APPENDIX A

State of the Art: De-Identification

De-identification, which is also called anonymization in Europe, is an approach for stripping the PII from a dataset. This term encompasses a broad and diverse range of techniques for mitigating the risk of linkage attacks and other misuses of datasets that contain PII. There is a utility vs. privacy tradeoff, however, in that a greater level of difficulty to carry out a linkage attack against a de-identified dataset will most likely imply a reduced utility of the de-identified dataset for analysis and research purposes.

Currently popular de-identification techniques, such as field suppression (and other field specific perturbations) and guaranteeing k-anonymity, which preserve the utility of the dataset, often must sacrifice too great a level of privacy (needed to prevent linkage attacks and other potentially damaging uses of the datasets). In addition, it is difficult or most often impossible to quantify the amount of privacy that is lost with these techniques.

There is a growing body of academic research in the field of differential privacy which claim strict mathematical guarantees of privacy, although at a potentially greater loss of dataset utility. These techniques may hold great promise in the field of data de-identification, but only if the utility of the de-identified datasets that they produce can be substantially improved.

Several differentially private algorithms have been proposed for creating de-identified datasets that closely model the statistical properties of the original data while maintaining a formal privacy guarantee. The Multiplicative Weights and Exponential Mechanisms (MWEM) algorithm, one such well-known algorithm, maintains and corrects an approximating distribution to the true data through queries on which the approximate and true datasets differ. Since differential privacy is a probabilistic concept, any differentially private mechanism is necessarily randomized. Some of these, like the Laplace mechanism rely on adding controlled noise to the value of the function being computed while others, like the exponential mechanism and posterior sampling sample from a problem-dependent family of distributions instead.

Algorithms like MWEM can be created by combining randomized mechanisms such as the following non-exhaustive list, to improve the resulting privacy and similarity of the new dataset to the original dataset:

- **Laplace Mechanism** adds Laplace noise (i.e. noise from the Laplace distribution, which can be expressed by probability density function which has mean zero and standard deviation$\sqrt{2}$. (https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf, p.30)
- **Exponential Mechanism** protects privacy for non-numerical values, adding one-sided Laplace noise and privately returns the element of a range with the highest score. Read more here. (https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf , p.37)

- **Report Noisy Max Mechanism** which provides a method of privately releasing the maximal element of a set by adding regular Laplace noise and selecting the query with the greatest noisy value. (https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf , p.35)
- **Above Threshold Algorithm** takes a list of numerical queries as input and outputs the first query whose answer is above a certain threshold. Above Threshold is the critical component of the Sparse Vector technique. (https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf , p.55-58)
- **Sparse Vector** Technique (SVT), used in interactive and non-interactive settings, takes a long stream of queries and runs the Laplace mechanism on each query while only outputting the noisy answers whose noise is above some noisy (Laplace) threshold. It has the unique quality that one can output some query answers without apparently paying any privacy cost but most variants of SVT were found to actually not be privacy preserving. Read more here. https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf , p.55)

The above mechanisms are each privacy-preserving in a quantifiable sense, but none are a complete solution on their own. Rather, these mechanisms are more like the basic building blocks for more complete differentially private algorithms, wherein the composition order and choices of which mechanisms to apply at each point in the algorithm's execution are important. These mechanisms, when combined in different orders and quantities, while also considering the research question, can form an algorithm that does produce a de-identified dataset that is potentially useful for some research questions.

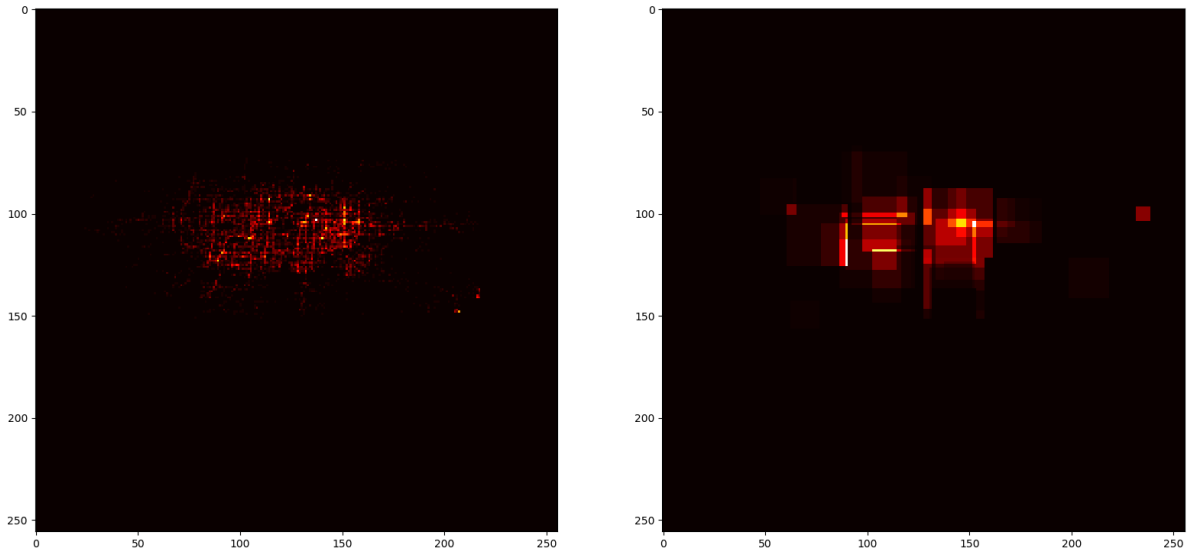
Privacy loss is quantified in the field of differential privacy through numerical parameters, or **Privacy Budgets**. As a dataset is being de-identified using a differentially private algorithm, information may be leaked about the original dataset at each point in the algorithm. These leaks add up and are irreversible. A cap on the total amount of information that will be allowed to leak is usually set, which is called a privacy budget, and is usually denoted by ε. Typical values for ε are usually set to be in the range of 0.1 to 2.0, although it is not really known from a practical point of view how the choice of this value corresponds to actual privacy loss. From the point of view of this competition we are seeking solutions that fit into this paradigm, or a generalization of this called (ε,δ)-differential privacy (see https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf for a detailed description of this generalization).

The constraint of a reasonable privacy budget can wreak havoc on utility as observed when training complex machine language (ML) models. In a paper by Fredreikson et al. in 2014, the authors used ML techniques on a public database linking Warfarin dosage outcomes to specific genetic markers to develop a dosing model. They tested the accuracy of the dosing model created by applying differentially private algorithms for various privacy budgets while training the model. The model performed very well when privacy budgets were large, but at those levels

sensitive patient information was allowed to leak; when privacy was adequately protected (corresponding to low values of ε), hypothetical patients were given lethal doses.

An example of an algorithm that does not exhibit great performance but, for illustrative purposes, is easier to describe, is the aforementioned MWEM algorithm, which is a combination of the Multiplicative Weights approach developed in 2010 by Hardt, Ligett, and McSherry. The purpose of the algorithm is to construct a synthetic dataset which models the original dataset as closely as possible with respect to a set of queries which the end user cares about. The algorithm is an iterative procedure that, during each loop, maintains a synthetic dataset which is its current "best guess" approximation to the original dataset. During each iteration a query is first selected using the Exponential Mechanism, the query answer is returned with some random noise added via the Laplace Mechanism, and finally the noisy query answer is used to update the algorithm's current "best guess" approximation to the original dataset. After a finite number of iterations, the algorithm terminates and returns its "best guess" approximation.

The algorithm was tested on range queries, contingency table release across a collection of statistical benchmarks, and datacube release. The algorithm is limited however, in terms of the quality of the approximation that the returned de-identified dataset provides to the original dataset, as measured by the standard measure, Kullback-Leibler Divergence, from information theory. Because the authors performed less than 100 iterative queries (out of thousands of possible queries) on their datasets, there will necessarily be a large degree of information loss, which means that it is a very rough approximation to the original. In a dataset with data universe of size N, you would need the synthetic dataset to be accurate on N queries to get a perfect reconstruction, and most data universes will be much larger than 100. A less than perfect reconstruction might need a much smaller number of queries on which it is accurate, but probably more than 100. These plots of the original dataset and the resulting de-identified dataset from this algorithm, with a reasonable privacy budget, show the poor quality of the differentially private de-identified set. Here are two additional plots of the results of the MWEM algorithm:



(The image on the left is a map of taxi cab ride data in Beijing, while the image on the right is of a de-identified dataset using the MWEM algorithm.)