A Deeper Look into the Fairness of Diferential Privacy

Capstone Research Paper
Presented to the Faculty of the
School of Engineering and Applied Science
University of Virginia

By

Youssef Errami

May 3, 2020

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed: _____Youssef Errami_____

Approved: _____ Date _____
David Evans, Professor of Computer Science, Department of Computer Science

Capstone Research Paper: A Deeper Look into the Fairness of Differential Privacy

Youssef Errami (ye4pg)
Advisor: Professor David Evans

I. Introduction

Fairness within the field of computing is rapidly starting to become a big talking point, especially with the rise of algorithmic decision making. With that said, fairness should be an integral component within the field of privacy as well. User's personal data should remain private and secure to them, no matter who they are. When certain members are more prone to privacy breaches then others, this is where the lack of fairness shines brightest. This capstone project focuses on the fairness of differential privacy by analyzing the members of data sets who were exposed by membership inference attacks. In this paper, we will not be focusing on the math and mechanisms surrounding differential privacy, as we will be treating it more so as a black box. We will be using the experiments and code provided by Bargav Jayaraman from "Evaluating Differentially Private Machine Learning in Practice" [1] in order to find the individual members who were exposed by the attacks. The contributions being made by this paper include shedding some light on the fairness of differential privacy through conducting our own experiments. These contributions are particularly important as it is crucial to reveal consequential flaws, such as a lack of fairness, with these different privacy definitions our communities interact with.

II. Background

Differential privacy (DP) is a mathematical definition of privacy that is used throughout the privacy-preserving machine learning field. DP provides a mathematically provable guarantee of privacy protection against a wide range of privacy attacks, including membership inference attacks (MIA), which we used for this paper. Along with DP comes an important guarantee which states that "anyone seeing the result of a differentially private analysis will essentially make the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis" [2]. What we are concerned with for this paper is whether this guarantee of privacy protection is truly guaranteed for all members of a dataset all the time or whether there is an imbalance of fairness at play here.

This capstone project is a continuation of the work that Jonah Weissman and I started this past summer (Summer 2019). We both worked along side Bargav Jayaraman, a current Ph.D Candidate here at the University of Virginia, and Professor David Evans. We started working alongside Bargav by conducting experiments related to the research project he was working on, "Evaluating Differentially Private Machine Learning in Practice" [1].

Throughout conducting these experiments, we started to wonder about certain fairness aspects with respect to these experiments. Our main concern was whether certain data points had higher risks of exposure compared to other data points. This question introduced lead us to start looking into possible ways that we can prove or disprove the possible correlation. The experiments we chose for this capstone project were to a) check and see if all exposed members shared certain characteristics and b) check and see if exposed members are within closer proximity of each other compared to members who weren't classified as being exposed.

III. Experiments and Results

For the two experiments I ran, I used the Purchase-100 data set. The Purchase-100 data set consists of 200,000 customer purchase records of size 100 each (corresponding to the 100 frequently-purchased items) where the records are grouped into 100 classes based on the customers' purchase style [3]. For both data sets, we use 10,000 randomly-selected instances for training and 10,000 randomly-selected non-training instances for the test set. For our experiments, we are only concerned with the test set, as that is what I performed all of the experiments with.

A. Experiment 1

For the first experiment, our main goal was to test whether the set of exposed members had any shared characteristics about themselves that allowed for them to be exposed. With respect to the Purchase-100 data set group, we defined "having a characteristic" as having a feature. The Purchase-100 data set contained 100 features which corresponded to 100 items. Each member either purchased the item or did not purchase the item, which made determining whether or not a member had a certain characteristic quite simple. Once we gathered the exposed members data, we came out with 54 members who were exposed out of the 10,000 members from the testing set. We ran through each of the 54 exposed members features and checked to see if there were any characteristics that they all shared. We

Fig. 1. Chart of the number of total members within 6.16 unit radius of exposed members

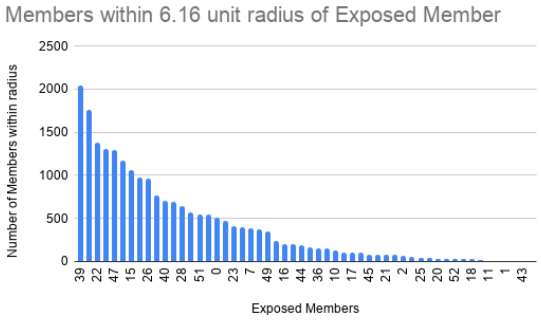
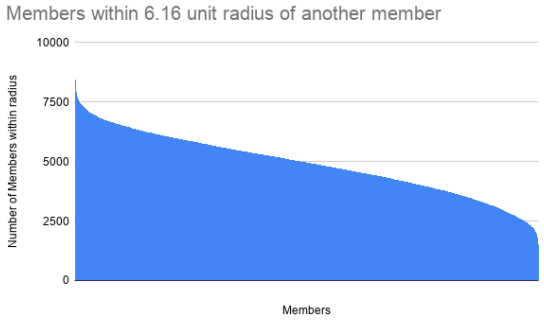


Fig. 2. Chart of the number of total members within 6.16 unit radius of any given member



would then cross check these features with the features of the non-exposed members to see if they didn't have those features. Our results showed that there were no set of characteristics that was exclusive to exposed members, meaning that there is no correspondence between a members characteristics and their exposure.

B. Experiment 2

For the second experiment, we started looking deeper into where the exposed members were in geometric space, and whether the exposed members happened to be laid out in a certain manner. We were mostly concerned with whether the exposed members were in close proximity of each other. We then came to the conclusion that the best way to determine whether this is the case would be to analyze the density patterns of exposed members. What we were looking for was whether the exposed members are closer to each other compared to non-exposed members. We determined the distance between members by calculating the euclidean distance between the points. Their "coordinates" were determined by the features (characteristics) they embodied. Therefore each point had an array of size 100 where each index represented whether they purchases the item (represented as a 1) or they did not purchase the item (represented as a 0). Once the euclidean distance between two said points is calculated, we can now determine whether those two points are "closer" or "farther" from each other relative to the other distances between other members.

Now that we had a way to calculate the distance between points, we sought out to find out the average distance between an exposed member and another exposed member. We ran the experiment and determined that the average distance from one exposed member to another is about 6.16 units. Upon completion of this experiment, it was clear that this did not really tell us anything about the region density of exposed members and whether this was the case.

C. Experiment 3

For the third experiment, we took what we learned about the average distance between exposed members and enhanced our definition of density with respect to members. We decided to define density as a circle around a given member (where the given member is at the center of the circle) and seeing how many other members are within said circle. For our experiments, the radius of said circle would be the average distance between exposed members, 6.16 units. We started by determining the amount of members within the 6.16 unit radius of a given exposed member. The results (Fig. 1.) showed that there were exposed members spread out. The number of members within the 6.16 unit-radius of an exposed members ranged from 2 to 2042 other members. To compare this data, we also graphed the number of total members within a 6.16 unit radius of any given member. We see a very similar graph shape which shows us that the densities are similar across the total number of members and exposed members.

D. Experiment 3 with different values for the radius

We also conducted Experiment 3 with different values for the radius to see how the densities changed with a larger and smaller radius. For this part of the experiment, we chose to test a 7.00 unit radius and a 5.00 unit radius. As seen in figure 3 and 4, a 7.00 unit radius is too large as it nearly includes all members and doesn't give us a meaningful density result. On the other-hand, as shown in figure 5 and 6, a 5.00 unit radius is too little as there are too few members within the 5.00 radius and there is very little meaningful data that can be used to come to a fairness conclusion.

E. Experiment 4

For the fourth and final experiment of this paper, we decided to see if there was a correlation between the the density and per-instance training loss value of members. We got the density information the same way we have been doing so for the past 3 experiments. The per-instance loss data was saved from the machine learning experiments that were conducted using code from Bargav Jayaraman's experiments [1]. What we found from this experiment was that there was some slight correlation between the two fields as the density

Fig. 3. Chart of the number of total members within 7.00 unit radius of exposed members

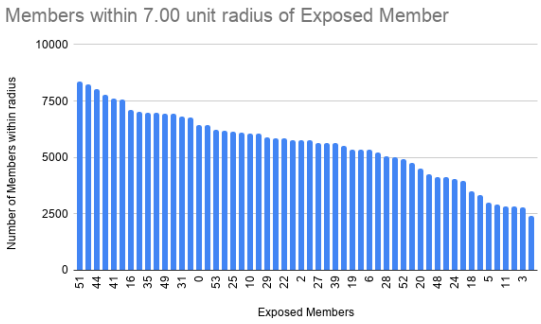


Fig. 4. Chart of the number of total members within 7.00 unit radius of any given member

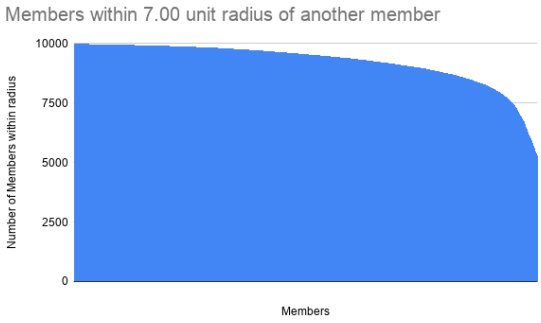
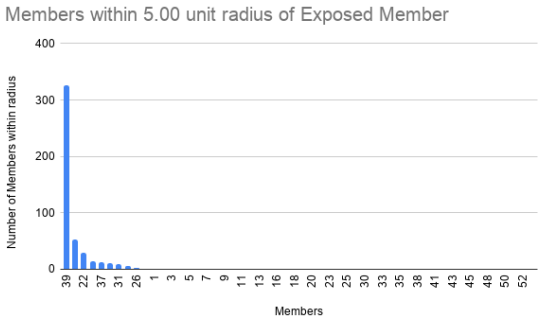


Fig. 5. Chart of the number of total members within 5.00 unit radius of exposed members



and per-instance loss both decreased across all 10000 members. This can be seen in Figure 7. An interesting continuation/follow up to this experiment would be to see if this slight correlation holds between density and per-instance loss of exposed members.

IV. Conclusions

The experiments that have been conducted and the results we found are in no way *definite* proof that there exists or doesn't exist some fairness inequality within differential privacy models but it is a step in the right direction towards figuring out the connection between fairness and differential privacy. It is important that researchers keep looking into this as fairness in privacy

Fig. 6. Chart of the number of total members within 5.00 unit radius of any given member

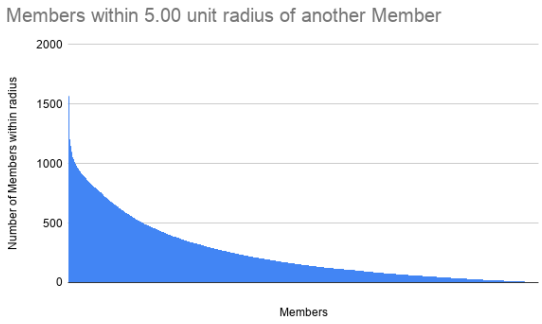
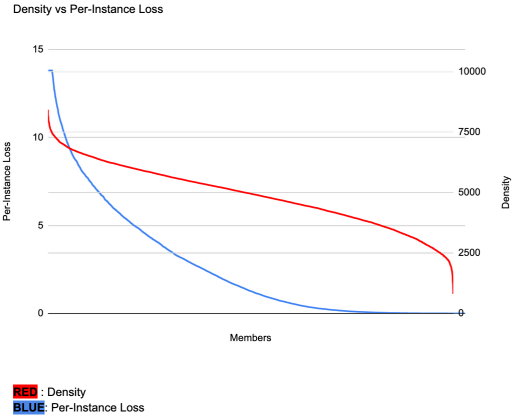


Fig. 7. Density vs Per-Instance Loss of 10000 members



is crucial. With the topic of fairness being brought up all the time in our modern-day society, it is important that our privacy standards abide by the same laws of equality and fairness we do.

References

[1] B. Jayaraman and D. Evans. (2019). [Online]. Available: https://arxiv.org/pdf/1902.08874.pdf

[2] K. Nissim. (2018). [Online]. Available: https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_new.pdf

[3] Kaggle. Acquired valued shoppers challenge. [Online]. Available: https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data