“Differential privacy in health research: A scoping review” Summary

"Differential Privacy in Health Research: A Scoping Review" discusses the importance of privacy in health research and the need to protect sensitive information. The review notes that some well-known privacy protection methods, such as the HIPAA safe harbor method and the Expert Determination method, may not always be effective. Differential privacy, which involves adding controlled noise to datasets, provides better protection against attackers who attempt to reconstruct the dataset to target specific individuals. The review examines the application of differential privacy in health research and identifies gaps in its evaluation and usage.

The review also highlights the increasing popularity of differential privacy as data has become a significant part of the tech world. The paper focuses on examining whether the current method of deploying this algorithm is sufficient. To determine the success of a study that applies differential privacy, they propose three main questions: the extent of differential privacy applications, the analytical purposes of algorithms, and the privacy-utility tradeoffs in specific health research contexts.

The scoping review analyzes over 50 research papers to examine the application of differential privacy specifically in health research. All 50 articles are under the scope of health or biomedical informatics, with 67% emphasizing a specific research or clinical application. The most common application areas were genomics at 24%, health surveillance with personal devices at 11%, and analysis of neuroimaging data at 7%. The rest are distributed toward different sub-areas. Some papers include datasets that include disease classification using neuroimaging data, medical phenotyping using electronic health records, and drug sensitivity prediction in cancer. This information is all very sensitive information to patients, thus, it is clear that this paper selected these papers carefully to emphasize the importance of differential privacy.

While differential privacy has been used in various health-related applications, they claim that there are gaps in its implementation. Especially in areas involving statistical inference and explanatory modeling. The authors recommend experimental deployment of each case to better understand the outcomes of real-world implementations. The articles also revealed privacy vulnerabilities among ethnicities with lower population sizes. Several ethnic groups with smaller sample sizes also experienced a significant loss of power for a differentially sensitive statistical test, potentially leading to greater ethical issues.

To promote ethical transparency and accountability for studies, the report recommends setting up an Epsilon Registry, basically, a method to find the best parameters when applying different differential privacy algorithms. As they found out these are rarely done before applying the method to a dataset in health-related areas, at least to the 50 sample papers. The paper excellently highlights the importance of privacy protection during medical studies. The paper is reliable as it carefully cites and organizes many other papers into different study cases, which are publicly available in the original paper. However, the paper falls short in proposing solutions, as it does not provide a specific solution for the gaps they identified. It only suggests the need for experimental deployment to understand the model outcomes of real-world implementations. This review also has a critical issue where the sample size of the papers it analyzed is not enough to represent the whole field.

Overall, the review provides valuable insights for researchers to consider when protecting sensitive information in health research and can serve as a guide for data analysts to consider privacy protection when selecting built-in functions from predefined libraries to add noise to their data, and it can contribute to the improvement of privacy protection in health research.

“Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality” Summary

This paper speaks about the concept of differential privacy and how it can be applied to data to protect confidentiality. It discusses the ongoing problem of the impact of COVID-19 on public health data. As COVID-19 rapidly infected the world, the need for the protection of people’s private health information became more apparent. Thus, researchers looked to differential privacy as a potential way of sharing and protecting patient privacy. They soon realized that differential privacy could not only protect private sector data but also public health data. The paper discusses the many solutions and use cases of differential privacy, particularly in public health data sharing.

The paper details current strategies that can ensure data privacy for public health. The first strategy is de-identification which is the process of removing personal identifiers and replacing them with token identifiers. It can remove or change identifying information from a dataset, such as a name and address, to protect individuals’ privacy. This makes it difficult to identify specific individuals in a dataset, particularly in cases where there is a small number of people reported in a geographic area or subgroup. The other strategies define ways to access, scramble, and decrypt data. While these strategies are effective at protecting data privacy to an extent, they have certain limitations that make them less effective than differential privacy methods. For example, data that has gone through de-identification can be potentially re-identifiable through indirect identification, using variables such as date of birth and postcode. A lot of these current strategies have vulnerabilities that make them more prone to re-identification than differential privacy.

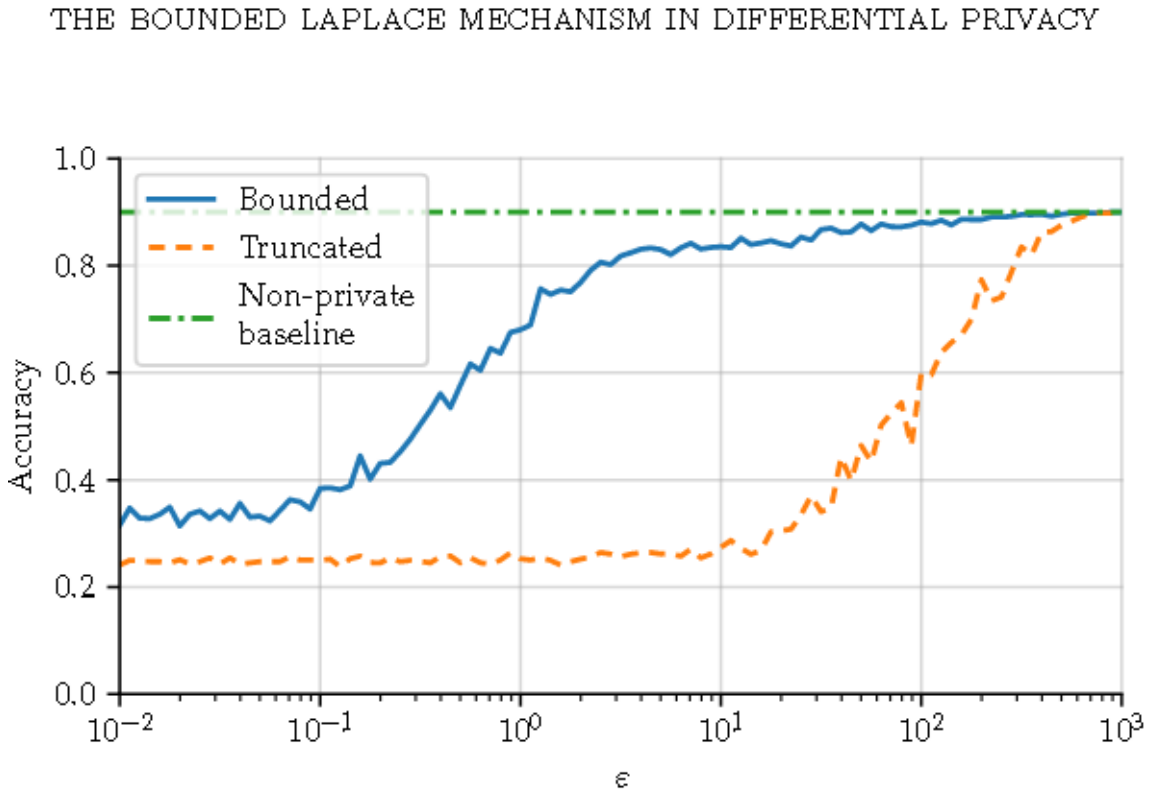
In the case study presented by the paper, an adversary, Eve, has access to unlimited computational resources about everyone in her community except Alice. She wants to analyze any data released about her community’s people to determine if Alice has COVID-19. Eve was able to deduce that Alice was diagnosed with COVID-19 because of binning. The goal of this case study is to find a solution that can protect Alice’s information. Because the Laplace mechanism is the proposed algorithm for our project, we focused on how the study utilizes the Laplace mechanism to apply noise to the statistics of their COVID-19 dataset. Random Laplace noise is added to the count of cases before being released to Eve. The effect is that it becomes more difficult for Eve to determine the exact number of cases in the community, so she will have a tougher time identifying if Alice has COVID-19. Bayes’ formula showed that there was a 47.5% chance that Eve’s guess was incorrect which is a drastic change from when there is no added Laplace noise. Thus, differential privacy makes it difficult to identify specific individuals in the dataset, but it still allows accurate analysis of overall patterns.

This paper provides a thorough explanation of differential privacy and its effectiveness in protecting individual privacy. However, it is not clear how effective the results would generalize to other contexts since the Laplace approach may not apply to other scenarios. But, we chose this paper since it related to our initial dataset and algorithm anyways. I thought it was very helpful that the paper discussed the advantages and limitations of using differential privacy for health data. Differential privacy is still a relatively new concept to us, so they helped us a lot in their explanations. They state that it is less suitable for datasets with low counts as the amount of noise applied will more significantly affect the results. Given this knowledge, we wanted to find a dataset with a larger number of counts and test out several parameters on it. Although the dataset we opted for is in a different area, we gained valuable knowledge from the paper on the applications and limitations of differential privacy for data.

Krishaan Patel

The Bounded Laplace Mechanism in Differential Privacy

The approach we are implementing is from the research paper “The Bounded Laplace Mechanism in Differential Privacy” by Naoise Holohan, Spiros Antonantos, Stefano Braghin, and Pol Mac Aonghusa. The paper starts by describing the notion of data privacy and its importance for data owners to consider whenever collecting, storing, sharing, and publishing user data. To address this, in recent years differential privacy arose as a popular privacy framework for introducing noise to data and thus adding privacy. Further expanding on this, the paper notes that “The Laplace Mechanism” is the workhorse of differential privacy due to its many applications where numerical data is processed and is said to have infinite support. Its popularity has come from the strength in mathematical and computational simplicity compared to other state of the art mechanisms. However, one pitfall that comes with the Laplace mechanism is that it lacks consistency in its outputs. As a result, problems can arise from that flexibility where the mechanism is able to return semantically impossible values such as negative results for a count query. To address this initial issue, the paper introduces the notion of bounding and the Bounded Laplace Mechanism which is a modified version of the standard Laplace Mechanism where constraints are added to the Laplace distribution to ensure the probability of the noise is within the allotted bounds for an allowed output.After introducing these initial concepts, the paper begins analysis and further investigation into the Bounded Laplace Mechanism. In Example 1.3, they first describe the setup for a naive Bayes classifier which is a probabilistic classifier that learns means and variances of each feature for each label. This will allow application of Bayes’ theorem on unseen examples to classify and predict them. Furthermore, differential privacy will be implemented here by adding appropriately-scaled noise to means and variances. Here, it is highlighted that the standard laplace mechanism could otherwise be used, but the bounded outputs are required to perturb variances as variances cannot be negative. Finally, they will implement a 80%/20% train/test split for the naive Bayes classifier to observe accuracy versus epsilon as an average over 100 simulations for each epsilon. This is conducted on the IRIS dataset from Anderson (1936) and Fisher (1936). The following figure results:



Here, we can observe the baseline (non-private) accuracy from the train test split on the Bayes classifier was at 90%. Overall, the paper asserts that bounding is shown to outperform truncation over a majority of the epsilon values. This was explained by the singularity produced in the Gaussian distribution at zero variance, since that is the lower bound of the output domain.

After going through analysis of the Bounded Laplace mechanism and reviewing results, the research paper transitions and introduces the main problem within the paper. The paper states that the bounded Laplace mechanism does not typically satisfy differential privacy when inheriting the parameters from the standard Laplace mechanism. Because of this, the paper states the need to accurately and effectively protect sensitive data while still maintaining capability of meaningful analysis. A solution is introduced here stating that an optimal noise scale must be determined for the parameter of the bounded Laplace mechanism in order to ensure differential privacy is maintained. In doing this, certain preliminaries and definitions are laid out prior to investigating how to determine the ideal scale parameter. After laying out the prerequisite material, the main result is then presented which is providing a definition of the scale parameter that is required for the bounded Laplace mechanism.

The main contribution and finding of the paper is laid out here with demonstrations of how to calculate this ideal scale parameter and laying out the algorithm needed for finding it.

Definition 4.1 (Fixed Point Operator). Given $\Delta Q > 0$, $\epsilon \geq 0$ and $0 \leq \delta \leq 1$, we define the fixed point operator $f : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ by

$$f(b) = \frac{\Delta Q}{\epsilon - \log \Delta C(b) - \log(1 - \delta)}. \tag{4.1}$$

Any positive fixed point of f (*i.e.* $b^* = f(b^*) > 0$) will act as a differentially private scale parameter for the bounded Laplace mechanism. In advance of examining f, we first define

$$b_0 = \frac{\Delta Q}{\epsilon - \log(1 - \delta)}. \tag{4.2}$$

Note that b_0 determines the variance required for the standard Laplace mechanism to achieve (ϵ,δ)-differential privacy.

We now present a number of lemmas concerning f, namely: (i) the value of $f(b_0)$; and (ii) the monotonicity of f. Proofs are given in Appendices A.5 and A.6.

Lemma 4.2. *$f(b_0) \geq b_0$, and $f(b_0) = b_0$ if and only if $\Delta Q = u - l$.*

Lemma 4.3. *$f'(b) \leq 0$ whenever $b \neq 0$, and $f'(b) = 0$ if and only if $\Delta Q = u - l$.*

This leads us to the main result of this section, that f has a unique fixed point b^*.

Here the paper lays out the notion of a fixed point operator which is needed for determining an ideal scale. Furthermore, an algorithm is laid out for describing how to obtain a needed fixed point from the fixed point operator.

Algorithm 1: A robust and precise method for finding b^*

input : Fixed point operator f (as given in (4.1)), b_0 as given (4.2)
output: Fixed point b^*, where $f(b^*) = b^*$

left $\leftarrow b_0$;
right $\leftarrow f(b_0)$;
intervalSize $\leftarrow$ (left + right) $\times$ 2;

while intervalSize $>$ right $-$ left **do**
    intervalSize $\leftarrow$ right $-$ left;
    b $= \frac{\text{left+right}}{2}$;
    if $f(b) \geq$ b **then**
        left $\leftarrow$ b;
    end
    if $f(b) \leq$ b **then**
        right $\leftarrow$ b;
    end
end

return b;

These are the main contributions outlined by the research paper to evaluate the specific scale value that would suit as a parameter to satisfy differential privacy for the bounded Laplace mechanism and replace the inherited parameters that are insufficient in the standard Laplace mechanism.

One of the strengths of this paper is the thoroughness and brevity of the theoretical analysis that provides a clear explanation of the limitations and capabilities of the bounded Laplace Mechanism. That being said, one weakness observed is that the analyzed results could be limited to simpler scenarios and may not necessarily apply to more complex real world situations and data. Further research could still be useful for analyzing the potential of the bounded Laplace mechanism in more complex situations and contexts. Overall, the paper serves as a valuable contribution to applying the Laplace mechanism and for differential privacy as a whole. A successful and robust new approach was provided to accurately and efficiently provide security and allow privacy researchers to deploy the bounded Laplace mechanism with confidence and certainty.

Reference: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/715/690