Differential Privacy - I

Prof. Dan A. Simovici

UMB

Releasing the data and just removing the names does nothing for privacy. If you know their name and a few records, then you can identify that person in the other (private) database.
Vitaly Shmatikov, Professor of Computer Science, University of Texas at Austin

Privacy in data analysis is treated from many perspectives:

- statistics,
- databases,
- philosophy,
- law,
- cryptography,
- theoretical computer science.

The developments in the

- internet,
- database technology, and
- data mining

brought to forefront the issues of privacy and has imposed limitations of the work of data analysts.

In general, data analysts do not have direct access to raw data and certain limitations are imposed on the content and number of exploring queries because accurate answers to too many questions will destroy privacy.

- exposure of medical records of governor William Weld of Massachusetts;
- identification of an user of AOL;
- identification of an user of Netflix based on movie ratings;
- massive data breaches at TJMaxx and other retailers.

Weld's Medical Records

A graduate MIT student, Latanya Sweeney managed to access the medical records of William Weld, the then governor of Massachusetts using poorly public anonymized medical records.
In his 2010 UCLA Law Review paper, "Broken Promises of Privacy, (Ohm, 2010) University of Colorado law professor Paul Ohm describes Sweeney's re-identification of Weld's hospitalization data as follows:

At the time GIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers. In response, then graduate student Sweeney started hunting for the Governors hospital records in the GIC data. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes. For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code. In a theatrical flourish, Dr. Sweeney sent the Governors health records (which included diagnoses and prescriptions) to his office.

The Searches of Mrs. Arnold on AOL

Starting from the public anonymized AOL records of the search history of an user it was possible to identify this user as Ms. Thelma Arnold.
Among a list of 20 million Web search queries collected by AOL and released on the Internet is were the searches of user No. 4,417,749. The number was assigned by the company to protect the searchers anonymity, but it was not much of a shield.
No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from numb fingers to 60 single men to dog that urinates on everything. Following searches for landscapers in Lilburn, Ga, several people with the last name Arnold and homes sold in shadow lake subdivision Gwinnett County Georgia, it became possible to identify the user as Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends medical ailments and loves her three dogs.

It might appear that Ms. Arnold fears she is suffering from a wide range of ailments. Her search history includes hand tremors, nicotine effects on the body, dry mouth and bipolar. But in an interview, Ms. Arnold said she routinely researched medical conditions for her friends to assuage their anxieties. Explaining her queries about nicotine, for example, she said: I have a friend who needs to quit smoking and I want to help her do it.

The search data was removed from the AOL site, and AOL apologized for its release. This incident shows how much people unintentionally reveal about themselves when they use search engines and how risky it can be for companies like AOL, Google and Yahoo to compile such data.
AOL chief technology officer resigned after a massive dataset of 20 million searches performed by 658,000 people was published for use in research. The data was believed to be anonymized, but revealed sensitive details of the searchers private lives, including Social Security numbers, credit-card numbers, addresses, and, in one case, apparently a searcher's intent to kill their wife.

Identification based on movie ratings

In a dramatic demonstration of the privacy dangers of databases that collect consumer habits, two researchers from the University of Texas at Austin have shown that a handful of movie ratings can identify a person as easily as a Social Security number.

The researchers – Arvind Narayanan and Vitaly Shmatikov, both from the Department of Computer Sciences at the University of Texas at Austin – claim to have identified two people out of the nearly half million anonymized users whose movie ratings were released by online rental company Netflix last year. The company published the large database as part of its $1 million Netflix Prize, a challenge to the world's researchers to improve the rental firm's movie-recommendation engine.

While Netflix's dataset did not include names, instead using an anonymous identifier for each user, the collection of movie ratings – combined with a public database of ratings (IMDb -standing for Internet Movie Database) was enough to identify the people.

Narayanan and Shmatikov identified movie ratings of two of the users in Netflix's data.

Exposing movie ratings that the reviewer thought were private could expose significant details about the person. For example, the researchers found that one of the people had strong – ostensibly private – opinions about some liberal and gay-themed films and also had ratings for some religious films.

Massive data breaches

In the past few years several massive data breaches that have leaked sensitive information on millions of people.

- Recently the head of HM Revenue & Customs, the United Kingdom's tax agency, resigned after two data discs containing sensitive, yet unencrypted, personal details of 25 million U.K. citizens were lost in the mail.
- Retail giant TJX Companies announced that data thieves had stolen the credit- and debit-card details on, what currently is estimated to be, more than 94 million consumers.

Conclusions

- Privacy research demonstrated that information that a person believes to be benign could be used to identify them in other private databases, a risk understood in privacy and intelligence circles.
- Even as early as decades as go, the U.S. government would classify aggregates of information, (because) you can take unclassified data and put them together to get something that is not unclassified.

- Those risks have long pitted privacy advocates against online marketers and other Internet companies seeking to profit from the Internets unique ability to track the comings and goings of users, allowing for more focused and therefore more lucrative advertising.
- The unintended consequences of all that data being compiled, stored and cross-linked is a ticking privacy time bomb.

- Differential privacy (DP) is an approach to privacy that guarantees that the distribution of outcomes of a computation that involves a database does not change significantly when an individual record is added or removed from a database.
- DP allows an investigator to learn information about a database without learning anything about an individual.
- DP database mechanisms can make confidential data widely available for accurate data analysis, without resorting to data clean rooms, institutional review boards, data usage agreements, restricted views, or data protection plans.

- DP ensures that the ability of an adversary to inflict harm is the same, independent of whether any individual opts in to, or opts out of, the dataset.
- DP focuses on the probability of any given output of a privacy mechanism and how this probability can change with the addition or deletion of any row. Thus, we concentrate on pairs of databases differing only in one row, meaning one is a subset of the other and the larger database contains just one additional row.

Probability Simplex

The probability simplex in $\mathbb{R}^m$ is the set $S_m \subseteq \mathbb{R}^m$ defined by

$$S_m = \{\mathbf{x} \in \mathbb{R}^m \mid x_i \geqslant 0 \text{ for } 1 \leqslant i \leqslant m \text{ and } \sum_{i=1}^{m} x_i = 1\}.$$

Randomized Algorithm

A *randomized algorithm with domain A and range B* is a triplet $\mathcal{M} = (A, B, M)$, where $M : A \longrightarrow S_m$ (for $m = |B|$) is a function such that on input $a \in A$, $\mathcal{M}$ produces an output $b = \mathcal{M}(a) \in B$ with the probability $M(a)_b$.

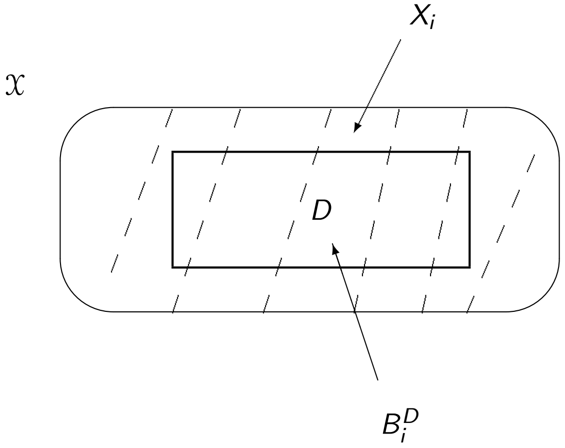
In other words, for every $a \in A$ a randomized algorithm defines a random variable:

$$\xi_a^{\mathcal{M}} : \begin{pmatrix} b_1 & \cdots & b_m \\ M_a(b_1) & \cdots & M_a(b_m) \end{pmatrix},$$

where $m = |B|$.

A universe is a pair $(\mathcal{X}, I)$, where $\mathcal{X}$ is a *set of records* and I is a *set of types* such that each record of $\mathcal{X}$ is associated with exactly one type in I.
Thus, there exists a partition τ of $\mathcal{X}$ whose blocks X_i (also known as *bins*) consist of records of type i.
The number of records of type i in $\mathcal{X}$ is denoted by $s(i)$, where the mapping $s : I \longrightarrow \mathbb{N}$ is a *histogram*.
A *database* is a set of records D drawn from a *universe* $\mathcal{X}$.

The blocks of the trace of τ on D, τ_D consist of records that have the same type i. This, in turn, defines a database histogram $s_D : I \longrightarrow \mathbb{N}$, where $s_D(i)$ is the number of records of type i contained by D.



The histogram of D is a mapping $s : I \longrightarrow \mathbb{N}$, where each entry $s(i) = s_D(i)$ represents the number of records of type i in D. Thus, the set of histograms of databases of the universe $\mathcal{X}$ is $\mathbb{N}^I$.
The norm of a histogram $s \in \mathbb{N}^I$ of a database D is $\| s \|_1 = \sum_{i=1}^{|I|} s(i)$.

- Under this definition $\| s \|_1$ is a measure of the size of the database D.
- If D_1, D_2 are two databases of the universe $\mathcal{X}$, then $\| s_{D_1} - s_{D_2} \|_1$ measures how many records are different in the two database histograms s_{D_1} and s_{D_2}.

A randomized algorithm $\mathcal{M} = (\mathsf{DB}(\mathcal{X}), \mathcal{X}, M)$ transforms the members of a database D of a universe $\mathcal{X}$ into members of $\mathcal{X}$ and, therefore into a new database D'. This is a random transformation in general, and the number of records of type i in the new database is a random variable.
Two databases $D_1, D_2 \in \mathcal{X}^n$ differ in one record if their symmetric difference consists of two records $x_1 \in D_1$ and $x_2 \in D_2$, that is, $D_1 \oplus D_2 = \{x_1, x_2\}$.

Definition

Let $\mathcal{M}$ be a randomized algorithm $\mathcal{M} : \mathcal{X}^n \longrightarrow B$. $\mathcal{M}$ is ϵ-differentially private if for every two databases D_1 and D_2 in $\mathcal{X}^n$ such that $|D_1 \oplus D_2| = 2$ we have:

$$P(\mathcal{M}(D_1) \in S) \leqslant e^{\epsilon} P(\mathcal{M}(D_2) \in S)$$

for all S in the range of $\mathcal{M}$.

If the roles of the databases are inverted we have $P[\mathcal{M}(D_2) \in S] \leqslant e^{\epsilon} P[\mathcal{M}(D_1) \in S]$. Thus, $\mathcal{M}$ is ϵ-differentially private if and only if for every two databases D_1, D_2 on $\mathcal{X}^n$ such that $|D_1 \oplus D_2| \leqslant 2$ we have

$$e^{-\epsilon} \leqslant \frac{P[\mathcal{M}(D_1) \in S]}{P[\mathcal{M}(D_2) \in S]} \leqslant e^{\epsilon}. \tag{1}$$

Note that, taking into account that for small values of ϵ we have $e^{\epsilon} \approx 1 + \epsilon$ and $e^{-\epsilon} \approx 1 - \epsilon$, Inequality 1 becomes

$$1 - \epsilon \leqslant \frac{P[\mathcal{M}(D_1) \in S]}{P[\mathcal{M}(D_2) \in S]} \leqslant 1 + \epsilon.$$

The quantity $\frac{\ln P[\mathcal{M}(D_1) \in S]}{\ln P[\mathcal{M}(D_2) \in S]}$ is the privacy loss incurred by observing the output $\mathcal{M}(D_1)$.

The definition of differential privacy ensures that seeing D_2 instead of D_1 can only increase the probability of any event by at most a small factor.

A more general concept is the notion of (ϵ, δ)-differential privacy.

Definition

Let $\mathcal{M}$ be a randomized algorithm $\mathcal{M} : \mathcal{X}^n \longrightarrow B$. $\mathcal{M}$ is (ϵ, δ)-*differentially private* if for every two databases D_1 and D_2 in $\mathcal{X}^n$ such that $|D_1 \oplus D_2| = 2$ and for all S in the range of $\mathcal{M}$ we have:

$$P(\mathcal{M}(D_1) \in S) \leqslant e^{\epsilon} P(\mathcal{M}(D_2) \in S) + \delta.$$

Example

Consider a database D of individuals draw from a population $\mathcal{X}$ that may or may not smoke. We present a technique that allows us to estimate the fraction p of individuals who smoke by using a randomized survey that preserves the privacy of individual responders. The individuals are instructed to answer `yes` or `no` when questioned about their smoking as follows:

i. flip a coin;
ii. if `tail`, then respond truthfully;
iii. if `head`, then flip a second coin and respond "`yes`" if `head` and "`no`" if `tail`.

Example cont'd

Let a be an individual who smokes. We have the following scenario:

1. If the first coin toss produces a `tail`, a responds truthfully and the answer is `yes`.
2. If the first coin toss produces `head`, a second coin is tossed and the answer is q dictated by tossing of the coin: if `head` the individual will respond `yes` (which is the truth); if `tail` the answer will be `no` (which is not truthful.

The distribution of the answer is

$$\xi_a^{\mathcal{M}} : \begin{pmatrix} \mathtt{yes} & \mathtt{no} \\ 3/4 & 1/4 \end{pmatrix}.$$

Example cont'd

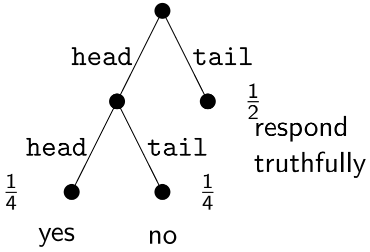
If a does not smoke the distribution of the answer is

$$\xi_a^{\mathcal{M}} : \begin{pmatrix} \texttt{yes} & \texttt{no} \\ 1/4 & 3/4 \end{pmatrix}.$$

Noise is introduced in this experiment through the spurious `yes` and `no` answers obtained by coin tossing.

Example cont'd

A `yes` answer is not incriminating because this answer occurs with probability at least $1/4$ regardless whether the respondent smokes. This provides plausible deniability to participants.



Example cont'd

- If p is the fraction of individuals who smoke, the expected number of $\mathtt{yes}$ answers in the histogram s' is $n = \frac{1}{4}(1-p) + \frac{3}{4}p = \frac{1}{4} + \frac{p}{2}$. Thus, p can be estimated as $2n - \frac{1}{2}$.
- The expected number of $\mathtt{no}$ answers in s' is $\frac{1}{4}p + \frac{3}{4}(1-p)$.
- The randomized algorithm discussed above has $(\ln 3, 0)$ privacy. Note that the range of $\mathcal{M}$ is the set $\{\mathtt{yes}, \mathtt{no}\}$.

Example cont'd

If D_1, D_2 are such that $\{x_1, x_2\} = D_1 \oplus D_2$, where $x_1 \in D_1 - D_2$ and $x_2 \in D_2 - D_1$ we may have the following four cases:

1. $\mathcal{M}(x_1) = \mathcal{M}(x_2) = \texttt{yes}$;
2. $\mathcal{M}(x_1) = \mathcal{M}(x_2) = \texttt{no}$;
3. $\mathcal{M}(x_1) = \texttt{yes}$ and $\mathcal{M}(x_2) = \texttt{no}$;
4. $\mathcal{M}(x_1) = \texttt{no}$ and $\mathcal{M}(x_2) = \texttt{yes}$.

Example cont'd

In the first two cases, we have

$$\frac{P[\mathcal{M}(D_1) \in S]}{P[\mathcal{M}(D_2) \in S]} = 1$$

for $S = \{yes\}$ or $S = \{no\}$.
In the third case, four subcases are possible:

i. both x_1 and x_2 are smokers;
ii. neither x_1 nor x_2 are smokers;
iii. x_1 is a smoker, but x_2 is not a smoker;
iv. x_1 is not a smoker but x_2 is one.

Example cont'd

The probabilities the first subcase are:

$$\frac{P(\mathcal{M}(x_1)=\texttt{no})}{P(\mathcal{M}(x_2)=\texttt{no})} = \frac{1/4}{1/4} = 1;$$
$$\frac{P(\mathcal{M}(x_1)=\texttt{yes})}{P(\mathcal{M}(x_2)=\texttt{no})} = \frac{3/4}{1/4} = 3;$$
$$\frac{P(\mathcal{M}(x_1)=\texttt{no})}{P(\mathcal{M}(x_2)=\texttt{yes})} = \frac{1/4}{3/4} = 1/3;$$
$$\frac{P(\mathcal{M}(x_1)=\texttt{yes})}{P(\mathcal{M}(x_2)=\texttt{yes})} = \frac{3/4}{3/4} = 1.$$

In the second subcase we have:

$$\frac{P(\mathcal{M}(x_1)=\texttt{no})}{P(\mathcal{M}(x_2)=\texttt{no})} = \frac{3/4}{3/4} = 1;$$
$$\frac{P(\mathcal{M}(x_1)=\texttt{yes})}{P(\mathcal{M}(x_2)=\texttt{no})} = \frac{3/4}{1/4} = 3;$$
$$\frac{P(\mathcal{M}(x_1)=\texttt{no})}{P(\mathcal{M}(x_2)=\texttt{yes})} = \frac{1/4}{3/4} = 1/3;$$
$$\frac{P(\mathcal{M}(x_1)=\texttt{yes})}{P(\mathcal{M}(x_2)=\texttt{yes})} = \frac{1/4}{1/4} = 1.$$

Example cont'd

For the third subcase (x_1 is a smoker, but x_2 is not a smoker) we can write:

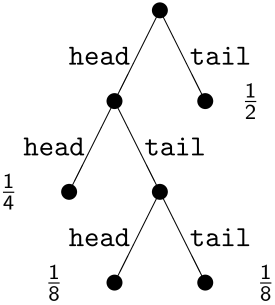
$$\frac{P(\mathcal{M}(x_1)=\texttt{no})}{P(\mathcal{M}(x_2)=\texttt{no})} = \frac{1/4}{1/4} = 1;$$
$$\frac{P(\mathcal{M}(x_1)=\texttt{yes})}{P(\mathcal{M}(x_2)=\texttt{no})} = \frac{3/4}{1/4} = 3;$$
$$\frac{P(\mathcal{M}(x_1)=\texttt{no})}{P(\mathcal{M}(x_2)=\texttt{yes})} = \frac{1/4}{3/4} = 1/3;$$
$$\frac{P(\mathcal{M}(x_1)=\texttt{yes})}{P(\mathcal{M}(x_2)=\texttt{yes})} = \frac{3/4}{3/4} = 1.$$

Example cont'd

Finally, for the fourth subcase (x_1 is not a smoker but x_2 is one):

$$\frac{P(\mathcal{M}(x_1)=\texttt{no})}{P(\mathcal{M}(x_2)=\texttt{no})} = \frac{3/4}{1/4} = 3;$$
$$\frac{P(\mathcal{M}(x_1)=\texttt{yes})}{P(\mathcal{M}(x_2)=\texttt{no})} = \frac{1/4}{1/4} = 1;$$
$$\frac{P(\mathcal{M}(x_1)=\texttt{no})}{P(\mathcal{M}(x_2)=\texttt{yes})} = \frac{3/4}{3/4} = 1/3;$$
$$\frac{P(\mathcal{M}(x_1)=\texttt{yes})}{P(\mathcal{M}(x_2)=\texttt{yes})} = \frac{1/4}{3/4} = 1/3.$$

A more general randomization algorithm can be developed using a binary tree, that is, a tree where every vertex with the exception of the leaves has two descendants. These descendants correspond to results of flipping a coin, that is, to `head` and `tail`.



Definition

Let $\mathcal{X}$ be a universe and let $f : \mathcal{X}^n \longrightarrow \mathbb{R}^d$ be a function. The L_1-sensitivity of f is the smallest number $S(f)$ such that for all $D, \tilde{D} \in \mathcal{X}^n$ which differ in a single entry we have

$$\parallel f(D) - f(\tilde{D}) \parallel_1 \leqslant S(f) d(D, \tilde{D}),$$

where d is the Hamming distance on $\mathcal{X}^n$.

In particular, if D, D' are two databases that differ in one position, $\parallel f(D) - f(\tilde{D}) \parallel_1 \leqslant S(f)$.

Example

If $\mathcal{X} = \{0, 1\}$ and $f(D) = \sum_{i=1}^{n} x_i$, then the sensitivity of f is 1.

Example

Suppose that a domain $\mathcal{X}$ has been partitioned into d bins $X_1, \ldots, X_n$. The histogram $s : \mathcal{X}^n \longrightarrow \mathbb{R}^d$ that computes the number of points that fall into each bin has sensitivity 2 independent of d because changing one point in a database can change at most two of these points: one bin loses a point and another gains one.

Example

Consider the analyst of a private student database, where x_t is the total number of students, x_A, x_B, x_C, x_D, x_F are the numbers of students who received A, B, C, D, or F, respectively, and x_p is the number of students with passing grades (D or higher).
A query that seeks to determine $(x_A, x_B, x_C, x_D, x_F)$ has sensitivity one because adding or removing a student changes exactly one of the variables.
If wee seek to determine $(x_A, x_B, x_C, x_D, x_F, x_t, x_p)$ the sensitivity is 3 (one change could affect three return values!

Laplace Distribution

Definition

A random variable has the Laplace (μ, b) distribution if its probability density function is

$$h(x|\mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$

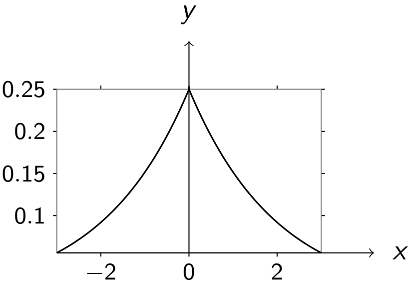
for $x \in \mathbb{R}$, where $b > 0$.

In particular, the *Laplace distribution* centered ar 0 with scale 1 is the distribution with the probability density given by

$$h(x|0, b) = \frac{1}{2b} e^{-\frac{|x|}{b}}.$$

This function will be denoted simply by $h_b(x)$.

The Laplace distribution can be regarded as a symmetric version of the exponential distribution.
It is easy to see that $\int_{\mathbb{R}} \text{Lap}(x|\mu, b)\, dx = 1$. The mean of this distribution is μ, while the variance is $2b^2$. The probability density function for the Laplace distribution with $\mu = 0$ and $b = 1$ is shown below.



The privacy preserving Laplace mechanism computes the query $f(D) \in \mathbb{R}^k$ and perturbs each coordinate $f(D)_i$ with noise drawn from a Laplace distribution Y_i.

Definition

Given a query $f : \mathrm{DB}(\mathcal{X}^n) \longrightarrow \mathbb{R}^k$, the Laplace mechanism is defined as

$$\mathcal{M}(D, f, \epsilon) = f(D) + (Y_1, \ldots, Y_k)$$

where $Y_1, \ldots, Y_k$ are independent, identically distributed Laplace variables from $\mathrm{Lap}\left(\frac{S(f)}{\epsilon}\right)$.

Example

Suppose $D \in \{0,1\}^n$ and the user wishes to learn $f(D) = \sum_{i=1}^{n} x_i$, that is, the total number of 1s in D.
If we add random Laplace noise $Y \sim Lap(1/\epsilon)$ (that is, a Laplace random variable with the parameter $b = \frac{1}{\epsilon}$ and the probability density function $h(x) = \frac{\epsilon}{2} e^{-\epsilon|x|}$), then the algorithm will produce $T(D)$, where $T(D) = \sum_{i=1}^{n} x_i + Y$. Note that $T(D) = t$ is equivalent to

$$Y = t - \sum_{i=1}^{n} x_i = t - f(D).$$

Example cont'd

Let D and $\tilde{D}$ be two databases that differ in a single entry. We have:

$$\begin{aligned}\frac{P(T(D)=t)}{P(T(\tilde{D})=t)} &= \frac{h(t-f(D))}{h(t-f(\tilde{D}))} \\ &\leqslant e^{\epsilon|f(D)-f(\tilde{D})} \leqslant e^{\epsilon}\end{aligned}$$

because the two databases D and $\tilde{D}$ differ in a single entry (which means that the sums $f(D)$ and $f(D')$ differ by 1) which shows that we have ϵ-privacy.

Definition

Given a query $f : \mathrm{DB}(\mathcal{X}^n) \longrightarrow \mathbb{R}^k$, the Laplace mechanism is defined as

$$\mathcal{M}(D, f, \epsilon) = f(D) + (Y_1, \ldots, Y_k)$$

where $Y_1, \ldots, Y_k$ are independent, identically distributed Laplace variables from $\mathrm{Lap}\left(\frac{\mathrm{GS}^n(f)}{\epsilon}\right)$.

Note that for the Laplace density function h we have:

$$\frac{h_b(y)}{h_b(y')} = \frac{e^{-\frac{|y|}{b}}}{e^{-\frac{|y'|}{b}}} = e^{\frac{|y'|-|y|}{b}} \leqslant e^{\frac{|y-y'|}{b}}$$

for $y, y' \in \mathbb{R}$.

Example

Suppose $D \in \{0,1\}^n$ and the user wishes to learn $f(D) = \sum_{i=1}^n x_i$, that is, the total number of 1s in D.
If we add random Laplace noise $Y \sim Lap(1/\epsilon)$ (that is, a Laplace random variable with the parameter $b = \frac{1}{\epsilon}$ and the probability density function $h(x) = \frac{\epsilon}{2} e^{-\epsilon|x|}$), then the algorithm will produce $T(D)$, where

$$T(D) = \sum_{i=1}^n x_i + Y.$$

Note that $T(D) = t$ is equivalent to

$$Y = t - \sum_{i=1}^n x_i = t - f(D).$$

Let D and $\tilde{D}$ be two databases that differ in a single entry. We have:

$$\frac{P(T(D) = t)}{\quad} = \frac{h(t - f(D))}{\quad}$$

Noise that must be added to a querying algorithm can be calibrated to achieve differential privacy according to the sensitivity of a query. If we have a querying mechanism $\mathcal{M}(D, f, \epsilon) = f(D) + \mathbf{y}$, where the noise $\mathbf{y}$ is drawn from $(Y_1, \ldots, Y_k)$, the density function of $(Y_1, \ldots, Y_k)$ at $\mathbf{y}$ is proportional to $e^{-\frac{\|\mathbf{y}\|_1}{b}}$. Thus, for all $t \in \mathbb{R}^d$ we have

$$\frac{P(\mathbf{z} + Y = t)}{P(\tilde{\mathbf{z}} + Y = t)} \in \{e^{\frac{\|\mathbf{z}-\tilde{\mathbf{z}}\|_1}{b}}, e^{-\frac{\|\mathbf{z}-\tilde{\mathbf{z}}\|_1}{b}}, \}$$

Thus, to release a perturbed value $f(D)$ while satisfying ϵ-differential privacy it suffices to add Laplace noise with standard deviation $\frac{S(f)}{\epsilon}$ in each coordinate.

Sequential Composition

Definition

Let $\mathcal{M}_i$ be two private algorithms that are ϵ_i differentially private, for $i = 1, 2$. The composition of $\mathcal{M}_1$ and $\mathcal{M}_2$ is the algorithm $\mathcal{M}_{1,2}$ defined as

$$\mathcal{M}_{1,2}(D) = (\mathcal{M}_1(D), \mathcal{M}_2(D))$$

Sequential Composition

Suppose that for $i = 1, 2$ $\mathcal{M}_i$ are private ϵ_i differential algorithms, respectively and that D, D' are databases that differ in one position. We have

$$\begin{aligned} P(\mathcal{M}_1(D) = S_1] &\leqslant e^{\epsilon_1} P(\mathcal{M}_1(D') = S_1), \\ P(\mathcal{M}_2(D) = S_2] &\leqslant e^{\epsilon_2} P(\mathcal{M}_2(D') = S_2). \end{aligned}$$

Then,

$$\begin{aligned} &P[(\mathcal{M}_1(D) = S_1), (\mathcal{M}_2(D) = S_2)] \\ &= P[(\mathcal{M}_1(D) = S_1)] P[(\mathcal{M}_2(D) = S_2)] \\ &\leqslant e^{\epsilon_1} P[(\mathcal{M}_1(D') = S_1)] e^{\epsilon_2} P[(\mathcal{M}_2(D') = S_2)] \\ &= e^{\epsilon_1 + \epsilon_2} P[(\mathcal{M}_1(D') = S_1)] P[(\mathcal{M}_2(D') = S_2)], \end{aligned}$$

hence the composition is $(\epsilon_1 + \epsilon_2)$-private.

Parallel Composition

Suppose that a database is partitioned in disjoint sets and the data in each of these sets is subjected to differential private analysis. The ultimate privacy guarantee depends only on the worst of the guarantees of each analysis. Let $\mathcal{D}$ and $\mathcal{D}'$ be two databases, and suppose that $\mathcal{D}$ is partitioned into $A_1, \ldots, A_n$ and $\mathcal{D}'$ is partitioned into $B_1, \ldots, B_n$.
If $\mathcal{D}$ and $\mathcal{D}'$ differ in one position, then at most one of the pairs (A_i, B_i) differ in one position, say j. Then

Parallel Composition

If $\mathcal{D}$ and $\mathcal{D}'$ differ in one position, then at most one of the pairs (A_i, B_i) differ in one position, say j. Then

$$\begin{aligned}
P(\mathcal{M}(D) = (r_1, \ldots, r_n)) &= \prod_{i=1}^{n} P(\mathcal{M}(A_i) = r_i) \\
&= \prod_{i=1, i \neq j}^{n} P(\mathcal{M}(A_i) = r_i) \cdot P(\mathcal{M}(A_j) = r_j) \\
&= \prod_{i=1, i \neq j}^{n} P(\mathcal{M}(B_i) = r_i) \cdot e^{\epsilon_j} P(\mathcal{M}(B_j) = r_j) \\
&= P(\mathcal{M}(D) = (r_1, \ldots, r_n)) e^{\epsilon_j}.
\end{aligned}$$

Since this inequality must be satisfied for all j, we have

$$P(\mathcal{M}(D) = (r_1, \ldots, r_n)) \leqslant e^{\max\{\epsilon_j \mid 1 \leqslant j \leqslant n\}} \cdot P(\mathcal{M}(D') = (r_1, \ldots, r_n)),$$

so the parallel composition is $\max\{\epsilon_j \mid 1 \leqslant j \leqslant n\}$-private.

Lemma

If $Y \sim Lap(b)$, then $P(|Y| > tb) = e^{-t}$.

Proof.

This fact follows immediately from

$$\begin{aligned} P(|Y| > tb) &= P(Y > tb) + P(Y < -tb) \\ &= 1 - \int_{-tb}^{tb} h_b(t)\, dt \\ &= 1 - 2\int_0^{tb} \frac{1}{2b} e^{-\frac{t}{b}}\, dt = e^{-t}. \end{aligned}$$

□

Theorem

Let $f : DB(\mathcal{X}^n) \longrightarrow \mathbb{R}^k$ *and let* $\mathbf{y} = \mathcal{M}(D, f, \epsilon)$, *where* $\mathcal{M}$ *is a* k*-dimensional Laplace mechanism. For every* $\delta \in (0, 1]$ *we have:*

$$P\left(\| f(\mathbf{x}) - y \|_\infty \geqslant \ln \frac{k}{\delta} \cdot \frac{S(f)}{\epsilon}\right) \leqslant \delta.$$

Proof

We have

$$
\begin{aligned}
& P\left(\| f(\mathbf{x}) - y \|_\infty \geqslant \ln \frac{k}{\delta} \cdot \frac{S(f)}{\epsilon}\right) \\
= \quad & P\left(\max_{1 \leqslant i \leqslant k} |Y_i| \geqslant \ln \frac{k}{\delta} \cdot \frac{S(f)}{\epsilon}\right) \\
\leqslant \quad & k \cdot P\left(\max_{1 \leqslant i \leqslant k} |Y_i| \geqslant \ln \frac{k}{\delta} \cdot \frac{S(f)}{\epsilon}\right) \\
= \quad & k \frac{\delta}{k} = \delta,
\end{aligned}
$$

where the inequality follows from the fact that each Y_i is distributed Lap$(\frac{S(f)}{\epsilon}$ and from the previous Lemma.

Example

Suppose that we have a list of 10,000 potential name and we wish to compute which first name were the most common in a national census. Take

$$k = 10,000, \delta = 0.05, \text{ and } \frac{S(f)}{\epsilon} = 1.$$

Note that the sensitivity of this query is 1 because every person may have only one first name. By the theorem, we can calculate the frequency of all 10,000 names with (1,0)-differential privacy and with the probability 95%, no estimate will be off am additive error of $\ln \frac{10000}{.05} \approx 12.2$.

The PINQ Architecture

PINQ was proposed as an architecture for data analysis with differential privacy. It presents a wrapper to C# LINQ language for database access, which enforces differential privacy.

- A data curator allocates a *privacy budget* ϵ for each user of the interface. The user can use this interface to execute aggregate queries such as count (`NoisyCount`), sum (`NoisySum`) and average (`NoisyAvg`). The wrapper uses Laplace noise to enforce differential privacy.
- When queries are executed on disjoint data sets, as we saw above, the privacy costs do not add up. This is achieved in PINQ using the operator `Partition` which divides a data set into multiple disjoint sets according to a user-defined function. This ensures a better use of the privacy budget allocated to the user.

Any data miner should plan ahead the number of queries and the values of ϵ to request for each, because the premature exhaustion of the privacy budget will block access to the data for the data miner.