}$ with noise distribution

$$Lap\left(\sum_i \frac{\Delta(q_i)}{\epsilon}\right)$$

on each computation [5].

In [4] Dwork et al. address the matter of overall privacy when the privatized output of multiple queries are released together.

Theorem 1. ([4]) A sequence of m computations over a database D, each providing ϵ_i-DP, satisfies $(\sum_i \epsilon_i)$-DP.

This is also called *sequential composition* in the literature (cf. [9]).

3 Noise and Utility

It is clear that the scale of the noise added to $q(D)$ is independent of the real magnitude of $q(D)$. Note that sensitivity is a characteristic of the computation (query) and does not depend on the database (cf. [3]).

To illustrate, for a count query q, whether $q(D)$ is for example 30 or 30,000 does not have any influence on the value of noise scale. This is because sensitivity is equal to 1 for any *count* query. Thus, for a given ϵ, the noise scale will be $\lambda = 1/\epsilon$. Therefore, an amount of noise, say 69 (a plausible value as we will show), might be added to the answer of the query with the same probability regardless of the magnitude of the true answer. While a noise value of 69 does not affect significantly the "utility" of the query answer of 30,000 magnitude, it renders the privatized answer of the query of 30 magnitude almost useless.

Therefore, we analyze in this section the practicality of differential privacy from a utility point of view.

Let $q(D)$ be the true answer and p be the released privatized answer to query q. We can evaluate the utility of the released answer using *relative error* as follows.

$$RE = \frac{|p - q(D)|}{q(D)} \tag{2}$$

This is similar to the definition of relative error in [6,10]. Practically, we can think of acceptable values of RE as 10% or 15%.

The cumulative distribution function of Laplace distribution in an interval $[-z, z]$ can be computed by the following integral.

$$Pr(-z \leq x \leq z) = \int_{-z}^{z} \tfrac{1}{2\lambda} e^{\frac{-|x|}{\lambda}} dx = 1 - e^{\frac{-z}{\lambda}}.$$

Therefore,

$$Pr(|x| \geq z) = e^{\frac{-z}{\lambda}}. \tag{3}$$

Let $pr = Pr(|x| \geq z)$. Value z for a specified cumulative probability pr can be calculated using equation (3) as

$$z = -\lambda \cdot \ln(pr),$$

and with $\lambda = \Delta(q)/\epsilon$ we have

$$z = -\frac{\Delta(q)}{\epsilon} \cdot \ln(pr). \tag{4}$$

Fig. 1 illustrates the minimum absolute noise z as a function of cumulative probability pr for three different values of ϵ when $\Delta(q) = 1$. Each point (pr, z) on the curve for a given ϵ means that

> pr percent of the time the random noise has an absolute value of at least z.

For example, for $\epsilon = 0.01$, we have that 50% of the time the absolute value of noise is at least 69, and 30% of the time it is at least 120. This means that the query answer needs to be considerably higher than 69, or even 120, in order for the privatized (released) answer to have some utility.

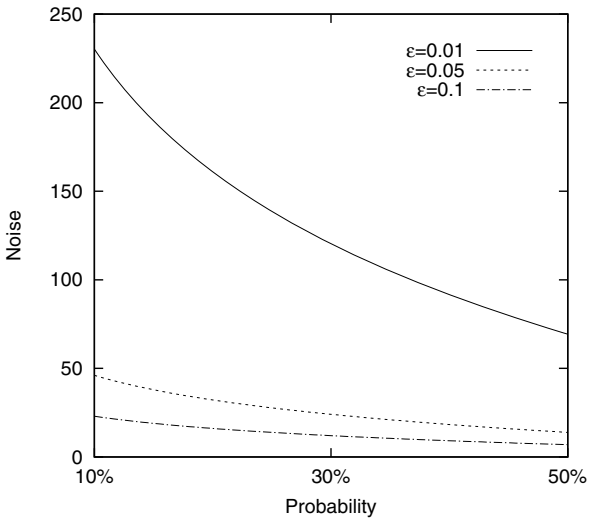

Fig. 1. Noise vs Probability ($\Delta(q) = 1$)

Next, we introduce measures to study the utility of the privatized (released) answers.

4 Single Query

Let q be a query with a single numerical output. Using Equations (2) and (4) we can compute the *minimum true (query) answer* (MTA), such that, with probability $1 - pr$, RE is below a threshold re.

The RE formula (Equation (2)) can be written as

$$RE = \frac{|z|}{q(D)}, \tag{5}$$

where z is the amount of noise added to the true answer of query q on database D. In order for RE to be below a threshold re, by Equation (4), we should have

$$\begin{aligned} q(D) &\geq \frac{|z|}{re} \\ &= -\frac{\Delta(q) \cdot \ln(pr)}{\epsilon \cdot re}. \end{aligned}$$

Thus,

$$MTA = -\frac{\Delta(q) \cdot \ln(pr)}{\epsilon \cdot re}. \tag{6}$$

Example 1. Let q be a count query, and let $\epsilon = 0.01$. Let us consider $pr = 10\%$, i.e. we want to be $1 - pr = 90\%$ sure about the relative error statements.

Recall that the sensitivity for any count *query is 1. From Equation* (4)*, we have that*

$$z = -\frac{1}{0.01} \ln(10\%) \simeq 230.$$

That is, 10% of the time the absolute value of the random noise of scale $\lambda = \frac{1}{0.01} = 100$ is at least 230. Using Equation (6)*, for an application specific RE threshold $re = 10\%$, we get*

$$MTA = \frac{230}{10\%} = 2300.$$

In plain language, the query answer should have a magnitude of at least 2300 in order for the privatized answer to have an acceptable utility ($RE \leq 10\%$) 90% of the time.

Fig. 2 illustrates MTA as a function of ϵ for three different values of pr when $re = 10\%$. Each point (ϵ, mta) on the curve for a given pr shows that:

> The query answer should have a magnitude at least mta, in order for the relative error to not be higher than 10%, $1 - pr$ of the time under noise of scale $\lambda = 1/\epsilon$.

If we substitute re for ϵ in the x-axis of Fig. 2 we get MTA as a function of relative error, with ϵ fixed to 0.1. Each point (re, mta) on the curve of a given probability pr will show that:

> If a query answer is at least mta, adding random noise of scale $\lambda = 1/0.1 = 10$ ($\epsilon = 0.1$) will satisfy $1 - pr$ of the time the requirement of having relative error at most re.

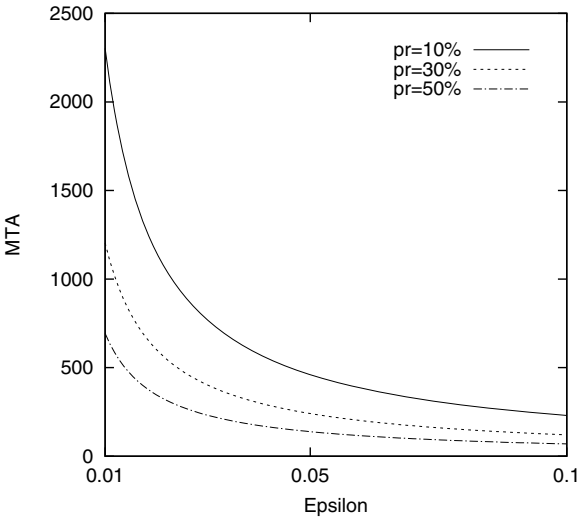

Fig. 2. MTA vs Epsilon ($re = 10\%$). We can substitute re for ϵ and get MTA as a function of relative error threshold.

5 Multiple Queries

In the case of two or more queries on a single database, a randomized algorithm satisfying ϵ-DP adds Laplace noise with scale $\lambda = \sum_i \Delta(q_i)/\epsilon$ (see Section 2). The MTA in this case will be possibly larger than what it would be if only one of the query answers were to be released. This is because the noise scale, being equal to $\lambda = \sum_i \Delta(q_i)/\epsilon$, is larger.

For example, in the case of two count queries, if a record affects both counts (at most by one), then λ will be double, $\lambda = \frac{\Delta(q_1)+\Delta(q_2)}{\epsilon} = \frac{1}{\epsilon} + \frac{1}{\epsilon}$. However, applying noise with *identical* scale to both may distort one of the answers a lot more than the other.

Example 2. Let $\epsilon = 0.01$. *Suppose* $q_1(D) = 3000$ *and* $q_2(D) = 30,000$ *are the true answers to two count queries on a database* D. *We have* $\Delta(q_1) + \Delta(q_2) = 2$, *and* $\lambda = \frac{2}{0.01} = 200$. *Based on Equation* (4), *with probability* 10%, *the absolute noise will be at least*

$$z = -200 \cdot \ln(10\%) \simeq 460.$$

Observe that each query answer is distorted twice than if they were considered in isolation. If the threshold on relative error is set to 10%*, then the MTA for each query needs to be at least 4600.*

Now, adding noise 460 to the true query answers, we have

$$RE_1 = \frac{460}{3000} \simeq 0.15$$
$$RE_2 = \frac{460}{30000} \simeq 0.015.$$

This example shows that whereas adding this amount of noise to the answer of q_2 *is reasonable for threshold* $re = 10\%$*, it distorts the answer of* q_1 *too much, thus failing to satisfy threshold* re*.*

Differential Noise Problem. Can we find an algorithm for a set of queries that satisfies DP by adding noise that has different scales for different queries?

In the sequel, we introduce a simple method called *Differential* (*DIFF*). *DIFF* satisfies DP and adds noise to each query answer guided by a "(differential) query index". A query index is a set of numbers, one for each query, used to express our preference for the scale of noise used for each query.

We note here that Xiao et al. have also proposed a method in [11], called Proportional, which is similar to what we propose here. However, in their method, the noise scales depend on each database that the randomized algorithm is applied to. They show that Proportional does not satisfy DP.

6 Differential

Let $\{q_1, \ldots, q_m\}$ be m queries. Also, let ϵ be the privacy degree we seek for the query set. In the DP mechanism of [5] (analyzed in Section 5), all queries in the set will have the same noise scale λ. Another way to view this mechanism is as follows.

If we denote $\frac{\Delta(q_i)}{\lambda}$ by ϵ_i, for $i \in [1, m]$, we have

$$\epsilon = \frac{\sum_{i=1}^{m} \Delta(q_i)}{\lambda} = \sum_{i=1}^{m} \epsilon_i.$$

Thus, query q_i, for $i \in [1, m]$, bears weight ϵ_i toward achieving overall degree of privacy ϵ.

Here we propose *Differential* (*DIFF*) which assigns each query its own noise scale λ_i. Let $\Gamma = \{\gamma_1, \ldots, \gamma_m\}$ be the query index (a set of numbers, one for each query). *DIFF* sets each λ_i to be proportional to the corresponding γ_i value. That is,

$$\lambda_i = \alpha \cdot \gamma_i \tag{7}$$

where $i \in [1, m]$, and α is some constant.

Now, query q_i, for $i \in [1, m]$, bears weight $\epsilon_i = \frac{\Delta(q_i)}{\lambda_i}$ toward achieving overall degree of privacy $\epsilon = \sum_{i=1}^{m} \epsilon_i$.

Constant α can be computed by substituting (7) in equation $\epsilon = \sum_{i=1}^{m} \frac{\Delta(q_i)}{\lambda_i}$. We get

$$\alpha = \frac{1}{\epsilon} \sum_{i \in [1,m]} \frac{\Delta(q_i)}{\gamma_i} \tag{8}$$

and then each λ_i is computed using Expression (7).

A randomized algorithm with λ_i's thus computed will satisfy ϵ-DP. To verify this, let D_1 and D_2 be two such databases. The following equation shows that privatized answers p_i's for queries q_i's are almost as likely on D_1 as on D_2 with DP privacy degree of ϵ. Namely, we have

$$\begin{aligned}
&\frac{Pr[p_1, \ldots, p_m \ on \ D_1]}{Pr[p_1, \ldots, p_m \ on \ D_2]} \\
&= \frac{\prod_{i=1}^{m} \exp(-|p_i - q_i(D_1)|/\lambda_i)}{\prod_{i=1}^{m} \exp(-|p_i - q_i(D_2)|/\lambda_i)} \\
&= \prod_{i=1}^{m} \exp\left(\frac{-z_{i,1}}{\lambda_i} + \frac{z_{i,2}}{\lambda_i}\right) \\
&\leq \prod_{i=1}^{m} \exp\left(\frac{-z_i}{\lambda_i} + \frac{z_i + \Delta(q_i)}{\lambda_i}\right) \\
&= \exp\left(\sum_{i=1}^{m} \frac{\Delta(q_i)}{\lambda_i}\right) \\
&= \exp(\epsilon)
\end{aligned}$$

where $z_{i,j}$ is the noise added to the true answer $q_i(D_j)$, and the last step is based on equation $\epsilon = \sum_{i=1}^{m} \frac{\Delta(q_i)}{\lambda_i}$. Note that α is *not* a user defined constant and depends on ϵ, on the sensitivity of the queries, and on the query index.

One might be interested in knowing whether using indexed noise as above can satisfy for each query a *user given* relative error threshold re with a cumulative probability pr on database D.

From Equations (4), (5), (7), and (8) we have

$$\begin{aligned}
RE_i &= \frac{-\lambda_i \cdot \ln(pr)}{q_i(D)} \\
&= \frac{-\alpha \cdot \gamma_i \cdot \ln(pr)}{q_i(D)} \\
&= -\left(\frac{1}{\epsilon} \sum_{i=1}^{m} \frac{\Delta(q_i)}{\gamma_i}\right) \cdot \frac{\gamma_i \cdot \ln(pr)}{q_i(D)}
\end{aligned}$$

Thus, it can be inferred that if the equation

$$\sum_{i=1}^{m} \frac{\Delta(q_i)}{\gamma_i} \leq -\frac{re_i}{\gamma_i} \cdot \frac{\epsilon \cdot q_i(D)}{\ln(pr)} \tag{9}$$

is true for a set of m queries on a database D, then at least $1 - pr$ percent of the time *DIFF* satisfies a user given relative error threshold re_i for q_i. From this we derive the MTAs for each query q_i to be

$$MTA_i = -\frac{\gamma_i \cdot \sum_{i=1}^{m} \frac{\Delta(q_i)}{\gamma_i} \cdot \ln(pr)}{\epsilon \cdot re_i}. \tag{10}$$

Observe that in the case of one query, the above becomes the same as Equation (6).

Example 3. Consider again Example 2. Let $\Gamma = \{1, 10\}$ *be the query index for the two count queries. From* (10)*, we have*

$$MTA_1 = -\frac{1 \cdot (1/1 + 1/10) \cdot \ln(10\%)}{0.01 \cdot 0.10} \simeq 2533$$
$$MTA_2 = -\frac{10 \cdot (1/1 + 1/10) \cdot \ln(10\%)}{0.01 \cdot 0.10} \simeq 25328.$$

Since the answers of q_1 *and* q_2 *are greater than their respective MTAs, we can satisfy the relative error threshold by using mechanism DIFF.*

Specifically, we set $\lambda_1 = \alpha$ *and* $\lambda_2 = 10\alpha$*. Using Equations* (8) *we have* $\alpha = \frac{1}{\epsilon}(\frac{1}{1} + \frac{1}{10}) = 110$ $(\epsilon = 0.01)$ *and from Equation* (7)*,* $\lambda_1 = 110$ *and* $\lambda_2 = 1100$*.*

Using Equation (4) *we have that 10% of the time the noise added to the true answers of* q_1 *and* q_2 *has an absolute value of at least 253.3 and 2532.8, respectively (compare these values to noise value of 460 for both queries in Example 2). Now the noise added will not violate the error threshold of 10%.*

7 Conclusions

We have analyzed differential privacy from a utility perspective. We studied the connection between the cumulative probability of noise, and the privacy degree. Using the concept of relative error we explored the practicality of DP algorithms for single and multiple queries. Namely, we analyzed the circumstances when DP can be used reasonably without exceeding a given threshold for relative error. For multiple queries, we proposed the Differential (*DIFF*) method that adds noise with scales guided by a query index. We showed that *DIFF* satisfies DP.

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: PODS, pp. 273–282 (2007)
2. Ding, B., Winslett, M., Han, J., Li, Z.: Differentially private data cubes: optimizing noise sources and consistency. In: SIGMOD Conference, pp. 217–228 (2011)

3. 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)
4. 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)
5. 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)
6. Garofalakis, M.N., Kumar, A.: Wavelet synopses for general error metrics. ACM Trans. Database Syst. 30(4), 888–928 (2005)
7. Hay, M., Rastogi, V., Miklau, G., Suciu, D.: Boosting the accuracy of differentially private histograms through consistency. PVLDB 3(1), 1021–1032 (2010)
8. Li, C., Hay, M., Rastogi, V., Miklau, G., McGregor, A.: Optimizing linear counting queries under differential privacy. In: PODS, pp. 123–134 (2010)
9. McSherry, F.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: SIGMOD Conference, pp. 19–30 (2009)
10. Vitter, J.S., Wang, M.: Approximate computation of multidimensional aggregates of sparse data using wavelets. In: SIGMOD Conference, pp. 193–204 (1999)
11. Xiao, X., Bender, G., Hay, M., Gehrke, J.: ireduct: differential privacy with reduced relative errors. In: SIGMOD Conference, pp. 229–240 (2011)