Privacy Aware Learning

In Gaussian Mixture Model Using Expectation-Maximization method.

18.03.2018

Ankit Singh | 121501003

Ayush Mittal | 111501035

Suman Saurav Panda | 111501037

PROBLEM FORMULATION

While machine learning is one of the fastest growing technologies in the area of computer science, the goal of analyzing large amounts of data for information extraction collides with the privacy of individuals.

This is especially important for algorithms that rely on a large dataset containing previously analyzed data to learn from or which work by iteratively enhancing the global result by continuously increasing the amount of information.

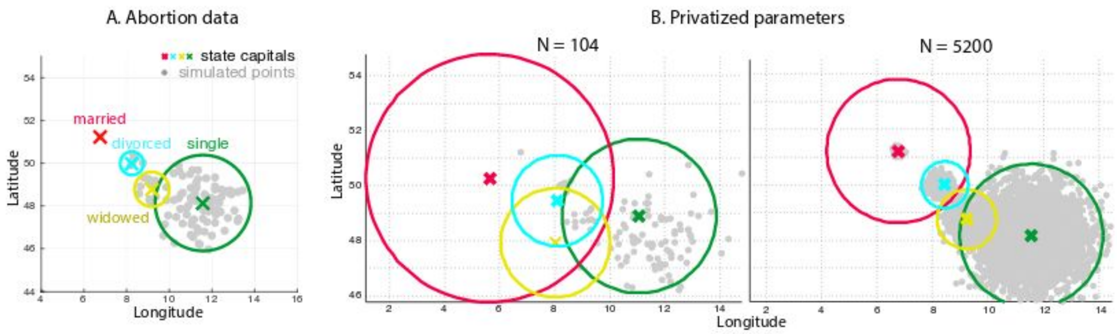
Simply anonymizing data or encrypting it is not helpful as the adversary always have access to auxiliary data hence he has the power to infer who has taken part in the survey for creating the desired model. But if we notice this there is a huge privacy risk if he is able to solve the membership problem of the user that is whether he has taken part in the data set or not he can know lot of information which the person wished to be keep as secret. For example let say we are studying AWGS (data set for genome study) if we know who has taken part in the study and then can infer what diseases he has then it will be a serious issue as the data was meant to be kept as secret.

In our attempt to overcome the same, in this project, we study a machine learning problem under a version of privacy in which the data is kept confidential even from the learner.

In this project we are going to implement privacy for parameter estimation of a clustering dataset where we use GMM EM algorithm to find the parameter of the data set.

Where is the privacy breach

Suppose say we have dataset where we have let say we have a particular cluster where only one data point is available. If we correctly find all the parameters of the given data set we know all the mean variance and probability of finding those clusters. With some auxiliary information if we know that there exist a data set which is the only data set of its cluster we can infer lot of things from this information. Here our privacy is compromised hence what we want is instead of giving the real parameter to the learner we will add some noise and then give it to him so that he won’t be able to correctly identify what was the data set present in the cluster.



Here is an example of what I have just mentioned in above image.

So if we know there is only one couple for the married cluster who have done abortion then we can find their exact location which is not desirable in our privacy preserving model.

Before proceeding I will formally present what is the definition of privacy in such statistical model.

LITERATURE SURVEY

For this project we have referred a lot of online material and videos but the best finding which was exactly solved in one paper is this microsoft research paper where they have discussed a new way implementing privacy in GMM-EM along with the traditional way of solving the same problem by perturbing the parameters in each iteration of the EM-algorithm to estimate the parameter.

https://arxiv.org/pdf/1605.06995.pdf

Some other good tutorials and video from where we are able to understand what is privacy learning is given below.

https://www.youtube.com/watch?v=Gx13lgEudtU

https://www.youtube.com/watch?v=hyDyOVFqm_U

https://www.youtube.com/watch?v=FYokHdzgJSg&t=1585s

http://www.win-vector.com/blog/2015/10/a-simpler-explanation-of-differential-privacy/

https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0ahUKEwixqMGf9-TZAhWMMY8KHdzzAbkQFgg4MAE&url=http%3A%2F%2Fcseweb.ucsd.edu%2F~kamalika%2Fpubs%2FWIFStutorial.pdf&usg=AOvVaw1PpQT-F8o0tRV3bLTeB8U3

Especially the videos on differential privacy by Cynthia Dwork from microsoft research were helpful.

https://www.youtube.com/watch?v=OfWj89oRD7g&t=1118s

RESEARCH BACKGROUND

WHAT IS PRIVACY

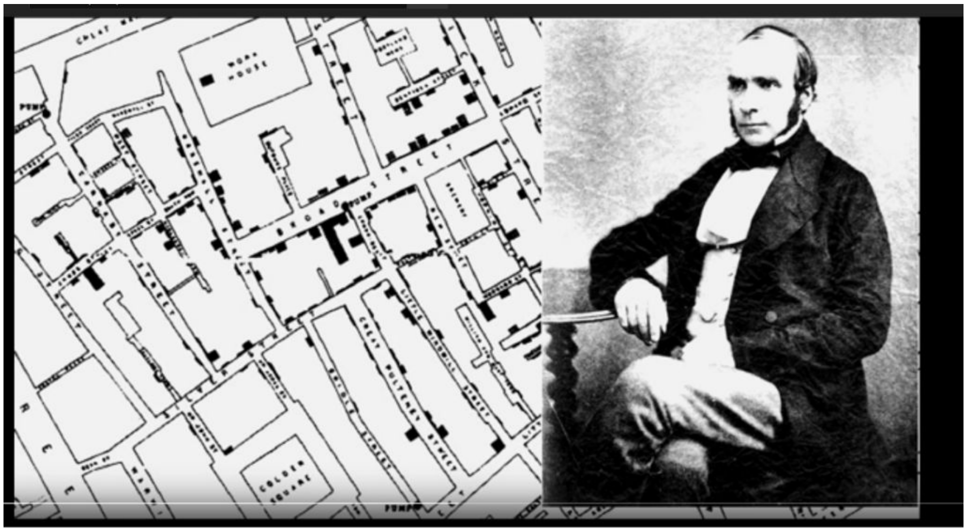
We have big data and we want to create a model where one can't extract personal data from the model or in other words by participating in the survey the person is not going to lose much in terms of privacy. Why we need such model, the answer is for public good.

For example let say we want to create a statistical model where we want to see the effect of smoking on some disease let say cancer. For that we need to survey some persons who are suffering cancer but this information is private and they are willing to share if no one can infer it by just looking the dataset for this model.

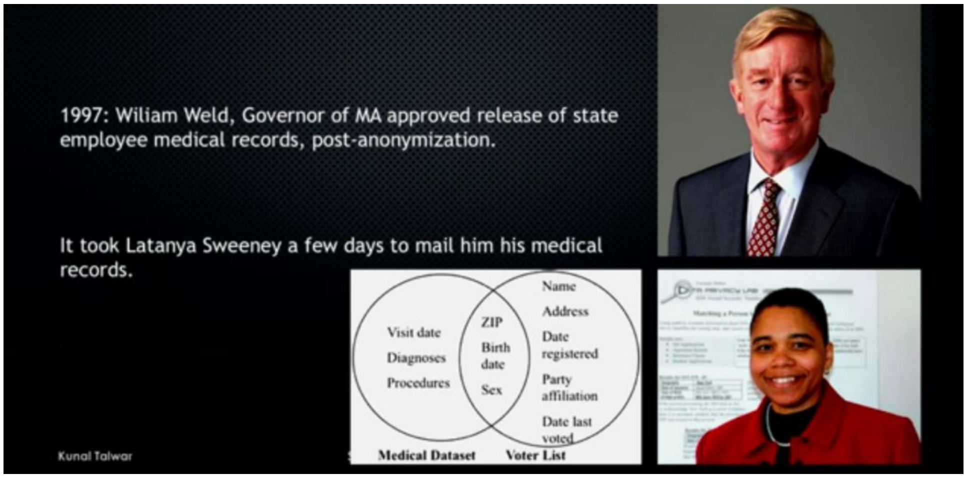
We can see this model is for public good from which we can infer smoking causes cancer but there is a risk of revealing the data of those personnel who has taken part for creating the dataset.

SOME GOOD EXAMPLE TO UNDERSTAND PRIVATE IN USE PUBLIC GOOD

The first attempt to do this was at 1860 in England where cholera broke out and we used private data to infer that a pump house was actual reason for the cholera



At 1997 we got the first medical report of government employee to create a model which will be helpful for public good. Whatever anonymization could have done at that time was done for the privacy of data but the auxiliary information like voter id detail was able to identify whose medical report is this from the anonymized data.



1997: Wiliam Weld, Governor of MA approved release of state employee medical records, post-anonymization.

It took Latanya Sweeney a few days to mail him his medical records.

Visit date
Diagnoses
Procedures

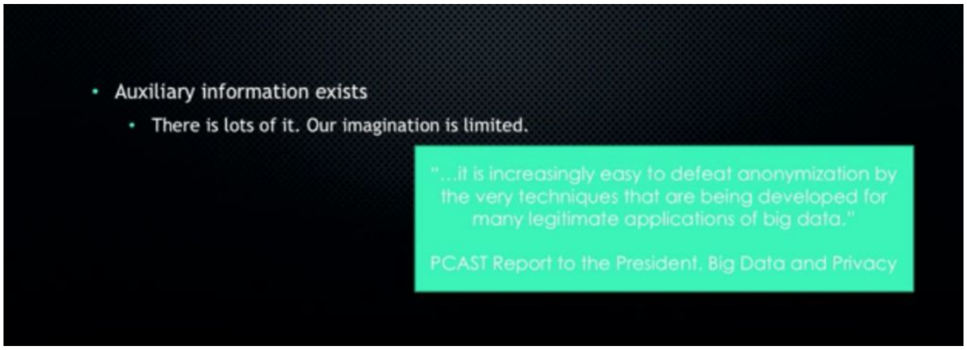
ZIP
Birth date
Sex

Name
Address
Date registered
Party affiliation
Date last voted

Medical Dataset
Voter List

Kunal Talwar

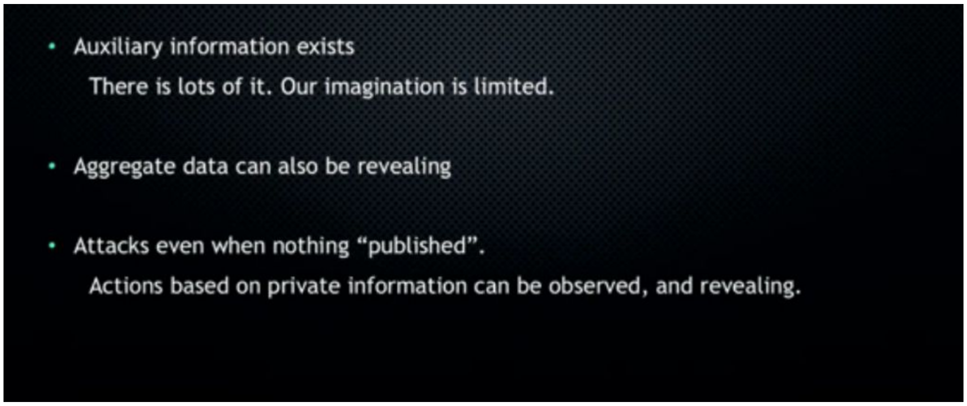
The same was done when netflix published out the movie database for rating having the goal to create a movies recommender for its user. That also gave unwanted data even they removed all the personal information from the matrix by the help of some auxiliary information.



- Auxiliary information exists
  - There is lots of it. Our imagination is limited.

"…it is increasingly easy to defeat anonymization by the very techniques that are being developed for many legitimate applications of big data."

PCAST Report to the President. Big Data and Privacy



- Auxiliary information exists

  There is lots of it. Our imagination is limited.

- Aggregate data can also be revealing

- Attacks even when nothing "published".

  Actions based on private information can be observed, and revealing.

SOME PROPOSED MODEL TO SOLVE THIS PRIVACY PROBLEM

We will answer to only large amount of data as a collection. But there is a loophole we can ask 2 large data and then subtract it to find particular data of an individual.

We will add random noise to the actual answer but this also fails why?

We can ask the same question again and again and the average the data which will theoretically average out the noise and give us the actual answer.

We will check whether the same question is asked as query and will always generate same answer for that query. This also fails?

Because we don’t know what is a same query if the query language is really rich then it is an np-hard problem to find out whether 2 queries are same or not.

Some insight to data leakage

We use machine learning to extract data from machine learning model itself. It is like using the same weapon to use against another weapon.

So we will train a neural network to solve the membership problem of a set. We will infer from a model that whether a member is trained with this particular data or not. First we will use our auxiliary knowledge to train a neural network to infer the membership problem **after solving this we will try to make an oracle which will tell us what are the missing attribute of that particular tuple**.

How can we do that?

Yes by brute force way we can do this by guessing all possible subset of the missing value and then ask the membership oracle to infer that whether this data is present or not in the model.

So by knowing a little about one person's data we can infer all other data from a model. Which is really a matter of concern.

THE NEED OF DIFFERENTIAL PRIVACY

If we carefully see this we can understand that these two things tells exactly same thing about a model. Here we don't have a tradeoff for privacy and a good machine learning model. We do have tradeoff between utility and privacy but here if we make model which don't overfit the data then we can achieve both privacy and high predictive power for machine learning model.

Overfitting is the enemy of both privacy and high predictive power of model. If the model don't learn about individual data then we never would be able to get information about private data in other word we won't have any privacy breach.

How do we do that?

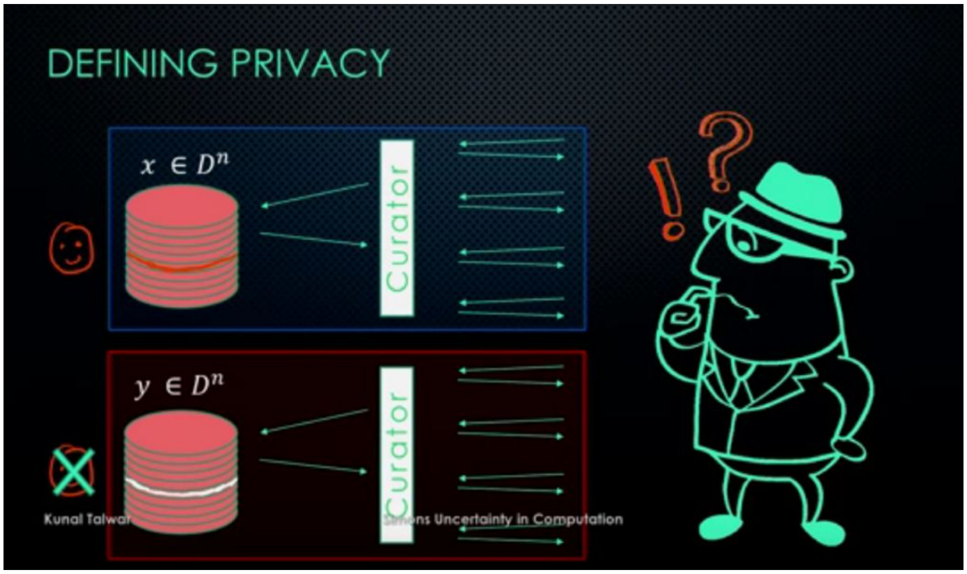
The answer is differential privacy.

If we give such guarantee that if one individual is not worsen off more than a tiny quantity by participating in the survey of the statistical model then our privacy goal can be achieved.

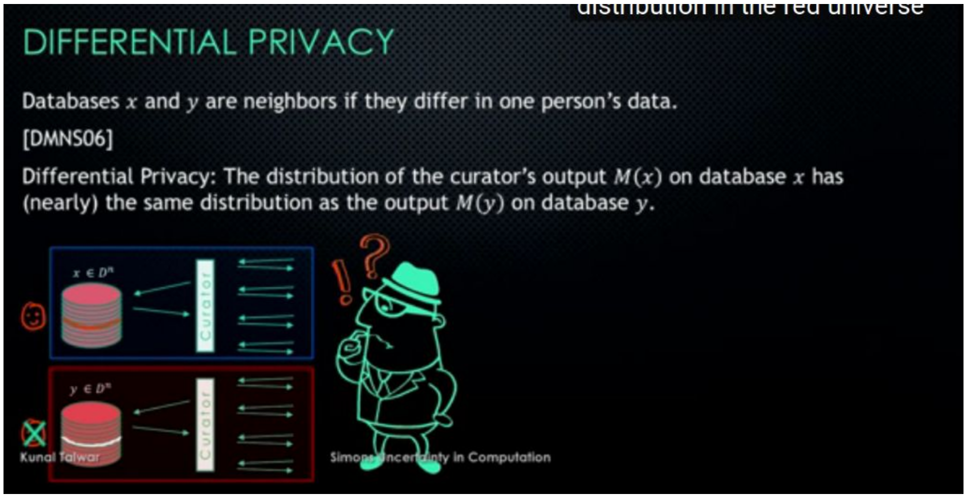
FORMAL DEFINITION OF DIFFERENTIAL PRIVACY

Privacy is something the way you achieved not what you publish but how you publish. For example take a cypher text it is important how you generated the cypher not what is the cypher.

Differential privacy is something where there are 2 universe. In one of the universe one the data is missing but the adversary is not able to identify in which universe he is.



DEFINING PRIVACY



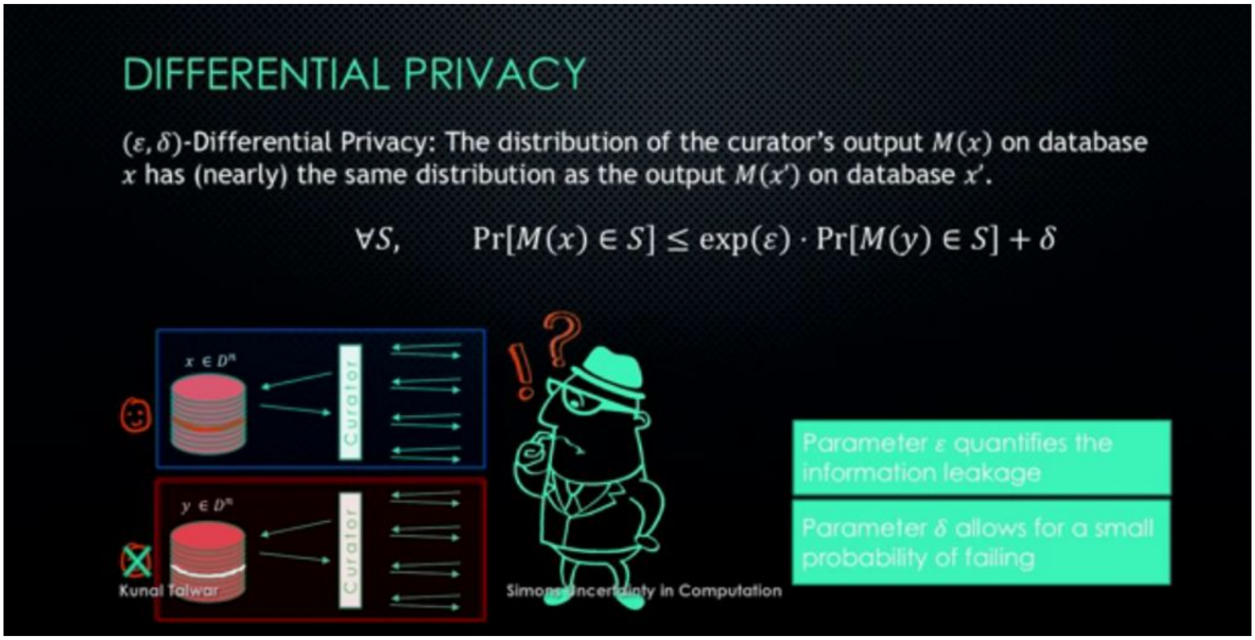
distribution in the red universe

DIFFERENTIAL PRIVACY

Databases x and y are neighbors if they differ in one person's data.

[DMNS06]

Differential Privacy: The distribution of the curator's output $M(x)$ on database x has (nearly) the same distribution as the output $M(y)$ on database y.



DIFFERENTIAL PRIVACY

(ε, δ)-Differential Privacy: The distribution of the curator's output $M(x)$ on database x has (nearly) the same distribution as the output $M(x')$ on database x'.

$$\forall S, \qquad \Pr[M(x) \in S] \leq \exp(\varepsilon) \cdot \Pr[M(y) \in S] + \delta$$

Parameter ε quantifies the information leakage

Parameter δ allows for a small probability of failing

We can define epsilon-differential privacy through the following game:

- A learner implements a summary statistic called A().
- A (notional) adversary proposes two data sets S and S’ that differ by only one row or example, and a test set Q.
- A() is called *epsilon-differentially private* iff:
- | log(Prob[A(S) in Q] / Prob[A(S’) in Q]) | ≤ epsilon
- for all of the adversary’s choices of S, S’ and Q. The probabilities are defined over coin flips in the implementation of A(), not over the data or the adversary’s choices.

The adversary’s goal is to use A() to tell between S and S’, representing a failure of privacy. The learner wants to extract useful statistics from S and S’ without violating privacy. Identifying a unique row (the one which changed markings) violates privacy. If the adversary can tell which set (S or S’) the learner is working on by the value of A(), then privacy is violated.

Notice S and S’ are “data sets” in the machine learning sense (collections of rows carrying information). Q is a set in the mathematical sense: a subset of the possible values that A() can return.

UNDERSTANDING GLOBAL SENSITIVITY AND ROLE OF NOISE IN ACHIEVING DIFFERENTIAL PRIVACY

The adversary has chosen two sets S and S’ of size *n* = 100

- S = {0,0,0,...,0} (100 zeros)
- S’ = {1,0,0,...,0} (1 one and 99 zeroes)

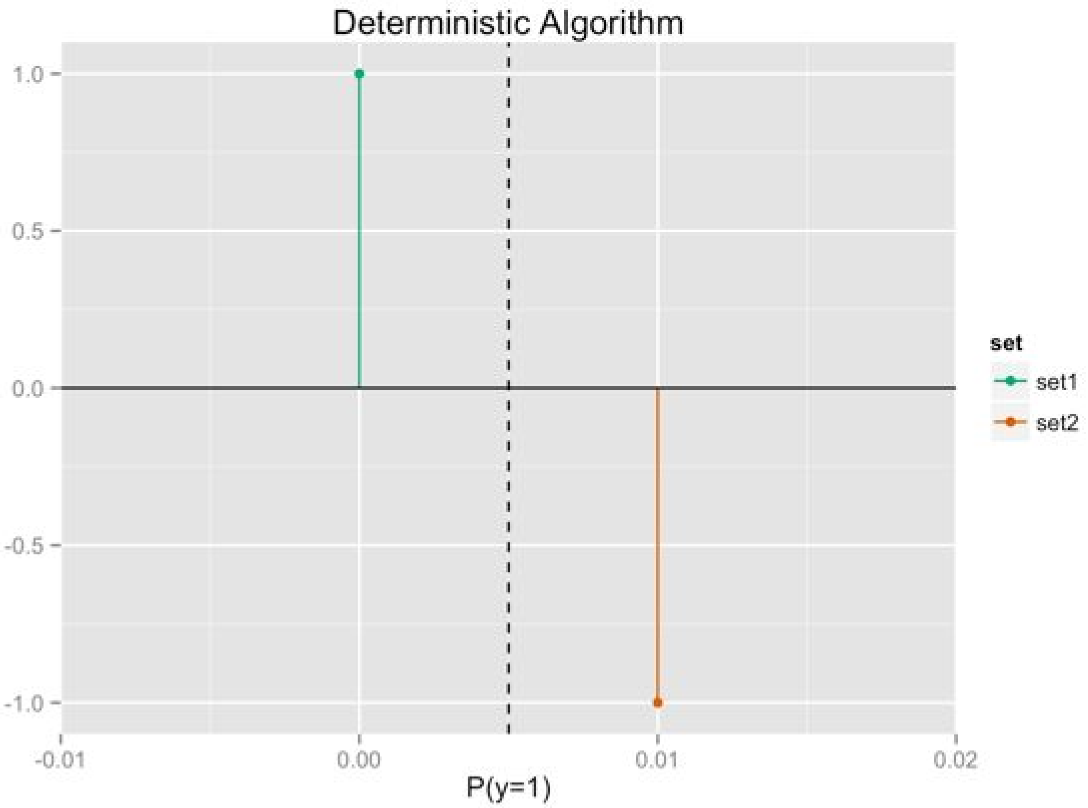
The adversary’s goal is to pick threshold T such that when he sees that A(*s*) ≥ T, he knows that A() has just evaluated S’. The learner has two (competing) goals:

- To pick an algorithm A() such that A(S) and A(S’) are so “close” that the adversary can’t pick a reliable T, to preserve differential privacy. “Close” is defined by epsilon.
- To have A() be a good estimate of the expectation, for performing useful analysis.

If we deterministically find the mean then the adversary can always find out which is s and which is s’.

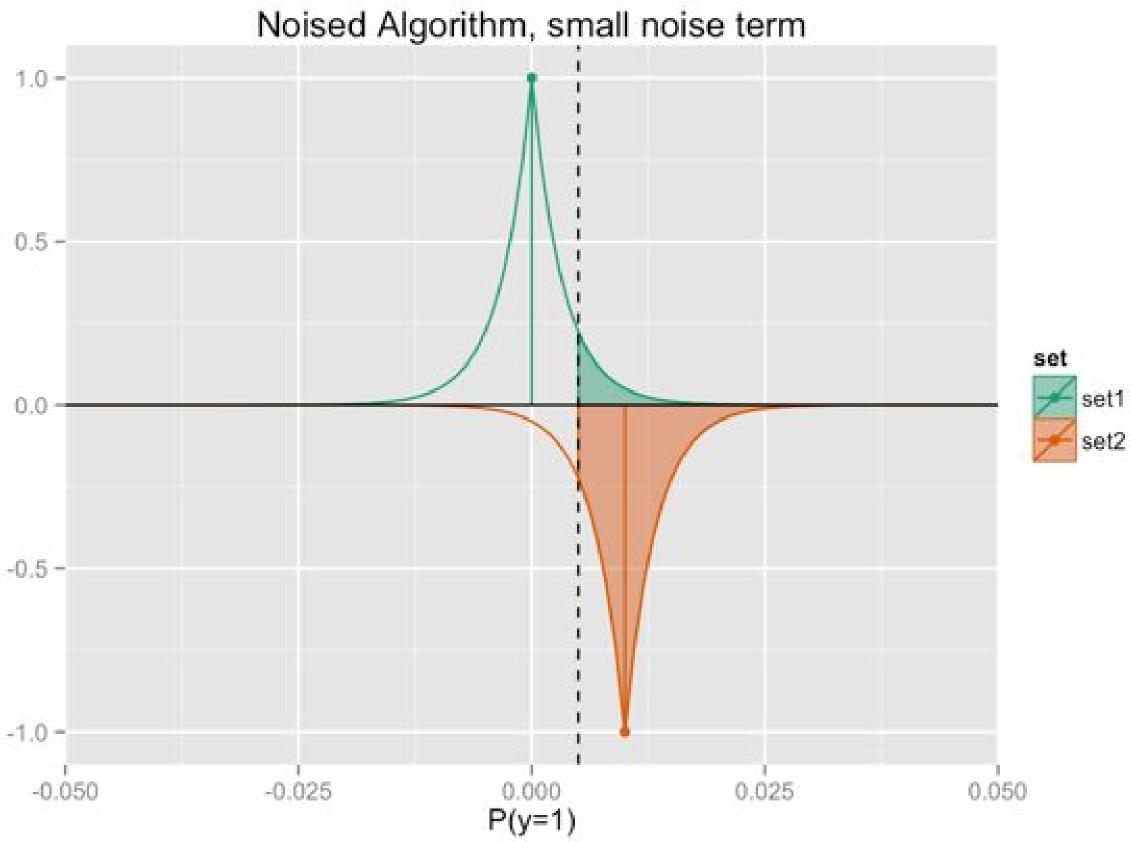
It always returns A(S) = 0 when evaluating S, and A(S’) = 0.01 when evaluating S’. This is clearly not differentially private for any value of epsilon. If the adversary picks T = 1/200 = 0.005, then he can reliably identify when A() has evaluated the set S’, every time they play a round of the game.

So here we can say global sensitivity is 0.01 because that is the maximum difference one dataset can change by being in the dataset or not.(note that we are taking binary values in the dataset either one or zero)



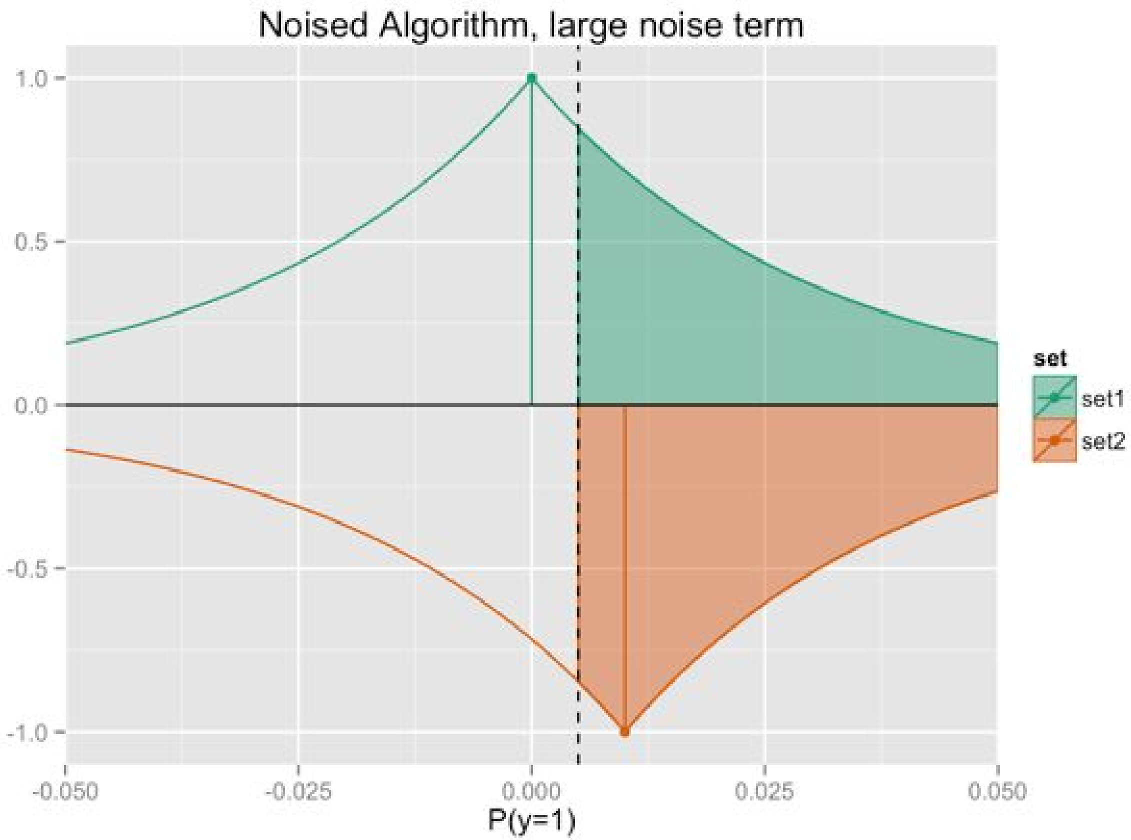
Adding Noise to give Privacy

One way for the learner to obscure which set she is evaluating is to add a little random noise to the estimate, to “blur” it. Following Dwork et.al.’s methods, we’ll add noise from a Laplace distribution, which is symmetric and falls off slower than gaussian noise would. Here we show adding just a little bit of noise (of “width” sigma = 1/3n):



The shaded green region represents the chance that A(S) will return a value greater than the adversary's threshold T — in other words, the probability that the adversary will mistake S for S' in a round of the game. We've now made that probability non-zero, but it's still much more likely that if A(*s*) > T, then *s* = S'. We need more noise.

In particular, we need the noise to be bigger than the gap A(S')-A(S) = 1/n, or 0.01. Let's pick sigma = 3/n = 0.03:



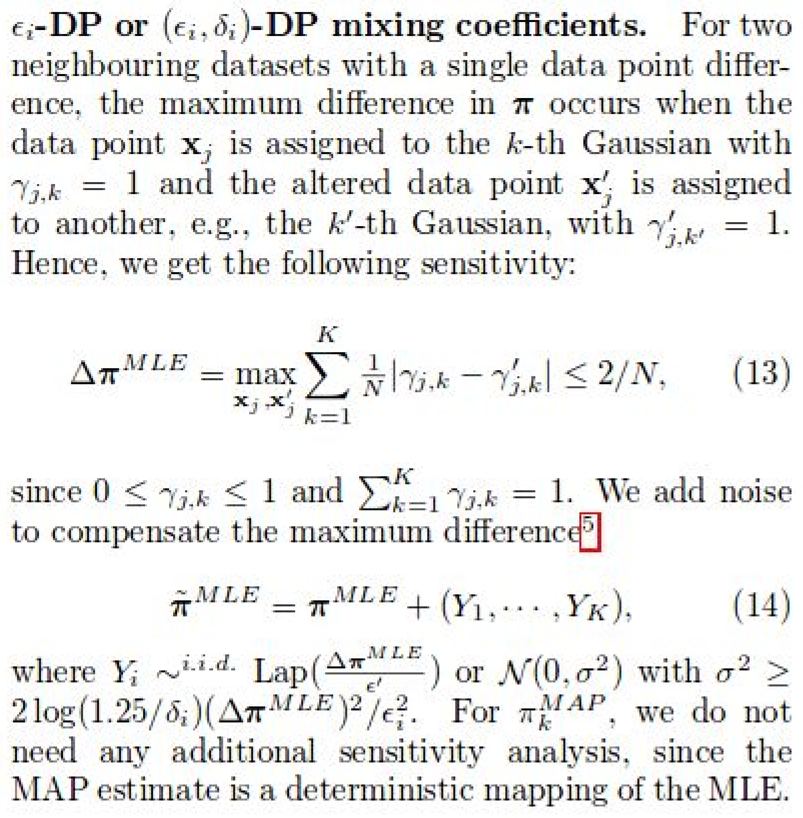
Now we see that the shaded green region has almost the same area as the shaded orange region — you can think of epsilon as expressing the difference between the shaded green region and the shaded orange region. In fact, the absolute value of the logratio of the two areas is epsilon. In other words, observing that A(*s*) > T is no longer strong evidence that *s* = S', and we have achieved differential privacy, to the degree epsilon.

ALGORITHM TO PRESERVE PRIVACY IN GMM-EM

For solving this problem we need to find out the global sensitivity in finding the parameters of the model. The parameters are (pi,mu,sigma).

The maximum amount of change in 'pi' due to absence of one data set in GMM-EM can at max change the value by 2/n. How ? previously the data was assigned to one cluster but due to removal of a point now it got assigned to another cluster so the max is 2/n.

So we will add a laplacian noise of parameter ((2/n)/eps)



ϵ_i-DP or (ϵ_i, δ_i)-DP mixing coefficients. For two neighbouring datasets with a single data point difference, the maximum difference in $\boldsymbol{\pi}$ occurs when the data point $\mathbf{x}_j$ is assigned to the k-th Gaussian with $\gamma_{j,k} = 1$ and the altered data point $\mathbf{x}'_j$ is assigned to another, e.g., the k'-th Gaussian, with $\gamma'_{j,k'} = 1$. Hence, we get the following sensitivity:

$$\Delta\boldsymbol{\pi}^{MLE} = \max_{\mathbf{x}_j, \mathbf{x}'_j} \sum_{k=1}^{K} \tfrac{1}{N} |\gamma_{j,k} - \gamma'_{j,k}| \leq 2/N, \qquad (13)$$

since $0 \leq \gamma_{j,k} \leq 1$ and $\sum_{k=1}^{K} \gamma_{j,k} = 1$. We add noise to compensate the maximum difference[5]

$$\tilde{\boldsymbol{\pi}}^{MLE} = \boldsymbol{\pi}^{MLE} + (Y_1, \cdots, Y_K), \qquad (14)$$

where $Y_i \sim^{i.i.d.} \text{Lap}(\frac{\Delta\boldsymbol{\pi}^{MLE}}{\epsilon'})$ or $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 \geq 2\log(1.25/\delta_i)(\Delta\boldsymbol{\pi}^{MLE})^2/\epsilon_i^2$. For π_k^{MAP}, we do not need any additional sensitivity analysis, since the MAP estimate is a deterministic mapping of the MLE.

Likewise for mean and covariance we will follow the traditional method prescribed in the paper.

ϵ_i-DP or (ϵ_i, δ_i)-DP mean parameters. Using the noised-up $\tilde{N}_k$ obtained from the noised-up mixing coefficients, i.e., $\tilde{N}_k = N\tilde{\pi}_k$, the maximum difference in mean parameters due to one datapoint's difference is

$$\begin{aligned}\Delta_1 \boldsymbol{\mu}_k^{MLE} &= \max_{\mathbf{x}_j, \mathbf{x}'_j} \frac{1}{\tilde{N}_k} \left|(A_k + \gamma_{j,k}\mathbf{x}_j) - (A_k + \gamma'_{j,k}\mathbf{x}'_j)\right|_1, \\ &\leq 2\sqrt{d}/\tilde{N}_k, \qquad (15)\end{aligned}$$

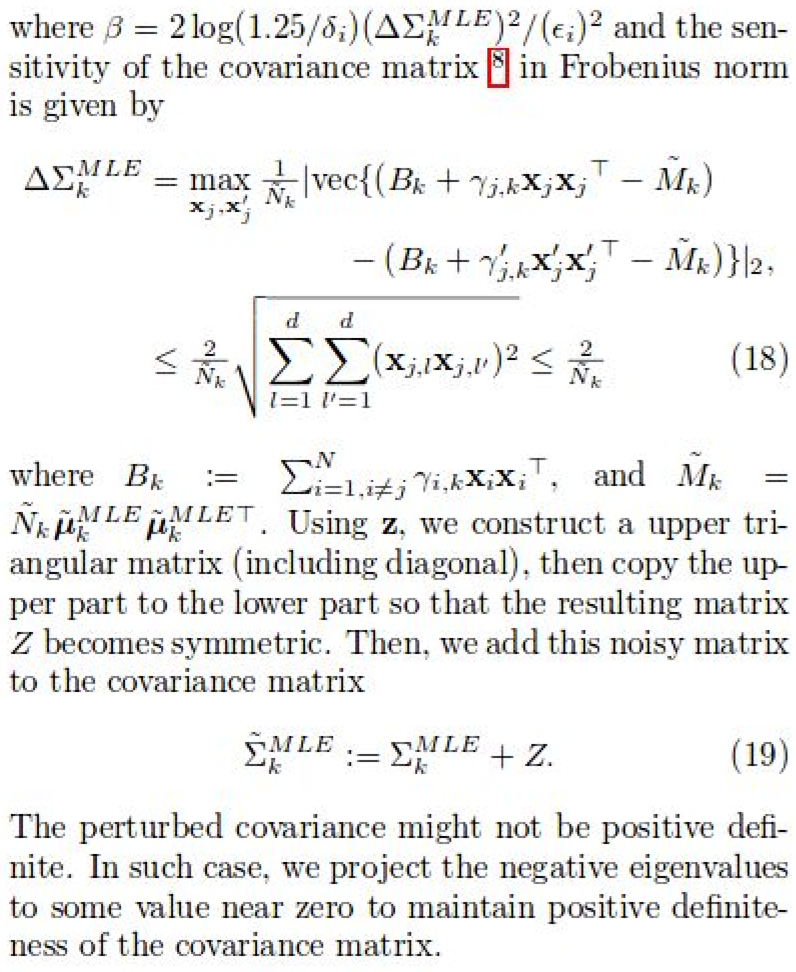
where $A_k := \sum_{i=1, i\neq j}^{N} \gamma_{i,k}\mathbf{x}_i$ and the L1 term is bounded by Eq (10). The $\sqrt{d}$ term is from the fact that each input vector is L2-norm bounded by 1.[6] We add noise to the MLE via[7]

$$\tilde{\boldsymbol{\mu}}_k^{MLE} = \boldsymbol{\mu}_k^{MLE} + (Y_1, \cdots, Y_d), \qquad (16)$$

where $Y_i \sim^{i.i.d.} \text{Lap}(\Delta_1 \boldsymbol{\mu}_k^{MLE}/\epsilon')$ or $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 \geq 2\log(1.25/\delta_i)(\Delta_2 \boldsymbol{\mu}_k^{MLE})^2/\epsilon_i^2$, where $\Delta_2 \boldsymbol{\mu}_k^{MLE} = 2/\tilde{N}_k$.

(ϵ_i, δ_i)-DP covariance parameters. For covariance perturbation, we follow the Analyze Gauss (AG) algorithm [18], which provides (ϵ_i, δ_i)-DP. We first draw Gaussian random variables

$$\mathbf{z} \sim \mathcal{N}\left(0, \beta I_{d(d+1)/2}\right), \qquad (17)$$



where $\beta = 2\log(1.25/\delta_i)(\Delta\Sigma_k^{MLE})^2/(\epsilon_i)^2$ and the sensitivity of the covariance matrix [8] in Frobenius norm is given by

$$\begin{aligned}\Delta\Sigma_k^{MLE} &= \max_{\mathbf{x}_j,\mathbf{x}_j'} \frac{1}{\tilde{N}_k}|\text{vec}\{(B_k + \gamma_{j,k}\mathbf{x}_j\mathbf{x}_j{}^\top - \tilde{M}_k) \\ &\qquad - (B_k + \gamma'_{j,k}\mathbf{x}'_j\mathbf{x}'_j{}^\top - \tilde{M}_k)\}|_2, \\ &\leq \frac{2}{\tilde{N}_k}\sqrt{\sum_{l=1}^{d}\sum_{l'=1}^{d}(\mathbf{x}_{j,l}\mathbf{x}_{j,l'})^2} \leq \frac{2}{\tilde{N}_k} \qquad (18)\end{aligned}$$

where $B_k := \sum_{i=1,i\neq j}^{N}\gamma_{i,k}\mathbf{x}_i\mathbf{x}_i{}^\top$, and $\tilde{M}_k = \tilde{N}_k\tilde{\boldsymbol{\mu}}_k^{MLE}\tilde{\boldsymbol{\mu}}_k^{MLE\top}$. Using $\mathbf{z}$, we construct a upper triangular matrix (including diagonal), then copy the upper part to the lower part so that the resulting matrix Z becomes symmetric. Then, we add this noisy matrix to the covariance matrix

$$\tilde{\Sigma}_k^{MLE} := \Sigma_k^{MLE} + Z. \qquad (19)$$

The perturbed covariance might not be positive definite. In such case, we project the negative eigenvalues to some value near zero to maintain positive definiteness of the covariance matrix.

We will try to implement both adding the laplacian noise and the gaussian noise and compare the utility accuracy.

May be we will try to see if we add noise to input data how that will change the accuracy and privacy.

The dataset to be experimented

Gowalla dataset

contains the social network users’ check-in locations in terms of longitude and latitude we will try to cluster the dataset.

More dataset yet to be decided