Considerations for Local Differential Privacy

Kediel Morales
Department of Computer Science and Engineering
New York University
New York, New York
km5655@nyu.edu

Abstract— Mobile technologies have improved so much that technology organizations are able to implement privacy mechanisms locally before they are collected for statistical research. Local Differential Privacy (LDP), the injection of randomized noise to ensure confidentiality through plausible deniability, is the most common method used today. This paper gives an overview of LDP differences based off of an organization's implementation. Finally, looking further into how LDP is affected by the epsilon metric and ways it can be fortified for the long term.

Keywords— *Privacy, epsilon, Local Differential Privacy, RAPPOR, DREAD*

I. Introduction

Local Differential Privacy (LDP) has been a part of Apple's ecosystem since 2016 and a research topic since 2006 when the seminal paper titled *Calibrating noise to sensitivity in private data analysis* was published. LDP builds off of differential privacy where the latter utilizes a central aggregator that houses the raw data, mathematically manipulates it, then utilizes it. This early version of differential privacy is just a data breach waiting to happen. Incorporating privacy to our everyday products is a good start to combat Orwellian situations and move forward with a free and open Internet society. Therefore technology organizations took it upon themselves to come up with new implementations for their mobile use case, among others. By keeping the collection and mathematical manipulation locally, organizations can guarantee privacy while they run statistical computations in-house.

Apple's specific strategy consists of implementing privacy as a feature, which allows them to differentiate their products from competitors. In an effort to bring awareness to the public they have mounted a substantial marketing campaign to illustrate the importance of privacy. However, in the age of big data and their respective end user leaks, trust alone will not suffice when touting the benefits of privacy and keeping users data confidential. As users progressively integrate mobile devices into their everyday lives there is only one question to ask. Will LDP provide confidentiality over the long term?

II. Related Research

A. Apple

Apple utilizes Local Differential Privacy with varying privacy budget levels (*e*) depending on the data type collected [7]. Some of their privacy budgets are set to such an exorbitant level (the iOS11 beta has been proven to use a privacy budget of 43 in some instances) that scholars have essentially deemed it ineffective[5].

B. Google

On the other side of the isle is Google's LDP solution. Their in-house Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR) algorithm assigns privacy budget levels (*e*) to each individual metric that is collected to produce their transformed dataset [8]. Assigning a budget to each individual metric before aggregation may be better than assigning a budget to a whole aggregated category but it is not guaranteed when epsilon is still used at high levels.

C. Internet Of Things

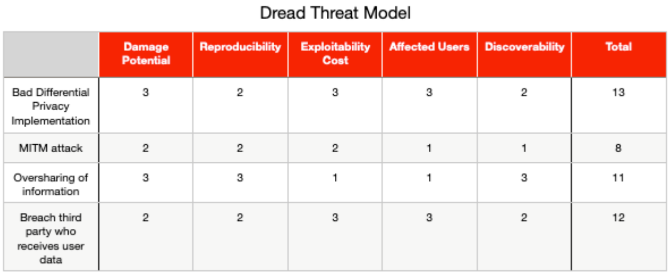
P. Kühtreiber et al distinguish Differential Privacy, Local Differential Privacy, and a Hybrid model in the context of the Internet of Things [1]. Epsilon and Privacy Budget is arguably the most important factor in the algorithms, and it is only mentioned once. When machine data is generated to be manipulated and analyzed, the important metrics should be isolated and taken into consideration the most. Let's take an in-depth view at how epsilon and privacy budget affects an algorithms ability to provide confidentially across it's lifecycle.

III. Motivating Example

First, let's assess a few cases of risk by breaking down potential attack vectors utilizing the DREAD Threat Model. As shown below in Fig. 1, each category has been labeled with a number where 1 is the lowest amount of risk and 3 is the highest. Taken as a whole all of the categories could present significant harm. However, the risk with the highest rating is a bad implementation of LDP, due to flaws from the very start of the process. Apple claims to not share confidential user data with third parties but after their encryption dispute with the FBI in 2016 it is uncertain how much data is being exposed legitimately or not. The good news is that even if a third party was breached,

XXX-X-XXXX-XXXX-X/XX/$XX.00

the data is encrypted in-transit. Moreover, it should be transformed by the time it is received outside of Apple.



Dread Threat Model

	Damage Potential	Reproducibility	Exploitability Cost	Affected Users	Discoverability	Total
Bad Differential Privacy Implementation	3	2	3	3	2	13
MITM attack	2	2	2	1	1	8
Oversharing of information	3	3	1	1	3	11
Breach third party who receives user data	2	2	3	3	2	12

Fig. 1

Now that risk has been analyzed with the DREAD model we should understand the differences in LDP implementation. Mathematically speaking the epsilon variable, which is the parameter that measures noise in a report, is the most important to note when understanding these complex privacy mechanisms. Apple sends reports once a day with an epsilon of 2 or, in iOS11's case, an upper limit epsilon of 43 [5]. Google's RAPPOR collects hundred of thousands of private hashed strings to learn which strings are most common [6]. Google uses a lower epsilon on average (2-9) [6] and does not store data in a central location. The key difference between both implementations is how they handle their "budget".

The budget, which prevents data from being recreated through multiple queries, in both systems is used as a counter measure to protect user privacy over long periods of time. Google's technique includes assigning a budget to each value for each metric, where as Apple assigns a budget to a whole category of metrics [6]. Academics argue that Apple's implementation addresses the issue of correlated metrics [6] but falling short when it comes to privacy protection over the long term based off of their epsilon usage [5]. This sounds trustworthy, but with Apple's code not being open-source it is hard to tell what other mechanisms are in place to guarantee user privacy over the long run [6].

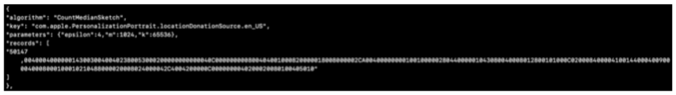
These two methods of implementing LDP tell us two things: High epsilon and assigning budgets to each value does not guarantee confidentiality over the long term. By utilizing low epsilon and assigning budgets to a category of data, confidentiality is held over the long term.

IV. Hypothesis and Empirical Evidence

Utilizing a lower epsilon value in Local Differential Privacy would provide users with more data confidentiality.

Let's take the Count Median Sketch (CMS) which is an algorithm that aggregates data into a histogram by count of a specific metric [11]. The client, in our case an iPhone, randomly samples a hash function and encodes a piece of data using size *m* turning it into a vector. In my use case, shown in Fig. 2, I had some location data sampled and aggregated on the server-side with *m* = 1024 bits with the last bit set to 1. Apple's algorithm then independently flips each bit with probability 1/(e^*e*/2)+1 to ensure differential privacy [11]. The vector and hash function are then sent to a server to be constructed into a sketch matrix M. By aggregating my privately transformed vector the matrix is generated with *k* rows and *m* columns. Corresponding to each hash function and size of the vector, respectively[11]. In my generated record I can see that *epsilon* is set to 4, the matrix has a column size of 1024 and row size of 65536. With epsilon set to 4 it shows that my data is less noisy than it could be with epsilon set to 1. This means that more of my data is exposed to potential vulnerabilities.

iPhone Xs Privacy Report Data



```
{
"algorithm": "CountMedianSketch",
"key": "com.apple.PersonalizationPortrait.locationDonationSource.en_US",
"parameters": {"epsilon":4,"m":1024,"k":65536},
"records": [
"50147
    ,004000400000001430030040040238005300020000000000000040C0000000000800404001000820000018000000002CA0040000000001001000002800440000001043080004000080128001010000C0200008400004100144000400900
    004000080001000102104880000020000802400004 2C4004200000C00000000402000200080100405010"
]
},
```

Fig. 2

To test the hypothesis I used Python to run 4 instances of Apple's Count Mean Sketch with different values of epsilon ranging from the base 1 to a ceiling of 167. The data set was synthetically created with 10 values and measured basic statistical metrics such as Mean, Variance, and Standard Deviation. Out of the 3 metrics Variance is the most meaningful due to its correlation with noise. The lower value of epsilon an instance has the higher the variance is, or in LDP's case, the more noise is in the output. Fig. 3 shown below contains the CMS Estimates on the y-axis and each instances epsilon value on the x-axis. Analyzing the violin plot it is clear that there is less noise with epsilon set to 4 as opposed to epsilon set to 1. The plot also shows that mean is virtually identical across the board showing that error due to noise is not a huge concern

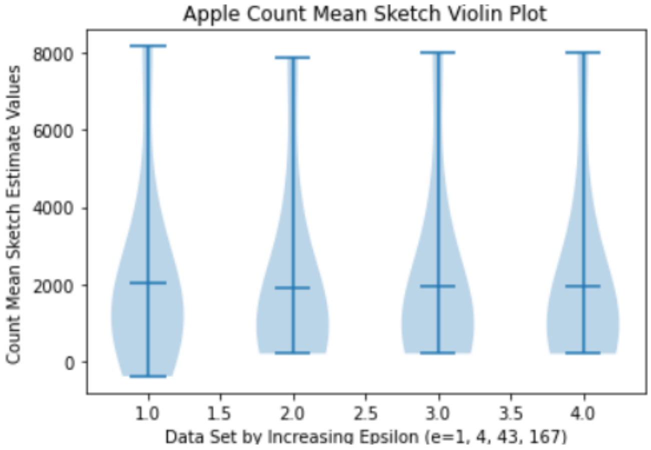


Fig 3.

V. Conclusion

Apple's illustration has thrown privacy into the public purview but do their products correctly implement it? My analysis has shown that real machine generated data is not being transformed with the highest amount of noise. I have proposed that Local Differential Privacy is only meaningful when using low values of epsilon to generate the most amount of noise in a dataset. The risk to organizations like Apple and Google is their ability to find meaningful trends out of the transformed data. However, as Petabytes of data are generated daily, end users should be given the highest priority when it comes to anonymity. It has been said that data is the new oil and if that's the case it can also be argued that unrefined oil is useless so let's treat data as such.

VI. Future Work

Future works may have access to raw data, larger time frames, and additional algorithms to compare LDP performance. A source in mind would be Samuel Maddock's python project [13] which includes Apple's Count Mean Sketch algorithm, Google's RAPPOR combined with Bloom filters, Hadamard Mechanism, and much more. Using their algorithm library in addition to Stochastic Gradient Descent to compare and test privacy over the long term has the potential to be an exciting avenue to go down.

VII. Bibliography

[1] P. Kühtreiber and D. Reinhardt, "Usable Differential Privacy for the Internet-of-Things," 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), 2021, pp. 426-427, doi: 10.1109/PerComWorkshops51409.2021.9431047. URL: https://ieeexplore-ieee-org.proxy.library.nyu.edu/stamp/stamp.jsp?tp=&arnumber=9431047&isnumber=9430860

[2] J. Domingo-Ferrer, D. Sánchez, and A. Blanco-Justicia, "The limits of differential privacy (and its misuse in data release and machine learning)," *Communications of the ACM*, vol. 64, no. 7, pp. 33–35, Jul. 2021, doi: 10.1145/3433638.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] J. R. Bambauer, K. Muralidhar, and R. Sarathy, "Fool's Gold: an Illustrated Critique of Differential Privacy," *papers.ssrn.com*, Sep. 15, 2013. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2326746

[4] B. Bebensee, "Local Differential Privacy: a tutorial," Jul. 2019. Accessed: Oct.24,2021. [Online]. Available: https://arxiv.org/pdf/1907.11908v1.pdf.

[5] A. Greenberg, "How One of Apple's Key Privacy Safeguards Falls Short," *Wired*, Sep. 15, 2017. https://www.wired.com/story/apple-differential-privacy-shortcomings/.

[6] B. Cyphers, "Differential privacy, part 3: Extraordinary claims require extraordinary scrutiny," *Access Now*, Nov. 30, 2017. https://www.accessnow.org/differential-privacy-part-3-extraordinary-claims-require-extraordinary-scrutiny/ (accessed Oct. 24, 2021).

[7] Apple, "Differential Privacy A privacy-preserving system." [Online]. Available: https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf .

[8] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response," *ACM SIGSAC Conference on Computer and Communications Security*, Nov. 2014, doi: 10.1145/2660267.2660348.

[9] M. Xu, B. Ding, T. Wang, and J. Zhou, "Collecting and analyzing data jointly from multiple services under local differential privacy," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 2760–2772, Aug. 2020, doi: 10.14778/3407790.3407859.

[10] Luisquintanilla, "Differential privacy in machine learning (preview) - Azure Machine Learning," *Differential privacy in machine learning (preview) - Azure Machine Learning | Microsoft Docs*. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/concept-differential-privacy.

[11] "Learning with Privacy at Scale," *Apple Machine Learning Research*. [Online].Available: https://machinelearning.apple.com/research/learning-with-privacy-at-scale.

[12] "ML: Stochastic Gradient Descent (SGD)," *GeeksforGeeks*, 13-Sep-2021. [Online]. Available: https://www.geeksforgeeks.org/ml-stochastic-gradient-descent-sgd/.

[13] S. Maddock, "Samuel-Maddock/pure-LDP: Python package for simple implementations of state-of-the-art LDP frequency estimation algorithms. Contains code for our VLDB 2021 Paper.," *GitHub*. [Online]. Available: https://github.com/Samuel-Maddock/pure-LDP.