Comparing Local and Central Differential Privacy Using Membership Inference Attacks

Daniel Bernau, Jonas Robl, Philip W. Grassal, Steffen Schneider, Florian Kerschbaum

Comparing local and central differential privacy using membership inference attacks

Daniel Bernau[1], Jonas Robl[1⋆], Philip W. Grassal[2⋆⋆⋆], Steffen Schneider[3⋆⋆], and Florian Kerschbaum[4]

[1] SAP, Karlsruhe, Germany, firstname.lastname@sap.com
[2] Heidelberg University, Heidelberg, Germany, philip-william.grassal@iwr.uni-heidelberg.de
[3] Procure.AI, London, United Kingdom, steffen.schneider@procure.ai
[4] University of Waterloo, Waterloo, Canada, florian.kerschbaum@uwaterloo.ca

Abstract. Attacks that aim to identify the training data of neural networks represent a severe threat to the privacy of individuals in the training dataset. A possible protection is offered by anonymization of the training data or training function with differential privacy. However, data scientists can choose between local and central differential privacy, and need to select meaningful privacy parameters ϵ. A comparison of local and central differential privacy based on the privacy parameters furthermore potentially leads data scientists to incorrect conclusions, since the privacy parameters are reflecting different types of mechanisms.
Instead, we empirically compare the relative privacy-accuracy trade-off of one central and two local differential privacy mechanisms under a white-box membership inference attack. While membership inference only reflects a lower bound on inference risk and differential privacy formulates an upper bound, our experiments with several datasets show that the privacy-accuracy trade-off is similar for both types of mechanisms despite the large difference in their upper bound. This suggests that the upper bound is far from the practical susceptibility to membership inference. Thus, small ϵ in central differential privacy and large ϵ in local differential privacy result in similar membership inference risks, and local differential privacy can be a meaningful alternative to central differential privacy for differentially private deep learning besides the comparatively higher privacy parameters.

Keywords: Anonymization · Membership Inference · Neural Networks.

1 Introduction

Neural networks have successfully been applied to a wide range of learning tasks, each requiring its own specific set of training data, architecture and hyperparameters to achieve meaningful classification accuracy and foster generalization.

⋆ Authors contributed equally to this research.
⋆⋆ This work was done during an internship at SAP.

In some learning tasks data scientists have to deal with personally identifiable or sensitive information, which results in two challenges. First, legal restrictions might not permit collecting, processing or publishing certain data, such as National Health Service data [5]. Second, membership inference (MI) [20,31,38] and model inversion attacks [15,16] are capable of identifying and reconstructing training data based on information leakage from a trained, published neural network model. A mitigation to both challenges is offered by anonymized deep neural network training with differential privacy (DP). However, a data scientist can choose between two categories of DP mechanisms: local DP (LDP) [40] and central DP (CDP) [9]. LDP perturbs the training data before any processing takes place, whereas CDP perturbs the gradient update steps during training. The degree of perturbation, which affects the accuracy of the trained neural network on test data, is calibrated for both DP categories by adjusting their respective privacy parameter ϵ. Choosing ϵ too large will unlikely mitigate privacy attacks such as MI, and setting ϵ too small will significantly reduce model accuracy. Balancing the privacy-accuracy trade-off is a challenging problem especially for data scientists who are not experts in DP. Furthermore, data scientists might rule out LDP when designing differentially private neural networks due to concerns raised by the comparatively higher privacy parameter ϵ in LDP. In this work, we compare the empirical privacy protection under the white-box MI attack of Nasr et al. [31] for LDP and CDP mechanisms for learning problems from diverse domains: consumer preferences, face recognition and health data. The MI attack indicates a lower bound on the inference risk whereas DP formulates an upper bound [24,43,44], but in practice even high privacy parameters such as experienced in LDP may already offer protection. This work makes the following contributions:

- Comparing LDP and CDP by the average precision of their MI precision-recall curve as privacy measure, and showing that under this measure LDP and CDP have similar privacy-accuracy trade-offs despite vastly different ϵ.
- Showing that CDP mechanisms are not achieving a consistently better privacy-accuracy trade-off on various datasets and reference models. The trade-off rather depends on the specific dataset.
- Analyzing the relative privacy-accuracy trade-off and showing that it is not constant over ϵ, but that for each data set there are ranges where the relative trade-off is greater for protection against MI than accuracy.

Section 2 revisits differential privacy and Section 3 formulates our approach for comparing LDP and CDP under membership inference. We describe evaluation datasets in Section 4. Findings are presented in Section 5 and discussed in Section 6. Related work and conclusions are provided in Section 7 and Section 8.

2 Differential Privacy

DP [8] anonymizes a dataset $\mathcal{D} = \{d_1, \ldots, d_n\}$ by perturbation and can be either enforced centrally to a function $f(\mathcal{D})$, or locally to each entry $d \in \mathcal{D}$.

2.1 Central DP

In central DP the aggregation function $f(\cdot)$ is evaluated and perturbed by a trusted server. Due to perturbation, it is no longer possible for an adversary to confidently determine whether $f(\cdot)$ was evaluated on $\mathcal{D}$, or some neighboring dataset $\mathcal{D}'$ differing in one element. Assuming that every participant is represented by one element, privacy is provided to participants in $\mathcal{D}$ as their impact on $f(\cdot)$ is limited. *Mechanisms* $\mathcal{M}$ fulfilling Definition 1 are used for perturbation of $f(\cdot)$ [9]. We refer to the application of a mechanism $\mathcal{M}$ to a function $f(\cdot)$ as *central differential privacy*. CDP holds for all possible differences $\|f(\mathcal{D}) - f(\mathcal{D}')\|_2$ by adapting to the global sensitivity of $f(\cdot)$ per Definition 2.

Definition 1 ((ϵ, δ)-central differential privacy). *A mechanism $\mathcal{M}$ gives (ϵ, δ)-central differential privacy if $\mathcal{D}, \mathcal{D}' \subseteq \mathcal{DOM}$ differing in at most one element, and all outputs $\mathcal{S} \subseteq \mathcal{R}$*

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta$$

Definition 2 (Global ℓ_2 Sensitivity). *Let $\mathcal{D}$ and $\mathcal{D}'$ be neighboring. The global ℓ_2 sensitivity of a function f, denoted by Δf, is defined as*

$$\Delta f = max_{\mathcal{D},\mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_2.$$

Definition 3 (Gaussian Mechanism [10]).
Let $\epsilon \in (0, 1)$ be arbitrary. For $c^2 > 2ln(\frac{1.25}{\delta})$, the Gaussian mechanism with parameter $\sigma \geq c\frac{\Delta f}{\epsilon}$ gives (ϵ, δ)-CDP, adding noise scaled to $\mathcal{N}(0, \sigma^2)$.

For CDP in deep learning we use differentially private versions[5] of two standard gradient optimizers: SGD and Adam [27]. We refer to these CDP optimizers as DP-SGD and DP-Adam. A CDP optimizer represents a differentially private training mechanism $\mathcal{M}_{nn}$ that updates the weight coefficients θ_t of a neural network per training step $t \in T$ with $\theta_t \leftarrow \theta_{t-1} - \alpha(\tilde{g})$, where $\tilde{g} = \mathcal{M}_{nn}(\partial loss/\partial\theta_{t-1})$ denotes a Gaussian perturbed gradient and α is some scaling function on $\tilde{g}$ to compute an update, i.e., learning rate or running moment estimations. Differentially private noise is added by the Gaussian mechanism of Definition 3 [1]. After T update steps, $\mathcal{M}_{nn}$ outputs a differentially private weight matrix θ which is used by the prediction function $h(\cdot)$ of a neural network. A CDP gradient optimizer bounds the sensitivity of the computed gradients by clipping norm $\mathcal{C}$ based on which the gradients get clipped before perturbation. Since weight updates are performed iteratively during training a composition of $\mathcal{M}_{nn}$ is required until the the training step T is reached and the final private weights θ are obtained. For CDP we measure privacy decay under composition by tracking σ of the Gaussian mechanism. After training we transform and compose σ under Renyi DP [29], and transform the aggregate again to CDP. We choose this accumulation method over other composition schemes [1,25] due to the tighter bound for heterogeneous mechanism invocations.

[5] We used Tensorflow Privacy: `https://github.com/tensorflow/privacy`

2.2 Local DP

We refer to the perturbation of entries $d \in \mathcal{D}$ as local differential privacy [40]. LDP is the standard choice when the server which evaluates a function $f(\mathcal{D})$ is untrusted. We adapt the definitions of Kasiviswanathan et al. [26] to achieve LDP by using local randomizers $\mathcal{LR}$. In the experiments within this work we use a local randomizer to perturb each record $d \in \mathcal{D}$ independently. Since a record may contain multiple correlated features (e.g., items in a preference vector) a local randomizer must be applied sequentially which results in a linearly increasing privacy loss. A series of local randomizer executions per record composes a local algorithm according to Definition 5. ϵ-local algorithms are ϵ-local differentially private [26], where ϵ is a summation of all composed local randomizer guarantees. We perturb low domain data with randomized response [41], a (composed) local randomizer. By Equation (1) randomized response yields $\epsilon = \ln(3)$ LDP for a one-time collection of values from binary domains (e.g., $\{\texttt{yes}, \texttt{no}\}$) with two fair coins [12]. That is, retention of the original value with probability $\rho = 0.5$ and uniform sampling with probability $(1-\rho) \cdot 0.5$.

$$\epsilon = \ln\left(\frac{\rho + (1-\rho)\cdot 0.5}{(1-\rho)\cdot 0.5}\right) = \ln\left(\frac{\Pr[\texttt{yes}|\texttt{yes}]}{\Pr[\texttt{yes}|\texttt{no}]}\right). \tag{1}$$

Definition 4 (Local differential privacy). *A local randomizer (mechanism) $\mathcal{LR} : \mathcal{DOM} \to \mathcal{S}$ is ϵ-local differentially private, if $\epsilon \geq 0$ and for all possible inputs $v, v' \in \mathcal{DOM}$ and all possible outcomes $s \in \mathcal{S}$ of $\mathcal{LR}$*

$$\Pr[\mathcal{LR}(v) = s] \leq e^{\epsilon} \cdot \Pr[\mathcal{LR}(v') = s]$$

Definition 5 (Local Algorithm). *An algorithm is ϵ-local if it accesses the database $\mathcal{D}$ via $\mathcal{LR}$ with the following restriction: for all $i \in \{1, \ldots, |\mathcal{D}|\}$, if $\mathcal{LR}_1(i), \ldots, \mathcal{LR}_k(i)$ are the algorithms invocations of $\mathcal{LR}$ on index i, where each $\mathcal{LR}_j$ is an ϵ_j-local randomizer, then $\epsilon_1 + \ldots + \epsilon_k \leq \epsilon$.*

Definition 6 (Laplace Mechanism [10]). *Given a numerical query function $f : DOM \to \mathbb{R}^k$, the Laplace mechanism with $\lambda = \frac{\Delta_f}{\epsilon}$ is an ϵ-differentially private mechanism, adding noise scaled to $Lap(\lambda, \mu = 0)$.*

In our evaluation we also look at image data for which we rely on the local randomizer by Fan [14] for LDP image pixelization. The randomizer applies the Laplace mechanism of Definition 6 with scale $\lambda = \frac{255 \cdot m}{b^2 \cdot \epsilon}$ to each pixel, thus fulfilling Definition 4. Parameter m represents the neighborhood in which LDP is provided. Full neighborhood for an image dataset would require that any picture can become any other picture. In general, providing DP or LDP within a large neighborhood will require high ϵ values to retain meaningful image structure. High privacy will result in random black and white images. Within this work we consider the use of LDP and CDP for deep learning along a generic data

science process (e.g., CRISP-DM [42]). In such a processes the dataset $\mathcal{D}$ of a data owner $\mathcal{DO}$ is (i) transformed, and (ii) used to learn a model function $h(\cdot)$ (e.g., classification), which (iii) afterwards is deployed for evaluation by third parties. In the following $h(\cdot)$ will represent a neural network. DP is applicable at every stage in the data science process. In the form of LDP by perturbing each record $d \in \mathcal{D}$, while learning $h(\cdot)$ centrally with a CDP gradient optimizer, or to the evaluation of $h(\cdot)$ by federated learning with voting. We leave learning with more than two parties, such as used in PATE [33] with CDP or amplification by shuffling for LDP [11] as future work. However, independent of the stage of application, the privacy-accuracy trade-off is of particular interest. We follow the evaluation of regularization techniques that apply noise to the training data to foster generalization [17,18,28] and measure utility by the test accuracy of $h(\cdot)$.

3 White-Box MI Attack

Membership inference (MI) attacks aim at identifying the presence or absence of individual records in the training data of data owner $\mathcal{DO}$. MI attacks are of particular importance for members of the training dataset when the nature of the training dataset is revealing sensitive information. For example, a medical training dataset containing patients with different types of cancer, or a training dataset that is used to predict the week of pregnancy based on the shopping cart [21]. A related attack building upon MI is attribute inference [44] where individual records are partially known and specific attribute values shall be inferred. In this work we solely consider MI since protection against MI offers protection against attribute inference. In specific, we consider white-box MI by Nasr et al. [31] which is stronger than previously suggested black-box MI attacks (e.g., Shokri et al. [38]). The MI attack assumes an honest-but-curious adversary $\mathcal{A}$ with access to a trained prediction function $h(\cdot)$, knowledge about the hyperparameters and DP mechanisms that were used for training. We refer to the trained prediction function as *target model* and the training data of $\mathcal{DO}$ as $\mathcal{D}^{\texttt{train}}_{\texttt{target}}$. Given this accessible information $\mathcal{A}$ wants to learn a binary classifier, the *attack model*, that allows to classify data into members and non-members w.r.t. the target model training dataset with high accuracy. The accuracy of an MI attack model is evaluated on a balanced dataset including all members (target model training data) and an equal number of non-members (target model test data), which simulates the worst case where $\mathcal{A}$ tests membership for all training records. White-box MI exploits that an ML classifier such as a neural network (NN) tends to classify a record $d = (x, y)$ from its training dataset $\mathcal{D}^{\texttt{train}}_{\texttt{target}}$ with different confidence $p(\mathbf{x})$ given $h(\mathbf{x})$ for features $\mathbf{x}$ and true label y than a record $d \notin \mathcal{D}^{\texttt{train}}_{\texttt{target}}$. White-box MI makes two assumptions about $\mathcal{A}$. First, $\mathcal{A}$ is able to observe internal features of the ML model in addition to external features (i.e., model outputs). The internal features comprise observed losses $L(h(x; W))$, gradients $\frac{\delta L}{\delta W}$ and the learned weights W of $h(\cdot)$. Second, $\mathcal{A}$ is aware of a portion of $\mathcal{D}^{train}_{target}$ and $\mathcal{D}^{test}_{target}$. These portions were set to 50% by Nasr et al. [31] and will be the same within this work to allow comparison. Second, $\mathcal{A}$ extracts internal

and external features of a balanced set of confirmed members and non-members. An illustration of the white-box MI attack is given in Figure 1. Again, $\mathcal{A}$ is assumed to know a portion of $\mathcal{D}_{target}^{train}$ and $\mathcal{D}_{target}^{test}$ and generates attack features by passing these records through the trained target model. $\mathcal{A}$ trains a binary classification attack model per target variable $y \in Y$ to map $p(\mathbf{x})$ to the indicator "$\texttt{in}$" or "$\texttt{out}$". The set $(L(h(x;W)), \frac{\delta L}{\delta W}, p(\mathbf{x}), y, \texttt{in}/\texttt{out})$ serves as attack model training data, i.e., $\mathcal{D}_{\texttt{attack}}^{\texttt{train}}$. Thus, the MI attack model exploits the imbalance between predictions on $d \in \mathcal{D}_{\texttt{target}}^{\texttt{train}}$ and $d \notin \mathcal{D}_{\texttt{target}}^{\texttt{train}}$. Attack model accuracy is computed on features extracted from the target model likewise.

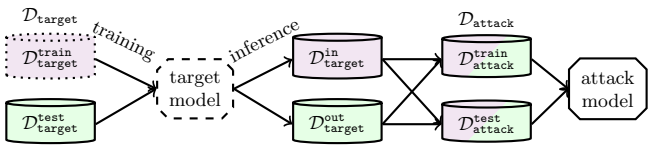


Fig. 1: White-box MI with attack features $(y^*, p(\mathbf{x}), L(h(x;W),y), \frac{\delta L}{\delta W})$. LDP perturbation on $\mathcal{D}_{\texttt{target}}^{\texttt{train}}$ (dotted) and CDP on target model training (dashed). Target model training is colored: training (violet) and validation (green).

3.1 Evaluating CDP and LDP under MI

DP has been shown to formulate a theoretical upper bound on the accuracy of MI adversaries [44], and thus the use of DP should impact the classification accuracy of $\mathcal{A}$. To illustrate the effect of the privacy parameter ϵ on the MI attack we focus on two questions related to the identifiability of training data within this work: "How many records predicted as $\texttt{in}$ are truly contained in the training dataset?" (precision), and "How many truly contained records are predicted as $\texttt{in}$?" (recall). For analysis we use precision-recall curves which depict the precision and recall for various classification thresholds, and thus reflect the possible MI attack accuracies of $\mathcal{A}$. We compare the precision-recall curves by their average precision (AP) to assess the overall effect of DP on MI. The AP approximates the integral under the precision-recall curve as a a weighted mean of the precision P per threshold t and the increase in recall R from the previous threshold, i.e.: $AP = \sum_t (R_t - R_{t-1}) \cdot P_t$. We prefer this non-interpolated technique over interpolated calculations of the area under curve, since the precision-recall curve is not guaranteed to decline monotonically and thus the linear trapezoidal interpolation might yield an overoptimistic representation [7,13]. Good MI attack models will realize an AP of close to 1 while poor MI attack models will be close to the baseline of uniform random guessing, hence $AP = 0.5$. The data owner $\mathcal{DO}$ has two options to apply DP against MI within the data science process introduced in Section 2. Either in the form of LDP by applying a local randomizer to the training data and using the resulting $\mathcal{LR}(\mathcal{D}_{\texttt{target}}^{\texttt{train}})$ for training, or

CDP with a differentially private optimizer on $\mathcal{D}^{\texttt{train}}_{\texttt{target}}$. A discussion and comparison of LDP and CDP purely based on the privacy parameter ϵ likely falls short and potentially leads data scientists to incorrect conclusions, since the privacy parameters are reflecting different types of mechanisms. Furthermore, data scientists give up flexibility w.r.t. applicable learning algorithms, if ruling out the use of LDP due to comparatively greater ϵ and instead solely investigating CDP (e.g., DP-SGD). We suggest to compare LDP and CDP by their concrete effect on the AP and the resulting privacy-accuracy trade-off. While we consider a specific MI attack our methodology is applicable to other MI attacks as well. Models that use CDP are represented by dashed lines in Figure 1. In the LDP setup, the target model is trained with perturbed records from a local randomizer, i.e., $\mathcal{LR}(\mathcal{D}^{\texttt{train}}_{\texttt{target}})$. However, in order to increase his attack accuracy $\mathcal{A}$ needs to learn attack models with high accuracy on the original data from which the perturbed records stem, i.e., $\mathcal{D}^{\texttt{train}}_{\texttt{target}}$. Perturbation with LDP is represented by dotted lines in Figure 1.

3.2 Relative Privacy-Accuracy Trade-off

We calculate the relative privacy-accuracy trade-off for LDP and CDP as the relative difference between $\mathcal{A}$'s change in AP to $\mathcal{DO}$'s change in test accuracy. Let AP_{orig}, AP_ϵ be the MI APs and ACC_{orig}, ACC_ϵ be the test accuracies for the original and DP target model. Furthermore, let ACC_{base} be the baseline test accuracy of uniform random guessing $1/\mathbb{C}$, where $\mathbb{C}$ denotes the number of classes in the dataset, and AP_{base} be the baseline AP at 0.5. We fix ACC_{base}, AP_{base}, since $\mathcal{A}$ or $\mathcal{DO}$ would perform worse than uniform random guessing at lower values. Rearranging and bounding the cases where AP and ACC increases over ϵ yields:

$$
\begin{aligned}
\varphi &= \frac{(AP_{orig} - AP_\epsilon)/(AP_{orig} - AP_{base})}{(ACC_{orig} - ACC_\epsilon)/(ACC_{orig} - ACC_{base})} \\
\varphi &= \frac{\max(0, AP_{orig} - AP_\epsilon) \cdot (ACC_{orig} - ACC_{base})}{\max(0, ACC_{orig} - ACC_\epsilon) \cdot (AP_{orig} - AP_{base})} \\
\varphi &= \min\left(2, \frac{\max(0, (AP_{orig} - AP_\epsilon) \cdot (ACC_{orig} - ACC_{base}))}{\max(0, (ACC_{orig} - ACC_\epsilon) \cdot (AP_{orig} - AP_{base}))}\right)
\end{aligned}
$$

To avoid φ from approaching infinitely large values when the accuracy remains stable while AP decreases significantly, and the undefined case of $ACC_{orig} \leq ACC_\epsilon$, we bound φ at 2. In consequence, when the relative gain in privacy (lower AP) exceeds the relative loss in accuracy, it applies that $1 < \varphi \leq 2$, and $0 \leq \varphi < 1$ when the loss in test accuracy exceeds the gain in privacy. Hence, φ quantifies the relative loss in accuracy and the relative gains in privacy for a given privacy parameter ϵ and captures the relative privacy-accuracy trade-off as a ratio which we seek to maximize.

4 Datasets and Learning Tasks

We consider four datasets for experiments. The datasets have been used in related work on MI and face recognition. The reference datasets are mostly unbalanced w.r.t. the amount of training data per training label, a characteristic that we found to benefit MI attacks. Each dataset is also summarized in Table 1 and the distributions for the two unbalanced datasets Texas Hospital Stays and Purchases Shopping Carts are provided in the appendix.

Texas Hospital Stays. The Texas Hospital Stays dataset [38] is an unbalanced dataset and consists of high dimensional binary vectors representing patient health features. Each record within the dataset is labeled with a procedure. The learning task is to train a fully connected neural network for classification of patient features to a procedure and we do not try to re-identify a known individual, and fully comply with the data use agreement for the original public use data file. We train and evaluate models for a set of most common procedures $\mathbb{C} \in \{100, 150, 200, 300\}$. Depending on the number of procedures the dataset comprises $67,330 - 89,815$ records and $6,170 - 6,382$ features. To allow comparison with related work [31,38], we train and test the target model on $n = 10,000$ records respectively.

Purchases Shopping Carts. This dataset is also unbalanced and consists of binary vectors with 600 features that represent customer shopping carts [38]. However, a significant difference to the Texas Hospital Stays dataset is that the number of features is almost 90% lower. Each vector is labeled with a customer group. The learning task is to classify shopping carts to customer groups by using a fully connected neural network. The dataset is provided in four variations with varying numbers of labels $\mathbb{C} \in \{10, 20, 50, 100\}$ and comprises $38,551 - 197,324$ records. We sample $n = 8,000$ records each for training and testing the target model. Again, this methodology ensures comparability with related work [31,38].

Labeled Faces in the Wild. The Labeled Faces in the Wild (LFW) dataset contains labeled images each depicting a specific person with a resolution of 250×250 pixels (i.e., features) [22]. The dataset has a long distribution tail w.r.t. to the number of images per label. We thus focus on learning the topmost classes $\mathbb{C} \in \{20, 50, 100\}$ with 1906, 2773 and 3651 overall records respectively. We start our comparison of LDP and CDP from a pre-trained VGG-Very-Deep-16 CNN faces model [34] by keeping the convolutional core, exchanging the dense layer at the end of the model and training for LFW grayscale faces. For LDP, we apply differentially private image pixelization within neighborhood $m = \sqrt{250 \times 250}$ and avoid coarsening by setting $b = 1$. We transform all images to grayscale before LDP and CDP training.

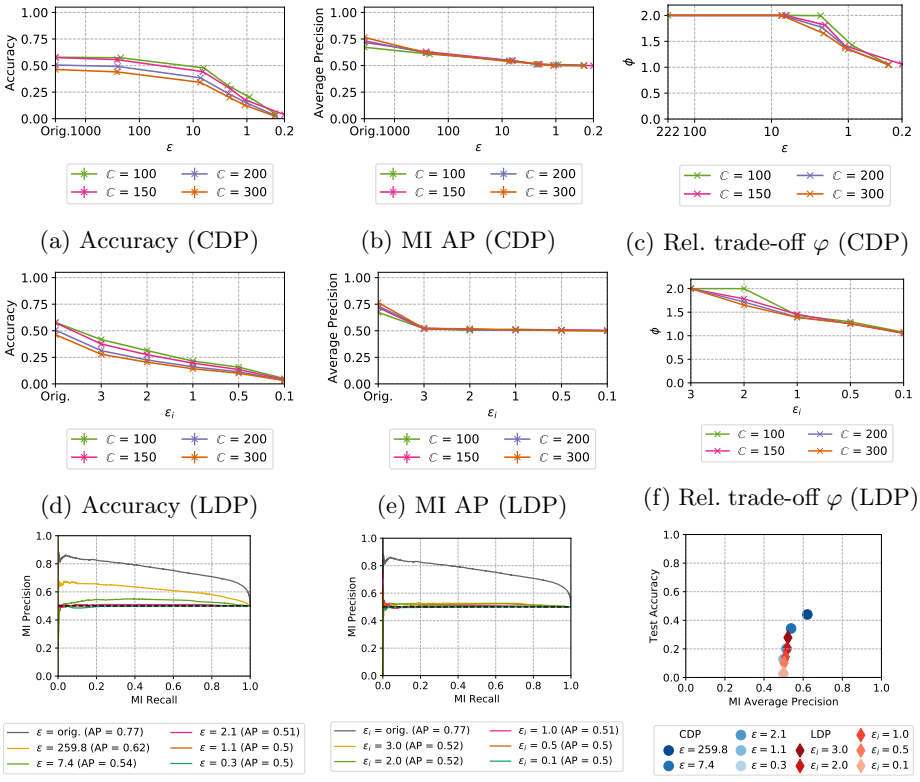
Skewed Purchases. We specifically crafted this balanced dataset[6] to mimic a transfer learning task, i.e., the application of a trained model to novel data which is similar to the training data w.r.t. format but following a different distribution. This situation arises for Purchases Shopping Carts, if for example not enough high-quality shopping cart data for a specific retailer are available yet. Thus, only few high-quality data (e.g., manually crafted examples) can be used for testing and large amounts of low quality data from potentially different distributions for training (e.g., from other retailers). In effect the distribution between train and test data varies for this dataset. Similar to Purchases Shopping Carts the dataset consists of $200,000$ records with 600 features and is analyzed for $\mathbb{C} \in \{10, 20, 50, 100\}$ labels. However, each vector x in the training dataset X is generated by using two independent random coins to sample a value from $\{0, 1\}$ per position $i = 1, \ldots, 600$. The first coin steers the probability $\Pr[x_i = 1]$ for a fraction of 600 positions per record x. We refer to these positions as indicator bits (*ind*) which indicate products frequently purchased together. The second coin steers the probability $\Pr[x_i = 1]$ for a fraction of $600 - (\frac{600}{|\mathbb{C}|})$ positions per record. We refer to these positions as noise bits (*noise*) that introduce scatter in addition to *ind*. We let $\Pr_{ind}[x_i = 1] = 0.8 \wedge \Pr_{noise}[x_i = 1] = 0.2, \forall x \in X_{train}$ and $\Pr_{ind}[x_i = 1] = 0.8 \wedge Pr_{noise}[x_i = 1] = 0.5, \wedge x \in X_{test}, 1 \leq i \leq 600$. The difference in information entropy between test and train data is ≈ 0.3.

5 Experiments

We perform an experiment which compares the privacy-accuracy trade-off for LDP and CDP by MI AP instead of privacy parameter ϵ per dataset. The results of each experiment are visualized by three sets of figures. First, we compare the relative privacy-accuracy trade-off φ resulting from test accuracy and MI AP over ϵ. We present this information for CDP per dataset in Figures 2 to 5 a,b,c and for LDP in Figures 2 to 5 d,e,f. The obtained information serves as basis to identify privacy parameters at which the MI AP is converging towards the baseline. Second, we state the precision-recall curves from which MI AP was calculated to illustrate the slope with which precision and recall are diverging from the baseline for LDP and CDP in Figures 2 to 5 g,h. Third, we compare the absolute privacy-accuracy trade-offs per dataset for both LDP and CDP in a scatterplot. We present this information in Figures 2 to 5i. For each dataset the model training stops once the test data loss is stagnating (i.e., early stopping) or a maximum number of epochs is reached. This design avoids excessive overfitting and increases real-world relevance. For all executions of the experiment CDP noise is sampled from a Gaussian distribution (cf. Definition 3) with scale $\sigma =$ *noise multiplier* $z \times$ *clipping norm* $\mathcal{C}$. We evaluate increasing noise regimes per dataset by evaluating noise multipliers $z \in \{0.5, 2, 4, 6, 16\}$ and calculate the resulting ϵ at a fixed $\delta = \frac{1}{n}$. However, since batch size, dataset size and number

[6] We provide this dataset along with all evaluation code on GitHub: `https://github.com/SAP-samples/security-research-membership-inference-and-differential-privacy`

of epochs are also influencing the Renyi differential privacy accounting a fixed z will inevitably result in different composed ϵ for different datasets. For LDP we use the same hyperparameters as in the original training and evaluate two local randomizers, namely randomized response and LDP image pixelization with the Laplace mechanism. For each randomizer we state the individual ϵ_i per invocation (i.e., per anonymized value). We apply randomized response to all datasets except LFW with a range of privacy parameter values $\epsilon_i \in \{0.1, 0.5, 1, 2, 3\}$ that reflect retention probabilities ρ from 5% – 90% (cf. Equation (1)). For LFW each pixel is perturbed with Laplace noise, and also investigate a wide range of resulting noise regimes by varying ϵ_i.



(a) Accuracy (CDP) (b) MI AP (CDP) (c) Rel. trade-off φ (CDP)

(d) Accuracy (LDP) (e) MI AP (LDP) (f) Rel. trade-off φ (LDP)

(g) PR curve $\mathbb{C} = 300$ (CDP) (h) PR curve $\mathbb{C} = 300$ (LDP) (i) Abs. trade-off $\mathbb{C} = 300$

Fig. 2: Texas Hospital Stays accuracy and privacy (error bars lie within points)

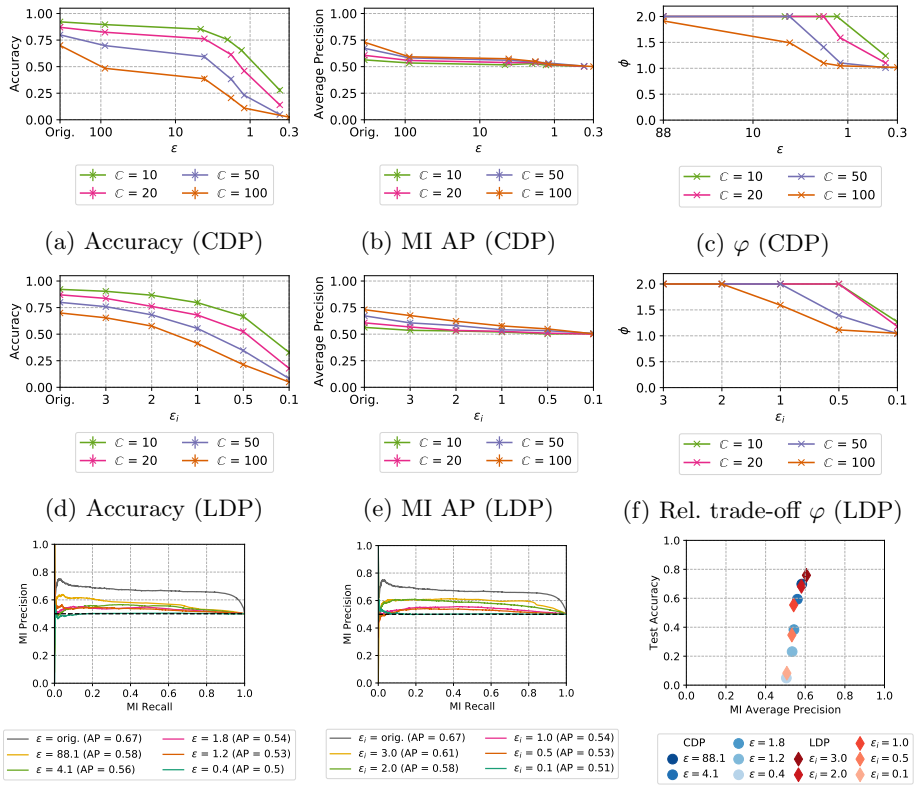
For sake of completeness we provide the resulting overall privacy parameters ϵ, z, hyperparameters and train accuracies for all datasets for LDP and CDP in Table 1 and 2 in the appendix. The experiment is repeated five times per

dataset to stabilize measurements and we report mean values with error bars unless otherwise stated. Precision-recall curves depict all experiment data.

Texas Hospital Stays. For Texas Hospital Stays we observe that LDP and CDP are achieving very similar privacy-accuracy trade-offs under MI. The main difference in LDP and CDP is observable in a smoother decrease of target model accuracy for CDP in contrast to LDP, which are depicted in Figures 2a and 2d. The smoother decay also manifests in a slower drop of MI AP for CDP in comparison to LDP as stated in Figures 2b and 2e. Texas Hospital Stays represents an unbalanced high dimensional dataset and both factors foster MI. However, the increase in dataset imbalance by increasing $\mathbb{C}$ is negligible w.r.t. MI AP. The relative privacy-accuracy trade-off for LDP and CDP is also close and for example the baseline MI AP of 0.5 is reached at $\varphi \approx 1.5$, as depicted in Figures 2c and 2f. In the example case of $\mathbb{C} = 300$ $\mathcal{DO}$ might prefer to use CDP, since the space of achievable MI APs in LDP is narrow while CDP also yields APs in between original and baseline as illustrated in the precision-recall curves in Figures 2g and 2h, and the scatterplot in Figure 2i. This observation is similar, though weaker, for all other $\mathbb{C}$.

Purchases Shopping Carts. CDP and LDP are achieving similar target model test accuracies on the Purchases dataset as depicted in Figures 3a and 3d. However, LDP is allowing a slightly smoother decrease in test accuracy over ϵ. Figure 3b illustrates that the CDP MI AP is somewhat resistant to noise and remains above 0.5 until a small $\epsilon \approx 1$. The LDP MI APs are significantly higher and decrease slower to the baseline as depicted by Figure 3e. A comparison of the relative privacy-accuracy trade-offs φ in Figures 3c and 3f underlines that CDP and LDP achieve similar trade-offs and LDP allows for smoother drops in the MI AP in contrast to CDP. Thus, LDP is the preferred choice for this dataset, if $\mathcal{DO}$ desires to lower the MI AP to a level *between* original and baseline. This is illustrated for example for $\mathbb{C} = 50$ in the precision-recall curves in Figures 3g, 3h and the scatterplots in Figure 3i. It is noticeable that while the overall ϵ for LDP and CDP differs by a magnitude of up to 10 times the relative and absolute privacy-accuracy trade-offs are close to each other. The observations also hold for other $\mathbb{C}$.

LFW. For LFW the target model reference architecture converges for both CDP and LDP towards the same test accuracy, which is reflecting the majority class. However, the target model test accuracy decay over ϵ is much smoother for CDP when comparing Figures 4a and 4d. Furthermore, the structural changes caused by LDP image pixelization seem to lead to quicker losses in test accuracy. W.r.t. the relative privacy-accuracy trade-off φ in Figures 4c and 4f CDP outperforms LDP. At MI AP = 0.5 CDP achieves $\varphi \approx 1.5$ for all $\mathbb{C}$ while LDP yields $\varphi \approx 1.1$ for all $\mathbb{C}$. The $\varphi = 0$ observed at $\epsilon_i = 10000$ for $\mathbb{C} = 100$ is due to an actual increase in AP that is comparatively larger than the decrease in test accuracy. The exemplary precision-recall curves for $\mathbb{C} = 50$ in Figures 4g and 4h furthermore illustrate that CDP can already have a large effect on MI AP at



(a) Accuracy (CDP) (b) MI AP (CDP) (c) φ (CDP)

(d) Accuracy (LDP) (e) MI AP (LDP) (f) Rel. trade-off φ (LDP)

(g) PR curve $\mathbb{C} = 50$ (CDP) (h) PR curve $\mathbb{C} = 50$ (LDP) (i) Abs. trade-off $\mathbb{C} = 50$

Fig. 3: Purchases accuracy and privacy (error bars lie within points)

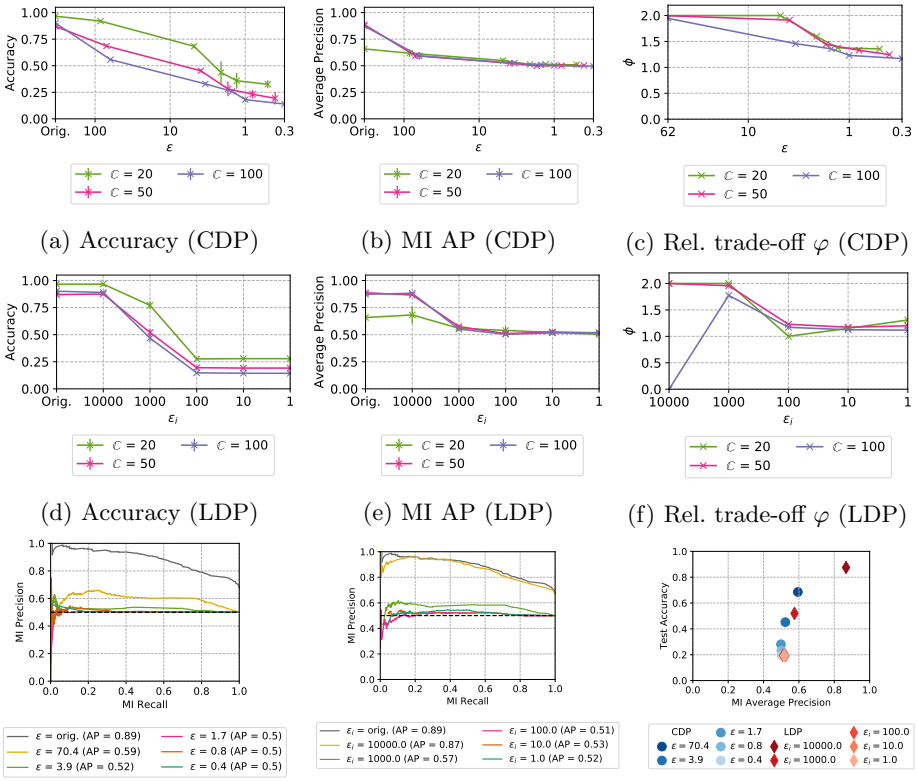
high ϵ. In addition, we observe from Figure 4i that CDP realizes a strictly better absolute privacy-accuracy trade-off under MI for $\mathbb{C} = 50$.

Skewed Purchases. The effects of dimensionality and imbalance of a dataset on MI have been addressed by related work [31,38]. However, the effect of a domain gap between training and test data which is found in transfer learning when insufficient high-quality data for training is initially available and reference data that potentially follows a different distribution has not been addressed. For this task we consider the Skewed Purchases dataset. Figures 5a and 5d show that the LDP test accuracy is in fact only decreasing at very small ϵ_i whereas CDP again gradually decreases over ϵ. This leads to a consistently higher test accuracy in comparison to CDP. W.r.t. the relative privacy-accuracy trade-off LDP outperforms CDP as depicted by φ in Figures 5c and 5f. However, we observe several outliers. Most notably for CDP, the MI AP decreases for $\mathtt{C} = 100$ and large ϵ values, but increases for small ϵ as shown in Figure 5b. This is a consequence of the target model resorting to random guessing for test records. Similarly, for

LDP the MI AP for $\mathbb{C} \in \{10, 100\}$ first decreases before recovering again as depicted in Figure 5e. We reason about the cause of these outliers by analyzing the target model decisive confidence values. LDP generalizes the training data towards the test data, however, at $\epsilon_i = 1.0$ LDP leads to nearly indistinguishable test and train distributions. Thus, the decisive softmax confidence of the target model increases in comparison to smaller and larger ϵ_i. For $\mathbb{C} = 10$ the absolute privacy-accuracy trade-off is also favorable for LDP as depicted in Figure 5i.

6 Discussion

Privacy parameter ϵ alone is unsuited to compare and select and compare DP mechanisms. We consistently observed that while the theoretic upper bound on inference risk reflected by ϵ in LDP is higher by a factor of hundreds or even thousands in comparison to CDP, the practical protection against a white-box MI attack is actually not considerably weaker at similar model accuracy.

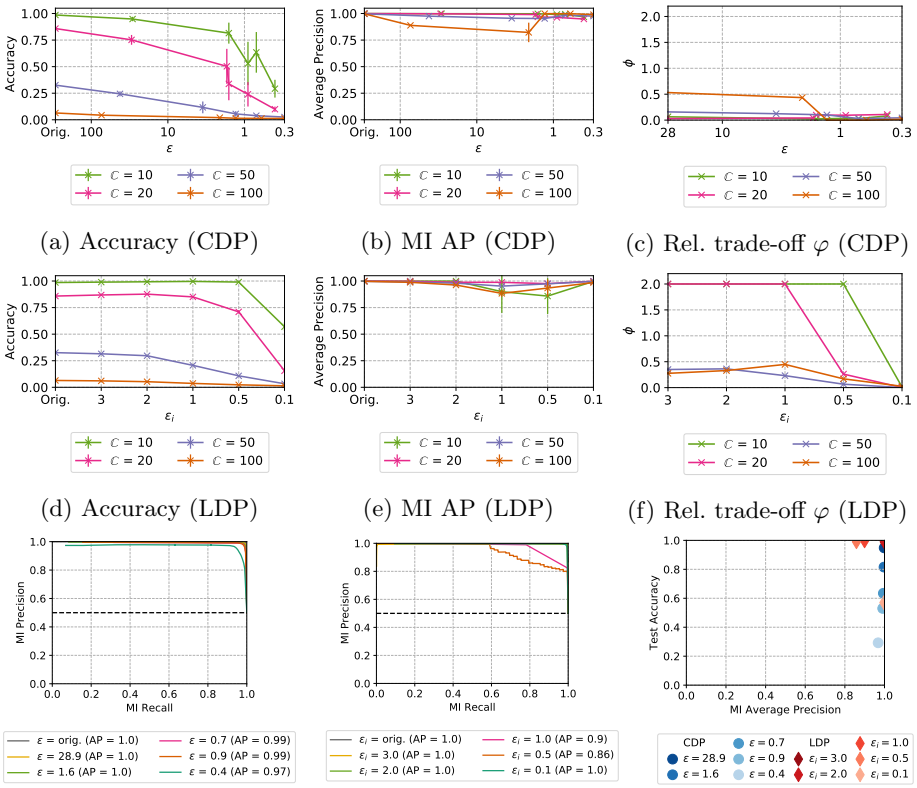


(a) Accuracy (CDP) (b) MI AP (CDP) (c) Rel. trade-off φ (CDP)

(d) Accuracy (LDP) (e) MI AP (LDP) (f) Rel. trade-off φ (LDP)

(g) PR curve $\mathbb{C} = 50$ (CDP) (h) PR curve $\mathbb{C} = 50$ (LDP) (i) Abs. trade-off $\mathbb{C} = 50$

Fig. 4: LFW accuracy and privacy (error bars lie within points)



(a) Accuracy (CDP) (b) MI AP (CDP) (c) Rel. trade-off φ (CDP)

(d) Accuracy (LDP) (e) MI AP (LDP) (f) Rel. trade-off φ (LDP)

(g) PR curve $\mathbb{C} = 10$ (CDP) (h) PR curve $\mathbb{C} = 10$ (LDP) (i) Abs. trade-off $\mathbb{C} = 10$

Fig. 5: Skewed Purchases accuracy and privacy (error bars lie within points)

For Texas Hospital Stays LDP mitigates white-box MI at an overall $\epsilon = 6382$ whereas CDP lies between $\epsilon = 0.9$ for $\mathbb{C} = 100$ and $\epsilon = 0.3$ for $\mathbb{C} = 300$. This observation at the baseline AP also holds for Purchases Shopping Carts where LDP $\epsilon = 60$ and CDP is between $\epsilon = 0.4$ for $\mathbb{C} = 10$ and $\epsilon = 0.3$ for $\mathbb{C} = 100$, and LFW (LDP $\epsilon = 62.5 \times 10^2$, CDP $\epsilon = 2.1$ to $\epsilon = 1.5$). Thus, we note that assessing privacy solely based on ϵ falls short. Given the results of the previous sections we rather encourage data scientists to also quantify privacy under an empirical attack such as white-box MI in addition to ϵ.

LDP and CDP result in similar privacy-accuracy curves. A wide range of privacy regimes in CDP and LDP can be compared with our methodology under MI. We observed for most datasets that similar privacy-accuracy combinations are obtained for well generalizing models (i.e., use of early stopping against excessive overfitting) that were trained with LDP or CDP. We also ran the experiments with black-box MI (i.e., only model outputs) and observed that the additional assumptions made by white-box MI (e.g., access to internal gra-

dient and loss information) only yield a small increase in AP (3 – 5%). The privacy-accuracy scatterplots depict that LDP and CDP formulate very similar privacy-accuracy trade-offs for Purchases Shopping Carts, LFW and Texas Hospital Stays. At two occasions on the smaller classification tasks Purchases Shopping Carts $\mathbb{C} = \{10, 20\}$ and Skewed Purchases $\mathbb{C} = \{10, 20\}$ LDP realizes a strictly better privacy-accuracy trade-off w.r.t. the practical inference risk. These observations lead us to conclude that LDP is an alternative to CDP for differentially private deep learning on binary and image data, since the privacy-accuracy trade-off is often similar at the same model accuracy despite the significantly larger ϵ. Thus, data scientists should consider to use LDP especially when required to use optimizers without CDP implementations or when training ensembles (i.e., multiple models over one dataset), since the privacy loss will accumulate over all ensemble target models when assuming that training data is reused between ensemble models. Here, we see one architectural benefit of LDP: flexibility. LDP training data can be used for all ensemble models without increasing the privacy loss in contrast to CDP.

The relative privacy-accuracy trade-off is favorable within a small interval. We observed that the privacy-accuracy trade-off as visualized in the scatterplots throughout this work allows to identify whether CDP or LDP achieve better test accuracy at similar APs. However, the scatterplots do not reflect whether target model test accuracy is decreasing slower, similar or stronger than MI AP decreases over the privacy parameter ϵ. For this purpose we introduced φ. We found that φ allows to identify ϵ intervals in which the AP loss is stronger than the test accuracy loss for all datasets. On the high dimensional datasets Texas Hospital Stays and LFW CDP consistently achieves higher φ than LDP. In contrast, φ values are similar for LDP and CDP on Purchases, and superior for LDP on Skewed Purchases.

7 Related Work

Our work is related to DP in neural networks, attacks against the confidentiality of training data and performance benchmarking of neural networks.

CDP is a common approach to realize differentially private neural networks by adding noise to the gradients during model training. Fundamental approaches for perturbation with the differentially private gradient descent (DP-SGD) during model training were provided by Song et al. [39], Bassily et al. [4] and Shokri et al. [37]. Abadi et al. [1] formulated the DP-SGD that was used in this work. Mironov [29] introduces Renyi DP for measuring the DP-SGD privacy loss over composition. Iyengar et al. [23] suggest a hyperparameter free algorithm for differentially private convex optimization for standard optimizers.

Fredrikson et al. [15,16] formulate model inversion attacks that use target model softmax confidence values to reconstruct training data per class. In contrast, MI attacks address the threat of identifying individual records in a dataset [3,36]. Yeom et al. [44] have demonstrated that the upper bound on

MI risk for CDP can be converted into an expected bound for MI Advantage. We state MI precision and recall, arguing that `in` is the sensitive information. Jaymaran and Evans [24] showed that the theoretic MI upper bound and the achievable MI lower bound are far apart in CDP. We observe, that LDP can be an alternative to CDP as the upper and lower bounds are even farther apart from each other. Shokri et al. [32] formulate an optimal mitigation against their MI attack [38] by using adversarial regularization. By applying the MI attack gain as a regularization term to the objective function of the target model, a non-leaking behavior is enforced w.r.t. MI. While their approach protects against their MI adversary, DP mitigates any adversary with arbitrary background information. Carlini et al. [6] suggest *exposure* as a metric to measure the extent to which neural networks memorize sensitive information. Similar to our work, they apply DP for mitigation. We focus on attacks against machine learning models targeting identification of members of the training dataset. Abowd and Schmutte [2] describe an economic social choice framework to choose privacy parameter ϵ. We compare LDP and CDP mechanisms aside from ϵ. Rahman et al. [35] applied a black-box MI attack against DP-SGD models on CIFAR-10 and MNIST. They evaluate the severity of MI attack by the F1-score which results in numerically higher scores, but assumes `out` labels to be sensitive.

MLPERF [30] and DPBench [19] are frameworks for machine learning performance measurements and evaluation of DP. We focus on comparing the privacy-utility trade-off and apply the core principles of both benchmarks.

8 Conclusion

We compared LDP and CDP mechanisms for differentially private deep learning under a white-box MI attack. The comparison comprises the average precision of the MI precision-recall curve and the target model test accuracy to support data scientists in choosing among available DP mechanisms and selecting privacy parameter ϵ. Our experiments on diverse learning tasks show that neither LDP nor CDP yields a consistently better privacy-accuracy trade-off. While MI only yields a lower bound on MI whereas ϵ in DP yields an upper bound, we observed that the lower bounds for LDP and CDP are close at similar model accuracy despite large difference in their upper bound. This suggests that the upper bound is far from the practical susceptibility to MI attacks and that data scientists should also consider to apply LDP despite the large privacy parameter values. Especially, since LDP does not require privacy accounting when training multiple models and offers flexibility w.r.t. optimizers. We consider the relative privacy-accuracy trade-off for LDP and CDP as the ratio of losses in accuracy and privacy over ϵ, and show that it is only favorable within a small interval.

Acknowledgements. This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 825333 (MOSAICROWN).

Appendix

Neural network models and composed ϵ for LDP are provided in Table 1. We state hyperparameters, composed ϵ for CDP, and training accuracies in Table 2. Texas Hospitals Stays and Purchases Shopping Carts provided by Shokri et al. are unbalanced in terms of records per class, as shown in Figures 6 and 7.

Table 1: Overview of datasets considered in evaluation.

Dataset	**Model**	**LDP**
Texas Hospital Stays [38]	Fully connected NN with three layers ($512 \times 128 \times \mathbb{C}$) [38].	$19,125 - 638$ ($6382 \times \epsilon_i$)
Purchases Shopping Carts [38]	Fully connected NN with two layers ($128 \times \mathbb{C}$) [38] (i.e., logistic regression).	$1800 - 60$ ($600 \times \epsilon_i$)
Labeled Faces in the Wild [22]	VGG-Very-Deep-16 CNN [34]	$62.5 \times 10^6 - 6,250$ ($250 \times 250 \times \epsilon_i$)
Skewed Purchases	Fully connected NN with two layers ($128 \times \mathbb{C}$) [38] (i.e., logistic regression).	$1,800 - 60$ ($600 \times eps_i$)

Table 2: Target Model training accuracy (from orig. to smallest ϵ), CDP ϵ values (from $z = 0.5$ to $z = 16$) and hyperparameters

		Texas Hospital Stays				Purchases Shopping Carts				LFW			Skewed Purchases			
$\mathbb{C}$		100	150	200	300	10	20	50	100	20	50	100	10	20	50	100
LDP		0.86	0.92	0.83	0.81	0.99	1.0	1.0	0.99	1.0	1.0	1.0	1.0	1.0	1.0	1.0
		1.0	1.0	1.0	1.0	0.97	0.97	0.95	0.94	1.0	1.0	1.0	1.0	1.0	1.0	0.99
		1.0	1.0	1.0	1.0	0.88	0.85	0.86	0.90	1.0	0.96	1.0	1.0	1.0	1.0	0.97
		1.0	1.0	0.98	0.92	0.64	0.58	0.69	0.79	0.22	0.18	0.13	1.0	0.99	0.97	0.89
		0.99	0.95	0.86	0.72	0.58	0.47	0.62	0.75	0.24	0.17	0.13	0.93	0.98	0.9	0.80
		0.82	0.71	0.59	0.53	0.44	0.38	0.49	0.51	0.25	0.17	0.13	0.52	0.55	0.71	0.45
CDP		0.86	0.92	0.83	0.81	1.0	1.0	1.0	0.99	1.0	1.0	1.0	1.0	1.0	1.0	1.0
		0.74	0.75	0.69	0.62	0.95	0.91	0.82	0.63	0.99	0.87	0.79	1.0	1.0	0.97	0.58
		0.57	0.54	0.48	0.42	0.91	0.84	0.71	0.51	0.76	0.5	0.35	1.0	0.96	0.6	0.1
		0.35	0.31	0.26	0.22	0.80	0.69	0.46	0.27	0.44	0.28	0.25	0.92	0.8	0.25	0.02
		0.22	0.19	0.16	0.13	0.69	0.51	0.28	0.14	0.36	0.23	0.18	0.89	0.64	0.12	0.02
		0.05	0.04	0.03	0.02	0.28	0.14	0.05	0.02	0.32	0.19	0.13	0.66	0.24	0.03	0.01
ϵ		222.6	259.8	251.5	259.8	88.1	88.1	88.1	88.1	84.3	70.4	62.4	28.9	29.8	42.2	73.5
		6.3	6.6	7.3	7.4	4.6	4.1	4.1	4.1	4.8	3.9	3.4	1.6	1.7	3.5	2.1
		2.3	2.0	2.2	2.1	2.0	1.8	1.8	1.8	2.1	1.7	1.5	0.7	1.6	1.3	1.3
		0.9	1.1	1.0	1.1	1.3	1.2	1.2	1.2	1.3	0.8	1.0	0.9	0.9	0.7	0.6
		0.3	0.2	0.3	0.3	0.4	0.4	0.4	0.3	0.5	0.4	0.3	0.4	0.4	0.3	0.3
Learning rate	Orig.	0.01	0.01	0.01	0.01	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	CDP	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.001	0.0008	0.0008	0.001	0.001	0.001	0.001
	LDP	0.01	0.01	0.01	0.01	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Batch size	Orig.	128	128	128	128	128	128	128	128	32	32	32	100	100	100	100
	CDP	128	128	128	128	128	128	128	128	16	16	16	100	100	100	100
	LDP	128	128	128	128	128	128	128	128	32	32	32	100	100	100	100
Epochs	Orig.	200	200	200	200	200	200	200	200	30	30	30	200	200	200	200
	CDP	1000	1000	1000	1000	200	200	200	200	110	110	110	200	200	200	200
	LDP	200	200	200	200	200	200	200	200	30	30	30	200	200	200	200
Clipping Norm	CDP	4	4	4	4	4	4	4	4	3	3	3	4	4	4	4

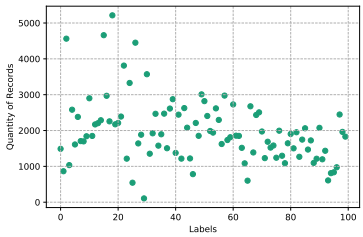
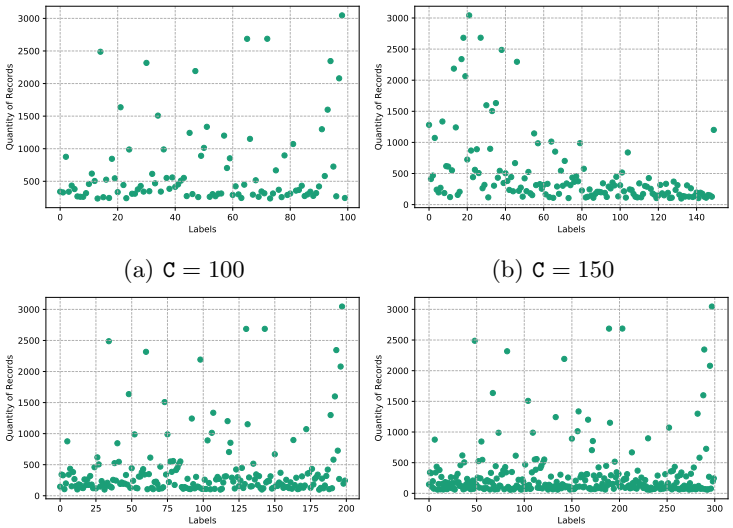


Fig. 6: Quantity of records per label for Purchases Shopping Carts



(a) $\mathtt{C} = 100$ (b) $\mathtt{C} = 150$

(c) $\mathtt{C} = 200$ (d) $\mathtt{C} = 300$

Fig. 7: The Quantity of records per Label for the Texas Hospital Stays Dataset

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep Learning with Differential Privacy. In: Proc. of Conference on Computer and Communications Security (CCS). ACM Press (2016)
2. Abowd, J.M., Schmutte, I.M.: An economic analysis of privacy protection and statistical accuracy as social choices. American Economic Review **109**(1) (2019)
3. Backes, M., Berrang, P., Humbert, M., Manoharan, P.: Membership Privacy in MicroRNA-based Studies. In: Proc. of Conference on Computer and Communications Security (CCS). ACM Press (2016)
4. Bassily, R., Smith, A., Thakurta, A.: Private Empirical Risk Minimization. In: Proc. of Symposium on Foundations of Computer Science (FOCS). IEEE Computer Society (2014)
5. BBC News: Google DeepMind NHS app test broke UK privacy law (2017), `https://www.bbc.com/news/technology-40483202`

6. Carlini, N., Liu, C., Kos, J., Erlingsson, Ú., Song, D.: The secret sharer: Measuring unintended neural network memorization and extracting secrets (2018)
7. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proc. of Conference on Machine Learning (ICML). Omnipress (2006)
8. Dwork, C.: Differential Privacy. In: Proc. of Colloquium on Automata, Languages and Programming (ICALP). Springer (2006)
9. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our Data, Ourselves: Privacy via distributed noise generation. In: Proc. of Conference on Theory and Applications of Cryptographic Techniques (EUROCRYPT). Springer (2006)
10. Dwork, C., Roth, A.: The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science **9**(3-4) (2014)
11. Erlingsson, U., Feldman, V., Mironov, I., Raghunathan, A., Talwar, K., Thakurta, A.: Amplification by shuffling: From local to central differential privacy via anonymity. In: Proc. of Symp. on Discrete Algorithms (SODA) (2019)
12. Erlingsson, U., Pihur, V., Korolova, A.: RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In: Proc. of Conference on Computer and Communications Security (CCS). ACM Press (2014)
13. Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge. International Journal of Computer Vision **88**(2) (2010)
14. Fan, L.: Image pixelization with differential privacy. In: Proc. of Conference on Data and Applications Security and Privacy (DBSEC). Springer (2018)
15. Fredrikson, M., Jha, S., Ristenpart, T.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In: Proc. of Conference on Computer and Communications Security (CCS). ACM Press (2015)
16. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T.: Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In: Proc. of USENIX Security Symposium. USENIX Association (2014)
17. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), `http://www.deeplearningbook.org`
18. Grandvalet, Y., Canu, S.: Comments on "Noise injection into inputs in back propagation learning". IEEE Transactions on Systems, Man, and Cybernetics **25**(4) (1995)
19. Hay, M., Machanavajjhala, A., Miklau, G., Chen, Y., Zhang, D.: Principled evaluation of differentially private algorithms using dpbench. In: Proc. of Conference on Management of Data (SIGMOD). ACM Press (2016)
20. Hayes, J., Melis, L., Danezis, G., De Cristofaro, E.: LOGAN: Membership Inference Attacks Against Generative Models. Proc. on Privacy Enhancing Technologies (PoPETs) **2019**(1) (2019)
21. Hill, Kashmir: How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did (2012), `https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/`
22. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., University of Massachusetts (2007)
23. Iyengar, R., Near, J.P., Song, D., Thakkar, O.D., Thakurta, A., Wang, L.: Towards practical differentially private convex optimization. In: Proc. of Symposium on Security and Privacy (S&P). IEEE Computer Society (2019)
24. Jayaraman, B., Evans, D.: Evaluating differentially private machine learning in practice. In: Proc. of the USENIX Security Symposium. USENIX Association (2019)

25. Kairouz, P., Oh, S., Viswanath, P.: The Composition Theorem for Differential Privacy. IEEE Transactions on Information Theory **63**(6) (2017)
26. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.: What can we learn privately? SIAM Journal on Computing **40** (2008)
27. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. ICLR (2015)
28. Matsuoka, K.: Noise injection into inputs in back-propagation learning. IEEE Transactions on Systems, Man, and Cybernetics **22**(3) (1992)
29. Mironov, I.: Rényi differential privacy. In: Proc. of Computer Security Foundations Symposium (CSF). IEEE Computer Society (2017)
30. MLPerf Website: MLPerf – Fair and useful benchmarks for measuring training and inference performance of ML hardware, software, and services (2018), `https://mlperf.org/`
31. Nasr, M., Shokri, R., Houmansadr, A.: Comprehensive Privacy Analysis of Deep Learning: Stand-alone and Federated Learning under Passive and Active White-box Inference Attacks (2018)
32. Nasr, M., Shokri, R., Houmansadr, A.: Machine learning with membership privacy using adversarial regularization. In: Proc. of Conference on Computer and Communications Security (CCS). ACM Press (2018)
33. Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Úlfar Erlingsson: Scalable private learning with pate (2018)
34. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference. BMVA Press (2015)
35. Rahman, M.A., Rahman, T., Laganière, R., Mohammed, N.: Membership inference attack against differentially private deep learning model. Transactions on Data Privacy **11** (2018)
36. Sankararaman, S., Obozinski, G., Jordan, M.I., Halperin, E.: Genomic privacy and limits of individual detection in a pool. Nature Genetics **41** (2009)
37. Shokri, R., Shmatikov, V.: Privacy-preserving Deep Learning. In: Proc. of Conference on Computer and Communications Security (CCS). ACM Press (2015)
38. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against ML models. In: Proc. of Symposium on Security and Privacy (S&P). IEEE Computer Society (2017)
39. Song, S., Chaudhuri, K., Sarwate, A.D.: Stochastic gradient descent with differentially private updates. In: Proc. of Conference on Signal and Information Processing. IEEE Computer Society (2013)
40. Wang, T., Blocki, J., Li, N., Jha, S.: Locally Differentially Private Protocols for Frequency Estimation. In: Proc. of USENIX Security Symposium. USENIX Association (2017)
41. Warner, S.L.: Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association **60**(309) (1965)
42. Wirth, R., Hipp, J.: Crisp-dm: Towards a standard process model for data mining. In: Proc. of Conference on practical applications of knowledge discovery and data mining. Practical Application Company (2000)
43. Yeom, S., Fredrikson, M., Jha, S.: The unintended consequences of overfitting: Training data inference attacks (2017)
44. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting (2018)