Frequency Oracle for Sensitive Data Monitoring (Student Abstract)

Richard Sances[1], **Olivera Kotevska**[2], **Paul Laiu**[2]

[1]Virginia Polytechnic Institute and State University
[2]Oak Ridge National Laboratory
richardsances3@gmail.com, kotevskao@ornl.gov, laiump@ornl.gov

Abstract

As data privacy issues grow, finding the best privacy preservation algorithm for each situation is increasingly essential. This research has focused on understanding the frequency oracles (FO) privacy preservation algorithms. FO conduct the frequency estimation of any value in the domain. The aim is to explore how each can be best used and recommend which one to use with which data type. We experimented with different data scenarios and federated learning settings. Results showed clear guidance on when to use a specific algorithm.

Introduction

Today, more and more devices and systems are connecting to the Internet that collect data ranging from health monitoring to national security infrastructure systems. These systems share data which can be of various range and type and use machine learning methods for intelligent decisions. However, their risks of privacy leakage and identity thief increase. We need to have a methods that can provide the correct privacy-protection for each of them.

Local differential privacy (LDP) is a privacy model for distributed architectures that can provide a strong guarantee for each entity (user or sensor) while collecting the data (Joseph et al. 2018). Frequency Oracles (FOs) (Cormode, Maddock, and Maple 2021) are a type of LDP algorithms. Given a domain D, a FO is a protocol which estimates the frequency of an element $d \in D$. We organized them in a taxonomy Table 1 and listed selected algorithms in each category.

The grouping here is based on whether the algorithm uses channels to split the values (e.g., hash-based algorithms). Heavy hitters only return the most frequent estimations and will not estimate infrequent estimations, while sketches estimate the frequency of the items in a data set. Encoding algorithms try to determine the probability that index x of the dataset is privatized by the client. Based on how it is calculated if the item is in the dataset and probability of returning as it is, there are segment pair and Hadamard groups.

We study each type of algorithm and understand under which circumstances they perform the best. Benchmark analysis was designed to compare them to a baseline of different data properties. We also perform a study to understand how they can be used in a federated learning (FL) setting. FL is a distributed machine learning process where each client trains the data locally and sends the model parameters to a central server. However, FL can have leaks that lead to a breach of privacy. These leaks can be patched by using another layer of privatization, privatizing the model parameters before it is sent to the server.

This study represents the initial research and foundation for establishing a privacy-preservation recommendation system for the FO algorithm.

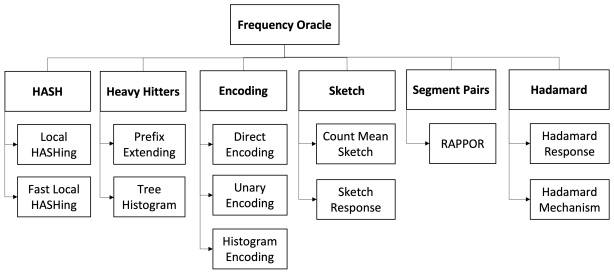

Figure 1: Taxonomy of a selected frequency oracles algorithms

Methodology

Data: (1) Smart meters in London (UKPowerNetworks 2018) contains energy consumption of 5567 households from 2012 to 2014; and *(2)* Energy consumption in Germany (Milojkovic 2021) of 955 households from 2018 to 2022.

Preprocessing: This work is split on two parts, first part explores the impact of data properties and the second part includes FL. (1) Each user was assigned a unique integer value and this new dataset would be privatized and aggregated by the client. (2) The data is split training 70% and testing 30%. This model creates weights for each client and uses them to predict the next value in the time series data.

Preliminary Results and Discussion

We studied how *data size*, *peak in data*, and the *privacy parameter (ϵ)* affected each FO. The research found that data size impacts the results (see Figure 2) and Direct Encoding

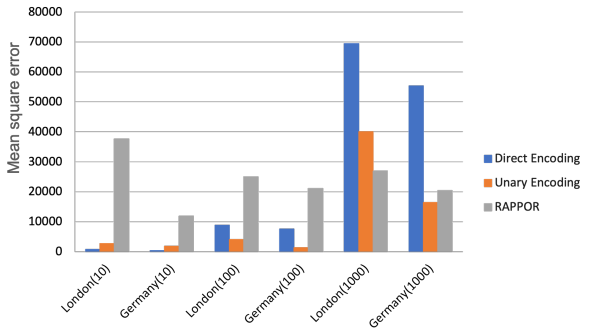

Figure 2: Impact of data size

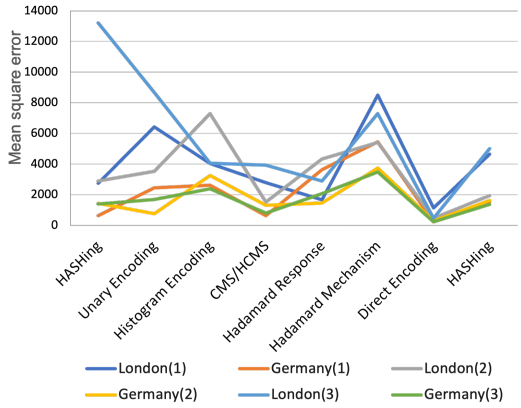

Figure 3: Impact of peak in the data

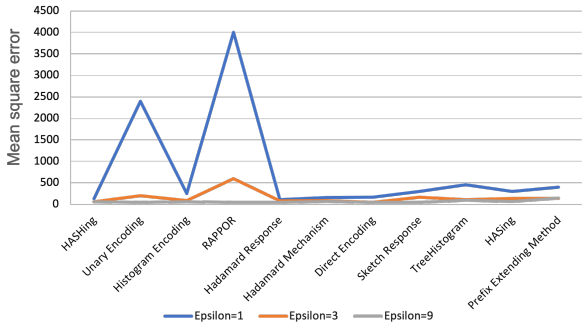

Figure 4: Impact of privacy loss (ϵ) parameter

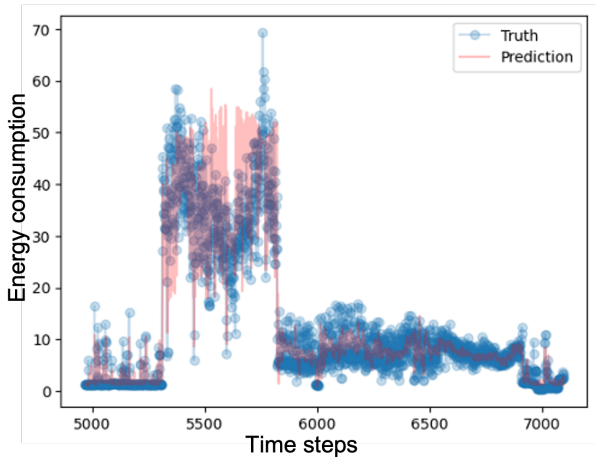

Figure 5: Federated Learning with Unary Encoding on weights and $\epsilon = 3$.

provides the highest accuracy for small data sets, Unary Encoding for large data sets, and RAPPOR for massive data sets. Peaks in the data had a small effect on the results (see Figure 3) and a larger epsilon resulted in more accurate frequency estimations (see Figure 4).

In FL context (see Figure 5), the research found that Unary Encoding, Histogram Encoding, Fast Local Hashing, and Hadamard Response were able to obtain accurate predictions even with the privatized data while Count Mean Sketch and Hadamard Mechanism struggled to make accurate predictions. Direct Encoding and Sketch Response were accurate when the epsilon value was high but inaccurate when the epsilon value was low. However, a change in epsilon had a small effect of RAPPOR's accuracy. Overall, privatizing the weights showed smaller error than privatizing the data on client side.

Conclusion and Future Work

Initial results showed a guidance on which algorithm to use based on the size of the dataset. However, changes in the range of data values had a negligible effect on the accuracy, and the accuracy was correlated with the privacy loss value. In the context of FL, some of the algorithm's performance did not change compared to non-FL settings, and some new behavior was noticed for Sketch Response and RAPPOR.

Further research is needed to investigate the algorithm performance with bigger and more diverse datasets. Also, to expand the range of algorithms and create a tool for automatic algorithm recommendation based on data properties and user preference.

Acknowledgments

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

References

Cormode, G.; Maddock, S.; and Maple, C. 2021. Frequency estimation under local differential privacy. *Proceedings of the VLDB Endowment*, 14(11): 2046–2058.

Joseph, M.; Roth, A.; Ullman, J.; and Waggoner, B. 2018. Local differential privacy for evolving data. *Advances in Neural Information Processing Systems*, 31.

Milojkovic, F. 2021. GEM HOUSE openData: German Electricity consumption in Many HOUSEholds over three years 2018-2020 (Fresh Energy).

UKPowerNetworks. 2018. SmartMeter Energy Consumption Data in London Households (Low Carbon London).