User Study for Differential Privacy

Sai Prahladh Padmanabhan
Carnegie Mellon University
saiprahp@andrew.cmu.edu

Sara B. Schwarz Iglesias
Carnegie Mellon University
sschwarz@andrew.cmu.edu

Annette Stawsky
Carnegie Mellon University
astawsky@andrew.cmu.edu

Abstract

Differential Privacy is implemented by companies without user input. This study aims at quantifying which values of epsilon are effective from a user perspective. We conduct a user study where we process participant data using a specific epsilon and local differential privacy. We then analyzed the users' answers for whether they felt that their privacy was preserved. We also measure the utility of using different epsilons, and identify an epsilon range that balances privacy and utility. We found that minimum utility (50 percent accuracy or greater) was achieved between epsilons 2.02 and 5.49. There were similar comfort levels for these two epsilons regarding the user's biggest fear, but there was a slightly greater comfort level for 2.02 when dealing with more sensitive information. Since most cases where differential privacy is used in practice deal with very sensitive information, our study shows that users would be comfortable with an epsilon around 2.02.

1 Introduction

1.1 Problem Statement

There has been an extensive amount research in the area of user data privacy and the methods which help us to increase data privacy. One common approach to make machine learning models private is to use differential privacy. Differential privacy adds some amount of noise to the data such that the distribution of neighboring datasets is similar to the original dataset. The amount of noise depends on a parameter epsilon which the implementer selects. The privacy of the user is inversely proportional to epsilon. When epsilon equals 0 there is perfect privacy and neighboring datasets are indistinguishable from the original dataset. Conversely when epsilon is infinity, there are no privacy guarantees and the result is equivalent to training a model normally without noise. There are two classes of differential privacy, global and local. In global differential privacy, all of the data is sent as is to a central aggregator which is where the noise is added. On the other hand, local differential privacy adds noise at user input level. The drawback of this approach is that the total noise is much larger. This is mitigated by using high values of epsilon in practice. The problem under consideration is to observe the trend of privacy guarantees provided by a local differential privacy mechanism through a user survey. The degree of privacy here is measured through user feedback in the form of a comfort scale. Another aspect of this user survey was to observe the utility v/s privacy trade-off trend based off the results of the user survey, in order to identify what amount of noise could strike a balance between utility and privacy guarantees.

1.2 Motivation

Most companies these days implement differential privacy as a default mechanism for privacy guarantees for warranting the collection and usage of user data. Companies have an incentive to use an epsilon that does not compromise the utility of the data, but at the same time provides privacy guarantees. Unfortunately there is a trade-off between privacy and utility when selecting epsilon. Companies have a history of using differential privacy for marketing while selecting an epsilon that's

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

too large to provide meaningful privacy guarantees. The main idea behind selecting local differential privacy is to have the mechanism be independent of 'trust' unlike the case in global differential privacy. This assures stronger privacy guarantees for the local differential privacy setting. Through this experiment, we wanted to observe our privacy guarantees and the utility guarantees for varying values of epsilon and identify the epsilon best suited for practical purposes.

1.3 Background

Lit Review We have looked into many published papers that discuss approaches for choosing epsilon and these were the three most influential to our research. Laud et al [3]. discuss the ideal choice of epsilon in terms of guessing advantage, as there is no common agreement on a sufficiently small epsilon. Using the guessing advantage, upper and lower bounds are calculated for the epsilon value and the problem of a very large space is solved by constraining the search space. J.Hsu et. al [1] talk about an economic method to choose epsilon which does not compromise the utility too much and still maintains the privacy guarantees. They establish a lower bound on epsilon for data universe X and probability of publishing a private record p* as epsilon >= ln(p*|Xl), and similarly set an upper bound as 1/N where N is the number of different rows in the neighboring database. Friedman et al. [2] also talk about the privacy utility trade off and show how to achieve middle ground.This paper focuses on decision tree induction as a sample application while considering the privacy and algorithmic requirements of the same. The authors have considered a global differential private mechanism whereby the aggregator adds some noise to each query of data. The adopted methodology falls under the umbrella of Pareto improvements. Our original framework was a decision tree like the last paper, but we eventually leaned more towards the first two papers and used their bounds for epsilon as a starting point. Our study builds on these findings because we are approaching it from a user perspective instead.

2 Approach

Our study consists of a game similar to 20 questions that predicts whether the participants' worst fear falls in one of 4 categories. The categories of fears are: people related (e.g. clowns or abandonment), animal related (e.g. spiders), object related (e.g. the ocean), and existential (e.g. death). For example, if the participants worst fear is being buried alive, our game would predict the existential category. The game is be hosted on a website and have only about 5 questions so that we can balance adding noise with utility.

2.1 High-Level Procedure

The participants of the study are asked to think of their worst fear and answer questions that we ask about their fear. We selected fears as a way of mimicking sensitive information without directly asking them personally identifiable information. We have chosen the 4 categories of fears based on a dataset of the 100 most common fears. All of the questions are yes/no questions. We select one epsilon per participant, and they always answer honestly. After they have completed the survey, we add noise to all of their answers depending on epsilon. We then use the noisy answers to predict what category their fear falls under. Once we make this prediction we ask the user whether they feel their privacy was preserved. We prompt the user to consider how comfortable they would be with this level of privacy being used with more sensitive information.

We designed the survey so that we are honest with the users from the beginning about the purpose of the study. We acknowledge the concern that alerting users to privacy issues would bias the participants to care more about privacy. However, it was not feasible to use deception and avoid collecting personal information for this study. Moreover, we do not believe that it is a negative thing to alert the users to privacy concerns in this case since differential privacy is implemented without their consent anyways. If we were measuring an action they might take, then the results would be biased, but we are measuring preferences.

We centered many design decisions around being able to make predictions in real time when the user finished the survey. We had considered training a model separately, testing it with the participants' answers, and then following up with the participant after they completed the survey. However, we were concerned that participants would lose interest if we had to follow up and not engage with the

second part of the study. Therefore, we implement our prediction algorithm on the website. The prediction is made based on what combination of questions were answered yes/no.

2.2 Procedure Details

We use a webserver and database through AWS where we host our game. We ask 5 questions in order to classify the participants' fears:

1. Would you expect someone to dress up as your fear for Halloween?
2. Does your fear become more likely/prominent as you age?
3. Would you expect to encounter your fear on vacation?
4. Is the movie 'Get Out' scarier than the movie 'Jaws' for you?
5. Would you expect to encounter your fear at a wedding?

The reasoning behind these questions is that we can make an equal number of predictions for each category from all of the possible answers. This means that they are good at identifying each category of fear in roughly the same way. For example, answering 'yes' to questions 1 indicates that the fear is probably a person and maybe an animal, whereas answering 'yes' to question 5 means that the fear is certainly not an animal.

We predict the each of the categories based on the following answer combinations (ex: 10000 means the answer to the first question was 'yes' and the answers to the remaining questions were all 'no')

Category Prediction

People : 10011, 11111, 11011,10111, 10010, 11110, 11010, 10110

Animals : 10000, 10100, 00000, 00100, 11000, 11100, 00110, 00010

Objects : 01011, 01111 01010, 01110, 01001, 01101, 01000, 01100

Existential : 00101, 00111, 10001, 10101, 11001, 11101, 00001, 00011

We use 5 values of epsilon which map to 5 different probabilities that we keep the participants' honest answer. We have a control group where we keep the participants' answer 100% of the time (corresponding to epsilon equals infinity). In the other 4 groups, we keep the participants' honest answer 90%, 75%, 60% and 51% of the time. These map to the following epsilons respectively: 2.197, 1.098, .405, and .04. Since we have 5 questions, our final epsilon will be 5 times the epsilon for any given question. Therefore the final epsilons we are considering are: 10.985, 5.49, 2.025, and .2 (in addition to the control group where epsilon is infinity).

Epsilon becomes 0 when we keep their values 50% of the time because that is no better than random which is completely private. Therefore, our reasoning was to have one value close to perfect privacy (epsilon = .04) and the others centered around 75% which is the midpoint between 50 and 100. We had 6 people in each of the 5 epsilons.

There are two post-game survey questions:

1. The more noise we add to your answers, the more your true fear is changed, and the more your privacy is protected. Imagine a company was predicting your fear and used the amount of noise that we used to make the above (noisy) prediction. Knowing your true fear category, how comfortable would you be that your privacy is being protected if companies thought your fear was [insert]?
2. Now imagine that companies were predicting your financial information using the same amount of noise that we used to predict your fear. How comfortable are you that your privacy is being protected now?

As explained above, we chose to inform the user about differential privacy and the purpose of the study. We informed them in the following way:

"**What this study is all about:**

Hi user, thanks for agreeing to participate in this little experiment. The key idea is to see how comfortable YOU are with a concept called differential privacy. You answer 5 questions honestly, we

add noise (change up the answers) to some percent of your answers, and only then make a prediction. The goal of local differential privacy is to limit the amount of sensitive information that the algorithm sees (here, your answers are the sensitive information that we are trying to protect). Don't be surprised if the experiment returns an answer diametrically opposite to what your actual answer was! This is an effect of protecting your privacy, and we will use different degrees of protection for different users. When we make a prediction, we want you to answer honestly how much you feel your privacy was protected. Having an accurate analysis for when differential privacy is helpful for the users can guide industry to implementing it in a useful way (and not just as a marketing ploy)."

2.2.1 Code

The analysis code used to generate all the following graphs can be found in this python notebook [5].

The noise adding mechanism is included below:

```
<?php
  function add_noise($var, $prob)
  {
    $x = rand(0,100);
    if ($x < $prob)
    {
        return $var;
    }
    else {
      if ($var == "yes")
      {
        return "no";
      }
      else
      {
        return"yes";
      }
    }
  }
  $answer1 = add_noise($row["question_1"], $epsilon);
  if ($answer1 == "yes")
  {
    $y = 1;
    $x = 4;
    $l = $y << $x;
    $comp = $comp | $l;
  }
 ?>
```

3 Results

As mentioned previously, we surveyed 30 people and asked them sensitive questions to predict a sensitive attribute and then questioned them on their comfort level due to the mechanism.

For the evaluation of our project we are looking specifically at two aspects:

- User's comfort level in regards to the epsilon value used by the mechanism
- Mechanism utility based on the the number of noiseless questions retained in regards to the epsilon value being used.

For the first evaluation we are expecting to see a decrease in comfort level as the epsilon value increases. As mentioned previously, the epsilon value's increase implies less privacy because there is less probability of adding noise to the answer. With less privacy it is expected there to be less comfort.

For the second evaluation, in order for our mechanism to maintain good utilization it should retain more than half of the true answers. We calculated for all users the number of answers flipped per

epsilon. It is expected that as the epsilon decreases more answers are flipped and the mechanism loses utilization.

3.1 Findings

Comfort Level For our first evaluation, we had 6 people per each epsilon answer questions of their comfort level due to the mechanism prediction accuracy.

> **First Question** As you can see in Figure 1, we found that for the first question the epsilon values equal to and larger than 2.025 had decreasing comfort level. We noticed a rare result where the epsilon value of 2, the value with most privacy, had the least comfort level out of all. There are many possible reasons for this that we explain in the Discussions section. For the first question it had an overall decrease of comfort level which is what was to be expected.
>
> **Second Question** Similar to the first question, we were expecting the comfort level to decrease as the epsilon increased. Looking at Figure 1, we can see this interaction more or less. Between epsilon values 2.025, 5.49 and infinity the comfort level decreases. We noticed both epsilon values, 2 and 10.985 to give odd results. Epsilon value 2 had a low comfort level where it should be high and epsilon value 10.985 had a high comfort level where it should be low. Again, there are many reasons for this which are mentioned in the Discussions section.

Mechanism Utilization For our second evaluation, we had 30 answers per epsilon and computed the ratio of how many true answers were flipped. As expected more true answers were retained as the epsilon increased. Looking at Figure 4, we can see that the graph makes an "S" curve seeing as how there is a larger difference between epsilon 2.025 and epsilon 5.49. The epsilon values equal to 2.025 and larger had an average of retaining more than 2 of the answers.

3.2 Plots

Figure 1 shows both averages of the answers of both questions for each epsilon. As mentioned previously, it was expected to see a decrease, which can be seen more or less.

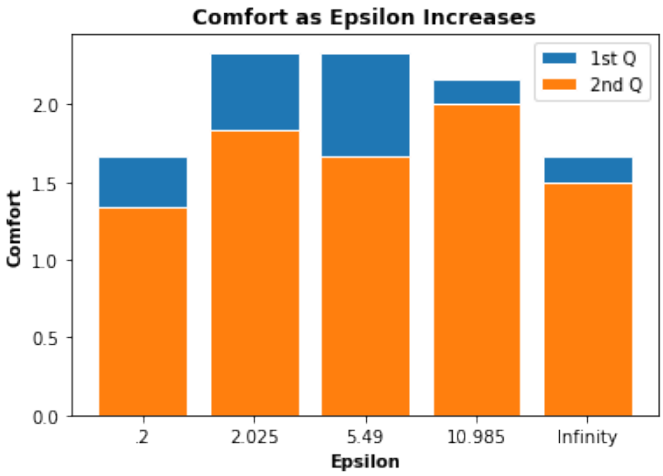

Figure 1: Average answers to the Comfort questionnaire. Blue: Question 1, Orange: Question 2

In order to comprehend, we generated a jitter graph, where the y axis organized the values by density based on the answers and simply by comfort level order. Each answer is plotted in both graphs spread among the specific epsilon values. The x axis shows the values of epsilon on a percentile scale.

For Figure 2 we note that the answer for 3, meaning the most comfort level, has the highest density implying it was the most answered value. But even more interesting is that the following value 1, the

least comfort level, was next in line in highest density. It is very interesting to note how our users were either very comfortable or not at all.

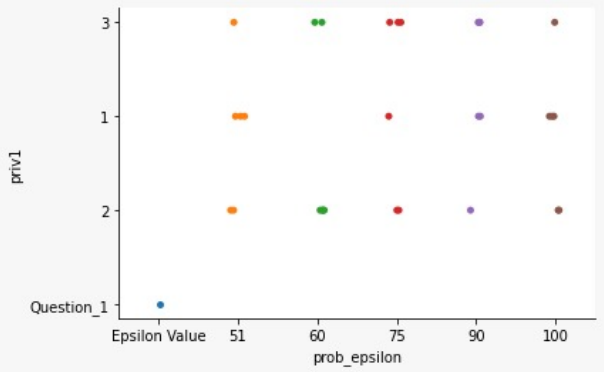

Figure 2: All answers to the first question regarding user Comfort level.

For Figure 3 we note that the answer for 2, meaning the medium between high and low comfort level, has the highest density implying it was the most answered value. For this question at least, being the one concerning finances, it seemed users were more insecure in whether to be comforted or not.

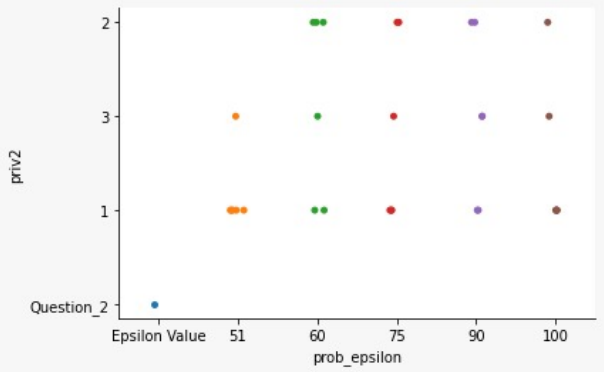

Figure 3: All answers to the second question regarding user Comfort level.

Figure 4 shows the average of flipped answers from all users per epsilon.

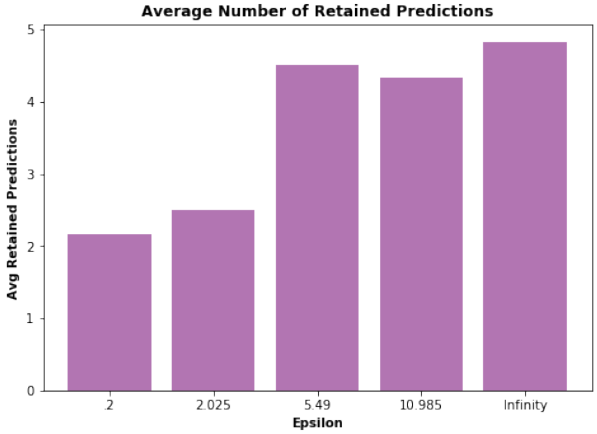

Figure 4: Average of flipped answers per epsilon

3.3 Intepretation

An important point to note, looking at Figure 2 and Figure 3, is the range of answers among all comfort levels for each epsilon. Since we only had 6 users, there is higher chance of overhead in our data.

For the first question on comfort level we saw expected results except for the oddity for epsilon value 2 in which almost all responses were that the users did not feel confident.

Part of our survey was to also ask the users for what their real prediction was, in order to evaluate any relation between the mechanism's lack of accuracy and the users' answers. For epsilon value 2, the mechanism only erred 2 out of the 6 users. This is not high enough to explain why most of the users had low comfort level. In the discussion we mention other possibilities out of our control that could have caused this.

For the second question, we were expecting people to be more cautious of their financial situation. This was the case since a lot less people responded to being very comfortable with the mechanism then when asked about having their fear known. But there was still two epsilon values that had odd comfort levels. As mentioned above, the mechanism erred 2 out of 6 users, thus not a real explanation to why they still did not feel comfortable using the mechanism. For epsilon value 10.985, the mechanism erred 4 out of the 6 users. This could explain why users felt comfortable even with a high epsilon value.

The evaluation for good utility is dependent of having more than half of the answers retained, thus showing that the best epsilon values in regards to utilization are those of 2.025 and above.

Because we showed that when the epsilon is smaller the comfort level of the users in regards to their data is higher, therefore we can safely say that for our experiment epsilon value 2.025 is the best epsilon. The value retains more than half the answers and it maintains users privacy.

4 Discussion

The experiment by itself underwent 3 iterations, mainly due to the explanation of the scale of rating privacy not being lucid enough for participants to mark their experienced comfort level during the experiment with proper understanding. The trend of comfort v/s utility observed in the finalized version of the experiment conformed with the general expectation of the comfort decreasing as the epsilon value increased. The random number mechanism which we implemented, had some of the noisy predictions to be the same as the true prediction in some cases, which explains the unexpected dip in the user comfort for epsilon value of 0.2. For a general scenario whereby the user database is far larger than the one used in this experiment, the above fact would not be observed because the average would not be greatly affected by a few unexpected observations. With more users for each epsilon, we would expect to see this phenomenon generalize well and verify a sudden dip in comfort level for some range of epsilon.

Broader Impact

Teamwork

The team consisted of Sai, Sara and Annette. We all brainstormed ideas and implementation details together, and we all distributed the survey. Sara used AWS credit to create the website, and Sai and Annette did analysis on the results using python. We split the writing evenly.

References

[1] J. Hsu et al., "Differential Privacy: An Economic Method for Choosing Epsilon," 2014 IEEE 27th Computer Security Foundations Symposium, Vienna, 2014, pp. 398-410, doi: 10.1109/CSF.2014.35.

[2] Arik Friedman and Assaf Schuster. 2010. Data mining with differential privacy. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data

mining (KDD '10). Association for Computing Machinery, New York, NY, USA, 493–502. DOI:https://doi.org/10.1145/1835804.1835868

[3] Peeter Laud and Alisa Pankova, Cybernetica AS. Interpreting Epsilon of Differential Privacy in Terms of Advantage in Guessing or Approximating Sensitive Attributes . CoRR volume 1911-127777, 08 Jan 2020.

[4] FearOf.Net. Phobia List – The Ultimate List of Phobias and Fears, 2020. [online] https://www.fearof.net/

[5] Python Notebook
https://colab.research.google.com/drive/1n-pnbDA5d5iDSJG5uOMpClA8wfPv-ctT?usp=sharing