Does Differential Privacy Prevent Backdoor Attacks in Practice? *

Fereshteh Razmi[1], Jian Lou[2], and Li Xiong[1]

[1] Emory University, Atlanta GA 30322, USA
[2] Zhejiang University, Hangzhou, Zhejiang 310027, China
{frazmim,lxiong}@emory.edu,
jian.lou@zju.edu.cn

Abstract. Differential Privacy (DP) was originally developed to protect privacy. However, it has recently been utilized to secure machine learning (ML) models from poisoning attacks, with DP-SGD receiving substantial attention. Nevertheless, a thorough investigation is required to assess the effectiveness of different DP techniques in preventing backdoor attacks in practice. In this paper, we investigate the effectiveness of DP-SGD and, for the first time, examine PATE and Label-DP in the context of backdoor attacks. We also explore the role of different components of DP algorithms in defending against backdoor attacks and will show that PATE is effective against these attacks due to the bagging structure of the teacher models it employs. Our experiments reveal that hyper-parameters and the number of backdoors in the training dataset impact the success of DP algorithms. We also conclude that while Label-DP algorithms generally offer weaker privacy protection, accurate hyper-parameter tuning can make them more effective than DP methods in defending against backdoor attacks while maintaining model accuracy.

Keywords: Differential Privacy · Backdoor Attack · Security

1 Introduction

Deep neural networks are vulnerable to backdoor attacks. The goal of a backdoor adversary is to misclassify the prediction of the target model on samples that contain a special pattern (trigger), while maintaining the inference performance on normal samples. To achieve this goal, backdoor attacks typically manipulate a small portion of training data with carefully designed triggers that lead to the mismatch between training features and labels [19]. Many studies have proposed countermeasures against this powerful attack or the more general data poisoning attacks. The most common approach in these studies is discovering abnormalities in model statistics or training data [7,8,31,35,36].

Differential privacy (DP) [12] is a fundamental concept of data privacy, guaranteeing that the inclusion or exclusion of individual data points doesn't significantly impact the outcome of any analysis. A common method to achieve DP in a

* This work was funded by National Institutes of Health (NIH) R01LM013712, and National Science Foundation (NSF) CNS-2124104, CNS-2125530.

deep learning model is by introducing calibrated randomness during the training process such as DP-SGD (Differentially Private Stochastic Gradient Descent), which adds noise to the gradients during the training. An alternative approach is PATE (Private Aggregation of Teacher Ensembles), which involves training multiple teacher models on disjoint subsets of the data and then using their aggregated outputs with added noise to train a student model with auxiliary data. Label differential privacy [14,34] is a variant of differential privacy that ensures that the learning process (and the resulting model) cannot reveal whether any individual's label was used or not. As the success of backdoor attacks relies on the influence of the triggered samples on the model, it is intuitive that the model might be more robust to backdoor attacks if the influence of each training sample is bounded. This concept of limiting the influence of individual samples aligns with the principles of DP. Thus a recent promising area of research focuses on using DP to build robust models against backdoor and poisoning attacks. This is accomplished by introducing randomness to the model through DP techniques, making it less sensitive to input.

There are a few works exploring this area in theory [5,25]. A few others have obtained experimental results either under a centralized setting using DP-SGD [4,10,17,40] or under the federated learning setting [24,27,28]. These studies provide some evidence that models trained with DP-SGD mitigate poisoning attacks, but they fall short of a comprehensive investigation and do not explore the power of other state-of-the-art DP models against backdoor attacks.

This paper aims to bridge the theory and practice and provide a comprehensive and in-depth understanding of whether and, more importantly, how various DP models and methods defend against backdoor attacks in practice given the theoretical promise and preliminary evidence in the literature. We study both the standard DP class of algorithms and the Label-DP variant, and compare them in their defense power against backdoor attacks. PATE and Label-DP are being examined for the first time against these types of attacks. We evaluate their performance empirically on two widely used datasets in the domain of backdoor attacks and differential privacy. To summarize, we make the following contributions:

1. **Comparative study of DP approaches against backdoor attacks, including standard DP-SGD approach and the less-studied PATE approach.** Existing studies use DP-SGD for training DP models to defend against poisoning or backdoor attacks. In this work, we explore the other well-known DP algorithm PATE against backdoor attacks in order to understand whether different DP algorithms (gradient perturbation vs. aggregation perturbation) have different powers against backdoors. We show that both of these classical DP approaches can provide robust models for backdoor attacks. Also, we will demonstrate that the ensemble structure of the PATE inherently makes it suitable against backdoors.
2. **A deeper understanding of the impact of noise and other parameters of DP approaches on backdoor attacks.** The effectiveness of DP approaches is affected by parameters other than noise. We explore the ori-

gin of these algorithms' resilience by examining whether randomness is the sole factor or if the other parameters have an impact. We empirically show that the randomness (privacy budget) contributes to mitigating the backdoor attack success rate, which is compatible with the theoretical results in the literature [39]. However, we demonstrate that the impact of other parameters can be significant on the outcome, especially for PATE, e.g., the threshold used to aggregate the teacher models' outputs.

3. **Comparative study of Label-DP approaches against backdoor attacks.** Label-DP protects the privacy of the labels of the training data by ensuring the output model is indistinguishable with respect to the label of a training sample. We study the Label-DP class of algorithms for the first time against backdoor attacks using two algorithms ALIBI [26] and LP-2ST [14]. We hypothesize that Label-DP also provides robustness against backdoor attacks while maintaining better utility than DP based on two observations. First, since Label-DP ensures the indistinguishability of labels, we expect a model with Label-DP to break the association between the backdoor triggers and their assigned target class (label). Second, Label-DP methods typically converge faster than standard DP algorithms while maintaining higher model utility. This is because indistinguishability is required only for the labels, rather than for both the features and labels, which necessitates less noise to achieve the same level of privacy.
   Our evaluations confirm that Label-DP makes the model more immune to backdoor attacks while preserving model accuracy. We show that Label-DP is superior to DP approaches in terms of convergence speed. Furthermore, we demonstrate that it can achieve better robustness accuracy trade-offs under certain settings. For instance, with a lower percentage of backdoors, ALIBI can eliminate the negative impact of the attack while achieving the highest accuracy among all approaches. For stronger attacks with higher percentage of backdoors, LP-2ST outperforms other approaches when the privacy budget is low.

2 Preliminaries

2.1 Backdoor Attacks

Backdoor attacks are a category of attacks that involve attaching a small patch to a portion of a base class of the training dataset along with flipping their labels to a specified target class. After the model has been trained using these backdoor samples, it would be vulnerable to the presence of the patch in the inputs. So as the next step of the attack, the attacker attaches the same patch to some desired test samples of the base class and passes it to the backdoored model, so that this combination of the base class pattern plus the patch pattern misleads the model to misclassify the sample as the target class. This form of backdoor attacks, initially introduced by Gu et al. [15], is a powerful attack that has gained much attention. Some other works tried to make some other type

of backdoor attacks that are less detectable or employ them in other domains including videos [32,41].

2.2 Differential Privacy and Label Differential Privacy

Differential Privacy (DP). DP is a privacy-preserving notion that makes an observer unable to tell if particular information contributes to the outcome [13]. In the context of machine learning, a DP method should not reveal whether a training sample has been utilized in the training process.

Let X and Y be the feature and label domain, respectively. Also, let the training dataset consists of n samples from a domain $U = (X \times Y)_n$. Given sample x, we have a classification task for the model M to predict y. A randomized training algorithm $\mathcal{M} : U \rightarrow R$ is (ε, δ)-DP if for any two adjacent datasets $D, D' \in U$ differing on at most one sample, it holds that:

$$\forall S \subset R, P[M(D) \in S] \leq e^{\varepsilon} P[M(D') \in S] + \delta. \tag{1}$$

A smaller ε guarantees stronger privacy but typically leads to a lower utility or accuracy of the model due to the randomization in the training. Using a DP property called **group privacy**, this definition can be extended to two datasets differing in k examples where k denotes more than one data point [11]. This is achievable by a linear increase in the privacy cost.

Label Differential Privacy (Label-DP). Label DP is an extension of DP that considers the labels as the only sensitive part of the training data that requires to be kept secret. So in contrast to (ε, δ)-DP which defines privacy for datasets D and D' differing on at most one sample, (ε, δ)-Label-DP considers D and D' differing on **the label** of at most one sample. Therefore, Label-DP can be seen as a relaxation of DP algorithms that guarantees only the privacy of the labels. One of the applications of Label-DP is recommendation systems where the user's profile or search queries are public, but the history of the user rating is sensitive.

2.3 DP and Label-DP Algorithms for Deep Learning

In this section, we explore the main methods for achieving DP (DP-SGD, PATE) and Label-DP (LP-MST and ALIBI) respectively, with Table 1 showing the critical parameters of the first two algorithms.

DP-SGD [1] is the most widely used algorithm for building DP models. DP-SGD restricts the privacy loss in each iteration of SGD (Stochastic Gradient Descent), by updating model in two steps: 1) clipping the L2 norm of the gradients, and 2) inserting calibrated Gaussian noise into those clipped gradients.

Table 1. Parameters of the DP algorithms

Method	Parameters
DP-SGD	1. **Noise multiplier** : Added randomness to the model's clipped gradients to provide DP 2. **Upper bound of the clipping norm (*Cnorm*)** : Bound to clip the L2-norm of the gradients to control their sensitivity to the noise
PATE	1. **Threshold** T : Queries exceeding this minimum teachers' aggregation are selected for training the student model 2. **Selection noise with variance** σ_1 : Gaussian noise added to the aggregator's votes before applying threshold to enforce privacy 3. **Result noise with variance** σ_2 : Noise added to the selected queries after applying threshold to guarantee DP 4. **Number of teacher models** 5. **Number of queries**

PATE [29] provides privacy through a teacher-student structure. First, an ensemble of teachers is trained on disjoint subsets of the private data. Then, given an unlabeled public dataset, a student model queries the teacher ensemble and uses their noisy aggregated vote as the label. The number of queries is restricted. Plus, their response is based on a noisy aggregation without access to any specific private data point. However, access to a public dataset forces a strong assumption on PATE compared to DP-SGD.

PATE was originally introduced with Laplacian noise [29]. Then it was revised to improve the utility and privacy trade-off through a more confident aggregated teacher consensus, called Confident-GNMax [30]. In this paper, we adopt the Confident-GNMax version of the PATE framework, which is based on Gaussian noise.

Label Private Multi-Stage Training (LP-MST) [14] is a work regarding differential privacy that achieves Label-DP for deep learning. It leverages a modified version of the Randomized Response (RR) algorithm to add noise to the labels [38]. RR outputs the actual class of a sample or randomly replaces it with one of the other classes. However, the randomness deteriorates the utility.

Ghazi et al. [14] modify the RR algorithm to compensate for the utility, by iteratively training the model on disjoint subsets of the dataset. Then they use the trained model from the previous stage to get the top-K predictions and limit the RR algorithm to those predictions. Similar to the main paper, we report our results on LP-2ST with two training stages.

Additive Laplace Noise Coupled with Bayesian Inference (ALIBI) [26] is another Label-DP method in ML that has been recently proposed. It first adds

Laplacian noise to one-hot labels, then uses these soft new labels to train the model while preserving Label-DP. Since post-processing does not affect differential privacy, Bayesian post-processing de-noises the soft labels iteratively during each step of SGD. The combination of additive Laplacian noise and iterative Bayesian inference increases the utility.

3 Related Work

DP has recently been highlighted for providing robust models to alleviate the negative impact of poisoning attacks. The rationale is that according to the definition of DP and group privacy, DP models are less sensitive to the impact of one or a group of poisoned data. In this section, we go through the literature to investigate where and how differentially private approaches used to defend against backdoor and poisoning attacks. We then identify the gaps in the literature, formulate those as research questions, and try to answer them and assess the results empirically.

There are two lines of work in the literature that consider the defensive power of DP methods on poisoning attacks; theoretical and practical studies.

Ma et al. [25] theoretically prove the robustness of DP models and provide a theoretical bound. They assume a training dataset D and an attacker with full knowledge creates some poisoned dataset $\tilde{D}$ from D. The poisoned model $\theta_{\tilde{D},b}$ is parameterized through the poisoned data $\tilde{D}$ and noise parameter b of the DP model. The attacker's objective loss $C : \Theta \rightarrow R$ aims to misclassify some targets or disrupt the overall classifier's functionality. Assuming the attacker does not know the exact realization of the noise, the attack is reduced to:

$$\min_{\tilde{D}} \quad J(\tilde{D}) = E_b \left[C(\theta_{\tilde{D},b}) \right] \tag{2}$$

Given k poisoned data, the authors utilize the property of differential privacy in Equation (1) and conclude:

$$J(\tilde{D}) \geq e^{-sign(C).k\varepsilon} J(D) \tag{3}$$

According to Equation (3) the attacker is unable to change $J(\tilde{D})$ arbitrarily because it is lower bounded by 0 if C is positive (for example, in case of Mean Squared Error) or it is unbounded from below if C is negative.

This paper provides insight into how DP methods may provide a natural immunity against data poisoning attacks. However, it has two limitations. First, the lower bound of $J(\tilde{D})$ is loose. Second, this paper implements and evaluates its theoretical findings on general attack loss functions and DP frameworks. Thus, the specific impact of Equation (3) on SOTA deep learning models (e.g. DP-SGD) and practical attacks (e.g. backdoor attacks) remains neglected.

To overcome the second limitation, a parallel set of works has employed DP-SGD as a practical usage of DP in deep learning to achieve protection against poisoning attacks [10,40,4]. Hong et al. [17] was one of the first works that considered DP-SGD against backdoor and other poisoning attacks. However, their

primary motive was not originated from the fact that DP-SGD is a private algorithm and Equation (1). Instead, they observed that during the training on a poisoned dataset, the gradients computed on poisoned samples have a higher magnitude and different orientation than those computed on clean samples. Hence they leveraged DP-SGD to offset the behavior of the model's gradients on both clean and poisoned data through the randomness of the gradients. Their results show some degree of protection against specific poisoning attacks, but their outcome is not promising on backdoor (insertion) attacks. Later, Jagielski and Oprea claimed that DP itself can not serve as a defense against poisoning attacks [18]. They argued that it is possible that the robustness of DP-SGD stems from some parameters other than noise.

4 Research Questions

The existing studies on DP-SGD are inconclusive, and there are no studies on other state-of-the-art DP approaches as a potential defense. It motivates us to extend current works by conducting more comprehensive experiments on DP-SGD and introducing other DP methods as a defense. Based on this primary motivation, we pose some research questions in this section and elaborate their significance. Then in the following sections, we will try to address them empirically.

Question 1. *Is DP-SGD a successful protective algorithm against backdoor attacks? Can PATE, as another main DP approach, mitigate backdoor attacks?*
Current studies have differing views on whether DP, particularly DP-SGD, can defend against backdoor attacks. It opens the door for a more comprehensive study of DP-SGD. It's not clear whether the robustness is achieved by the randomization introduced by DP methods in general or by other algorithm-specific parameters of DP-SGD. Additionally, this outcome can emphasize the gap between DP's theoretical and practical results against poisoning data.

So in this work, we first explore DP-SGD to understand why there is no consensus in the literature on DP-SGD as a defensive algorithm. Then for the first time, we explore PATE as a DP method against backdoor attacks to demonstrate if it confirms DP models' robustness. We examine the effectiveness of these algorithms by analyzing their hyper-parameters, even those that do not contribute to the randomness for DP. With this investigation, we hope to determine whether these algorithms are effective defense mechanism solely because they are DP.

Question 2. *Can other DP notions, such as Label-DP, also provide robustness and even better accuracy and robustness trade-off? How do different DP notions and algorithms compare in the trade-off?*
Answering the research question 1, leads us to two other major challenges with regard to DP-SGD and PATE. The first challenge is their prohibitive training time. Training an ensemble of teachers in PATE is heavily costly. Also, DP-SGD requires computation of per-sample gradient norms, which is extremely slow.

The other issue with the DP algorithms is the trade-off between the privacy budget and the utility, which means decreasing the privacy budget (i.e., achieving stronger DP) is accompanied by a drop in models' accuracy. We will show that lower privacy budgets usually lead to a lower attack success rate (ASR), which is necessary to defeat attacks. We call this simultaneous reduction in accuracy and ASR the *Accuracy-ASR trade-off*. We will define the criteria for attack success rate in Section 5. To address these challenges, we conduct a comparison between Label-DP and other DP algorithms by varying DP budgets and attack strengths.

5 Experimental Setup

Datasets and Models We evaluate each DP model on two datasets: MNIST [23] and CIFAR-10 [22]. We study end-to-end training and fine-tuning since both are common practices in modern machine learning. We use the same CNN architecture as [2] with two convolutional layers for MNIST and train it from scratch. Also, for CIFAR-10, as [37] suggests, we use ResNet50 [16] pretrained on ImageNet as a feature extractor and fine-tune its classification head.

Corresponding to each DP algorithm's specification, we find an optimizer and a learning rate using a grid search algorithm to ensure the training process achieves the highest accuracy. In addition, data augmentation reduces the effectiveness of all of the attacks [33,21], leading to a bias in our results. Therefore, we skip the data augmentation in our experiments. More details on the training process can be found in the appendix.

Attack and Threat Model All the DP models are in white-box settings. The backdoors are made based on the triggers introduced in BadNets [15]. To generate backdoors, we first randomly select two classes as base and target class. Then, we randomly select half of the samples from the base class, attach a 4×4 trigger patch to their bottom right corner and assign the target class as their labels [4]. We poison 50% base class to ensure the number of backdoors is high enough, and sufficient clean samples are left in the base class. Under this condition, the model learns both clean and backdoor data points.

Evaluation Metrics Attack success rate **(ASR)** is the metric to evaluate the success of the backdoor attacks. According to the definition of the backdoor attacks in Section 2.1, ASR indicates the number of test samples from base class that are patched with the backdoor trigger and misclassified as the target class. Thus, a defense method is considered more successful if it leads to a lower ASR.

The second defensive purpose is to maintain high **accuracy** for the clean test data. The original accuracy of our CIFAR-10 vanilla model over the clean test data is 91.24% and the backdoor ASR is 98.1%. The MNIST model's initial accuracy and ASR are 98.92% and 100%, respectively.

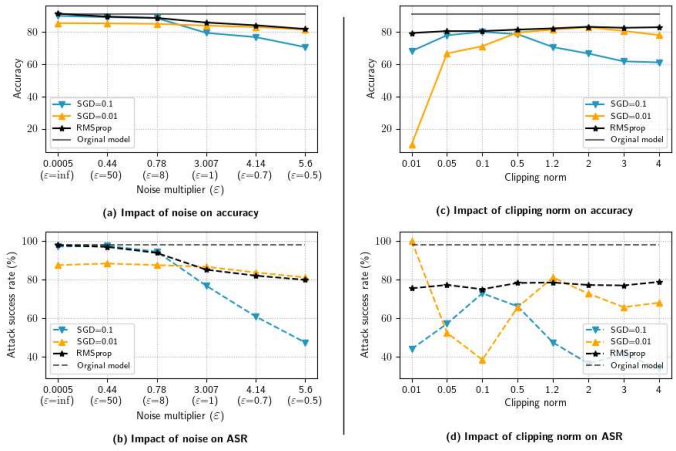

Fig. 1. Effectiveness of DP-SGD against backdoor attacks, w.r.t the noise multiplier, clipping norm, and the optimizer.

Experimental Roadmap This subsection provides an overview of the experiments in the forthcoming sections. In Section 6, we analyze two DP algorithms, DP-SGD and PATE, by assessing the impact of their privacy budget and other hyper-parameters on the attack success rate. This analysis helps us clarify the underlying reason for their defensive power. At the same time, we will show their resulting accuracy and attack success rate. Then, in Section 7, we compare all the DP and Label-DP algorithms in various circumstances to witness which one is prominent and whether the outcome alters in a different situation. Due to space constraints, we could not include all of our experiments and refer to the appendix for our findings on the exploration of parameters for Label-DP algorithms and the training procedure.

6 DP against Backdoors

This section investigates DP-SGD and PATE, against backdoor attacks. For each algorithm, we will evaluate their key hyper-parameters (introduced in Table 1) on CIFAR-10 dataset and show that some of them have a critical impact on the accuracy and ASR. The results of the MNIST dataset are very similar. So to be concise, we skip their reports here but use them to conduct the experiments in the subsequent sections.

6.1 DP-SGD vs. Backdoors

SGD is the dominant optimizer in practice paired with the DP-SGD algorithm, especially in defeating poisoning attacks [1,5,17,18]. So we consider different optimizers and learning rates to depict the sensitivity of DP-SGD performance

to these factors: RMSProp, SGD with a learning rate of 0.1, and SGD with a learning rate of 0.01. Based on the size of the dataset, we set the DP-SGD algorithm as $(\varepsilon, 10^{-5})$-DP and report ε as the privacy budget [30].

Fig. 1a and 1b show the impact of the noise multiplier by fixing the clipping norm to 1.2 (typical for CIFAR-10). Interestingly, the rate of the accuracy drop to the ASR drop differs for each optimizer. However, in general, higher noise levels reduce both accuracy and ASR simultaneously. This suggests that SGD can resist backdoor attacks more effectively by paying a slightly higher utility cost.

Fig. 1c and 1d illustrate the impact of different clipping norms on the accuracy (top) and ASR (bottom) using a fixed noise of 5.6. In contrast to RMSProp, for SGD optimizers, the choice of learning rate creates two different patterns of ASR with respect to the clipping norm. This reveals how SGD training without an adaptive learning rate can be affected by the norm of the gradients. Therefore, while the clipping norm significantly impacts the model utility and robustness, it is difficult to optimally adjust it when the defender is unaware of the attack specifications.

According to [6], the impact of the clipping norm on accuracy is not monotonic, which is manifested as a non-monotonic pattern of accuracy and ASR in Fig. 1c and 1d. Regarding the different pattern of ASR on the left side of Fig. 1d with SGD-0.01, we speculate that the small learning rate accompanied by a high noise and small clipping norm can hardly learn the normal images' manifold, and instead it retains the repetitive and striking patterns of the backdoor triggers.

Conclusion (Q1): In our evaluations, DP-SGD was successful in mitigating the impact of backdoor attacks. However, the noise multiplier, clipping norm and training parameters determine the extent of this success. As a result, differences in these parameters contribute to the varying results reported in previous studies on the effectiveness of DP-SGD as a defense mechanism.

6.2 PATE vs. Backdoors

In this section, we evaluate the robustness of PATE against backdoor attacks and the impact of different parameters including the number of teachers, number of queries, threshold, selection noise, and result noise. The results are shown in Fig. 2. Whenever noises or threshold are not evaluated, we fix their values to 0. In the case of the number of queries and number of teachers, the default values are fixed to 10000 and 200, respectively. For training PATE, we assume 1/5 (i.e. 10000 samples) of the training data is publicly available for training the student model, and the rest is private. In the original PATE paper [29], the number of queries is set to as low as 1000. However by doing so, we naturally remove a large fraction of poisoned data and make the comparison between different DP methods unfair. Therefore, we keep the default number of queries at 10000 and in the next sections, to compare the models, we analyze the impact of both noise and the number of queries on the PATE's utility and privacy budget.

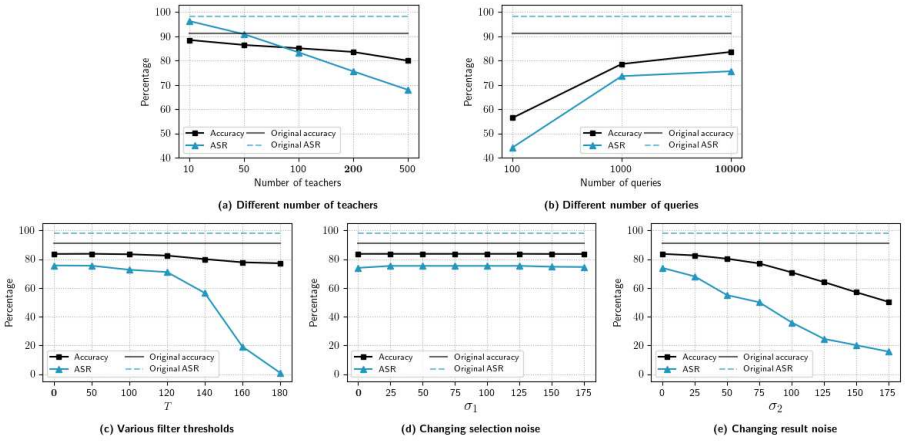

Fig. 2. The impact of number of teachers, number of queries, threshold, selection noise and result noise on the student model's accuracy and ASR from left to right and top to bottom, respectively).

Fig. 2a and 2b show that the number of teachers and the number of queries impact the accuracy and ASR in opposite ways. A higher number of teachers means fewer training data and lower accuracy for each teacher, hence less accurate consensus from the aggregator. This also compromises the consensus on assigning the target class to the backdoor samples and decreases the ASR, which aligns with the literature finding that bagging can hinder the success of the backdoor attacks [3,20,9]. Furthermore, in Fig. 2b, a lower number of queries is associated with less training data for the student model and fewer backdoors, hence lower accuracy and ASR.

Fig. 2c illustrates that the aggregation threshold is crucial in defeating backdoors and has minimal impact on utility loss. This finding complements previous results suggesting the use of bagging against poisoning attacks. The threshold forces the aggregation process to filter out uncertain data and backdoors, resulting in higher accuracy and lower ASR in the student model. To the best of our knowledge, this factor has not been considered in previous works as a major contributor to the effectiveness of bagging.

Fig. 2d and 2e demonstrate the effect of selection noise and result noise used in selecting and randomizing queries which form the basis of DP for PATE. We found similar trends when one of the noises is fixed to a random positive value. Based on these results, to defeat ASR we need a high result noise which leads to a drop in accuracy. Since we fixed the number of queries and only varied the noise values to control privacy, the privacy budget still remains as large as $\varepsilon = 4$ at a high noise level of 175.

Conclusion (Q1): PATE is very successful in defeating backdoor attacks. It can be more successful than DP-SGD but it is highly sensitive to the algorithm

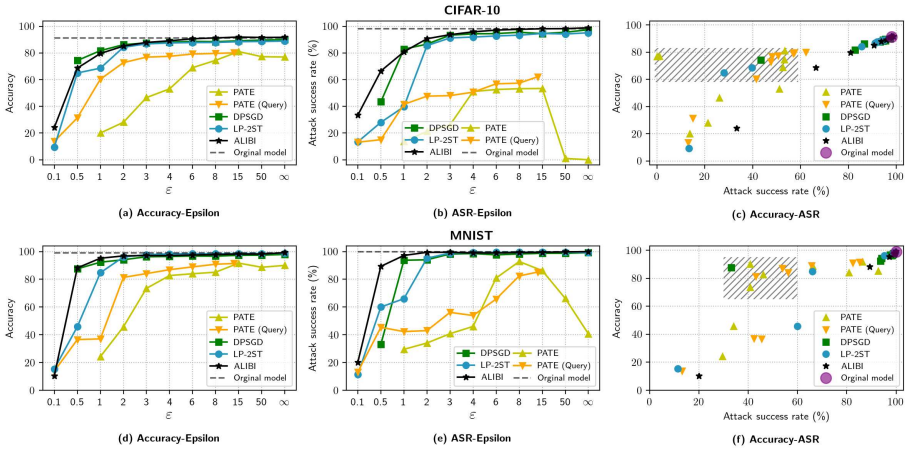

Fig. 3. The impact of epsilon on DP and Label-DP methods using MNIST (top) and CIFAR-10 dataset(bottom).

parameters. Result noise (σ_2) and the number of queries which are the most influential parameters on the privacy budget (ε) decrease the ASR but also cause a drastic decrease in the accuracy at the same time. Conversely, the best result is achieved through tuning the threshold, although it cannot provide any DP by thresholding alone.

7 Comparison of DP and Label-DP Methods

In this section, we compare all the DP and Label-DP algorithms to discover which one and under what conditions are more successful.

7.1 Privacy Budget Analysis

The ϵ in DP and Label-DP serves two different goals. So we do not directly compare the ϵ values of the two methods even though both can be reduced to label DP [14]. Instead, what we focus on is the trade-off between accuracy and ASR provided by varying ϵ of the two methods. We select the best parameters from the results in the previous section to conduct the current experiment. These best parameters lead to high accuracy and a low ASR. Wherever there is a trade-off between accuracy and ASR, we prioritize accuracy. For MNIST, we do not present those parameter selections due to the similar outcomes.

Fig. 3a,b compare the accuracy and ASR of the different methods for CIFAR-10 with varying ϵ while 3c shows the trade-off of accuracy and ASR of different methods (the ideal case correspond to 100%accuracy and 0% ASR). PATE can achieve different levels of privacy by varying two factors: 1) noises (lime green plots), and 2) number of queries (orange plots). The first observation is that

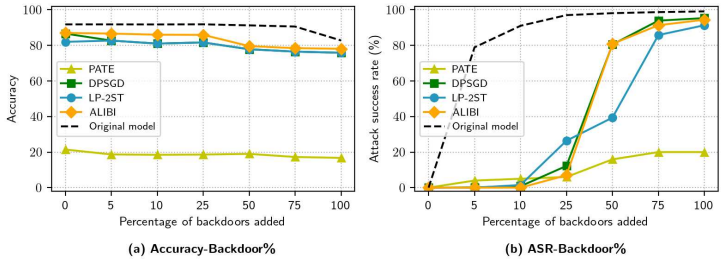

Fig. 4. The significant impact of poisoned data on DP-based defense methods. The epsilon is fixed to 1 and then all the methods are compared by varying the percentage of the training data that has been poisoned.

non-DP PATE outperforms all other results and methods (the rightmost point of the lime green plot). It indicates the power of bagging with a threshold against backdoor attacks. LP-2ST for some ϵ values works well. For instance, $\varepsilon = 1$ has high accuracy (78%) and a significantly decreased ASR (39%). However DP-SGD gives the best results when $\varepsilon = 0.5$. For ALIBI, both accuracy and ASR drop proportionally.

Fig. 3d,e,f show similar trends for MNIST. Fig. 3f combines the results of the two other columns by directly comparing the accuracy and corresponding ASR. The rectangular areas with the hatched pattern in the last column consist of the most desired results with high accuracy and dropped ASR regardless of their privacy budget. This area includes different private algorithms, but mostly PATE, which indicates the dominance of PATE.

Conclusion (Q2): The DP and Label-DP techniques effectively reduce the vulnerability of backdoor attacks, albeit at the cost of decreased accuracy. If the optimal approach is determined by the accuracy-to-ASR ratio, then the superiority of each DP or Label-DP model depends on the allocated privacy budget.

7.2 Attack Strength Analysis

We discussed the hyper-parameters and the privacy budget of the algorithm as two factors that impact the immunity of the DP approaches against backdoor attacks. A third factor that should be considered when assessing the level of immunity is the strength of the attack itself. So far, we have synthesized powerful attacks by poisoning 50% of the data with backdoors. However, in practice, the attacker conceals her malicious activity by limiting the percentage of poisoned data introduced into the pipeline. Therefore we change the percentage of the backdoors in the base class to develop a range of more realistic and more powerful (but less realistic) attacks.

Fig. 4 shows the accuracy and ASR with respect to the number of backdoors, when the privacy budget for all DP algorithms has been fixed to $\varepsilon = 1$. We observe that the accuracy does not drastically change with respect to the number

of backdoors, yet the ASR increases as the attack becomes more powerful. Looking at the pattern, we can see that the DP algorithms almost entirely diffuse the attack when the percentage of backdoors is sufficiently small. It should be noted that the low accuracy of PATE is a result of controlling its privacy budget by adding noise, rather than limiting the number of queries according to the reasoning we had in section 6.2.

Conclusion (Q2): These results illustrate the effectiveness of DP-SGD, LP-2ST, and ALIBI against more realistic backdoor attacks (with backdoor% ≤ 10). For such attacks, the accuracy drops by 10%, and the attack achieves no success. This is compatible with Equation (3) that shows that the attacker's loss limit in DP models is theoretically linked to the number of poisoned data.

7.3 Accuracy-Privacy Trade-off

To see the accuracy when a perfect defense is desired (close to 0 ASR), we have analyzed different privacy budgets for each DP method and found the greatest ε where the ASR does not exceed 1%. This small ASR is achievable when the number of backdoors is insignificant (we set it to 10%). By doing so, we achieve the least randomness that leads to a successful defense. After removing the impact of the attack, we can have a fair comparison of accuracy and training time.

Table 2. Comparison of the highest accuracy and epsilon that DP methods can achieve while ASR=0.

	DP-SGD	**PATE**	**ALIBI**	**LP-2ST**
Accuracy	88.67	85.02	**89.53**	79.9
Epsilon	2	**inf**	2	0.9
Time	140s	220s	**59s**	**58s**

Table 2 highlights the best values of accuracy, privacy budget, and training time in each row. The previous findings indicate that a deterministic version of PATE, with noise removed, is the most resilient against attacks. However, when the goal is to simultaneously defend against backdoors and protect privacy, this result is not favorable for PATE. DP-SGD and ALIBI, with the same privacy budget, can achieve better accuracy than PATE.

Finally, with respect to training time, two Label-DP methods demonstrate a considerable reduction in training time, surpassing other DP techniques. It is important to note that this experiment was conducted on a CIFAR-10 fine-tuning task, where training time is negligible. However, in more complex architectures with end-to-end settings, time may become a bottleneck for PATE and DP-SGD.

Conclusion (Q2): When a perfect defense is desired, Label-DP methods offer the best efficiency and comparable or better accuracy trade-off compared to DP approaches.

8 Discussion and Conclusion

This paper posed important questions regarding the ability of DP to provide robustness against backdoor attacks in practice. In addition to DP-SGD, we explored the other commonly used DP algorithm (PATE) and two Label-DP algorithms (LP-2ST and ALIBI) for the first time. We have several main findings.

First, the noise and randomness added to the private models can indeed decrease the attack success rate of the backdoors, but at the cost of utility drop for clean input. In a nutshell, a model trained with privacy guarantee has an inherent benefit in robustness against backdoor attacks. This statement holds for all four methods mentioned above. A somewhat unexpected outcome is that PATE delivers the best results, even without the use of noise (without DP guarantee) due to the ensemble based teacher-student structure.

Second, contrary to the claims of some previous studies, DP-SGD provides good resistance against backdoors while keeping the accuracy relatively high. We also observed the same phenomenon for Label-DP algorithms. The accuracy-ASR trade-off is diverse among the DP and Label-DP methods we analyzed. One model may outperform the others depending on the privacy budget, algorithm parameters, and attack specifications. Therefore, it is possible to use DP models as defense strategies. A proper selection of the above mentioned factors can adequately balance the accuracy and ASR.

This work was an empirical study on two benchmark datasets, MNIST and CIFAR-10. It offered new empirical insights into the connection between DP and backdoor attacks in relation to existing theoretical understandings. Future research could focus on exploring the impact of Label-DP on particular type of poisoning attacks focusing on labels such as label-based flipping attacks. Additionally, given the ability of DP methods to enhance robustness, there is an opportunity to develop modified DP algorithms that offer greater protection against poisoning attacks, and simultaneously fulfill both privacy and robustness objectives.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. pp. 308–318 (2016)
2. Andrew, G., Chein, S., Papernot, N.: Tensorflow privacy library (2020)
3. Biggio, B., Corona, I., Fumera, G., Giacinto, G., Roli, F.: Bagging classifiers for fighting poisoning attacks in adversarial classification tasks. In: International workshop on multiple classifier systems. pp. 350–359. Springer (2011)
4. Borgnia, E., Cherepanova, V., Fowl, L., Ghiasi, A., Geiping, J., Goldblum, M., Goldstein, T., Gupta, A.: Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3855–3859. IEEE (2021)

5. Borgnia, E., Geiping, J., Cherepanova, V., Fowl, L., Gupta, A., Ghiasi, A., Huang, F., Goldblum, M., Goldstein, T.: Dp-instahide: Provably defusing poisoning and backdoor attacks with differentially private data augmentations. arXiv preprint arXiv:2103.02079 (2021)
6. Bu, Z., Wang, Y.X., Zha, S., Karypis, G.: Automatic clipping: Differentially private deep learning made easier and stronger. In: ICML TPDP workshop (2022)
7. Chan, A., Ong, Y.S.: Poison as a cure: Detecting and neutralizing variable-sized backdoor attacks in deep neural networks. arXiv:1911.08040 [cs] (November 2019), arXiv: 1911.08040
8. Chen, B., Carvalho, W., Baracaldo, N., Ludwig, H., Edwards, B., Lee, T., Molloy, I., Srivastava, B.: Detecting backdoor attacks on deep neural networks by activation clustering. In: SafeAI@AAAI (2019)
9. Chen, R., Li, Z., Li, J., Yan, J., Wu, C.: On collective robustness of bagging against data poisoning. In: International Conference on Machine Learning. pp. 3299–3319. PMLR (2022)
10. Du, M., Jia, R., Song, D.: Robust anomaly detection and backdoor attack detection via differential privacy. In: ICLR 2020 (2020)
11. Dwork, C.: Differential privacy. vol. 2006, pp. 1–12. ICALP (2006)
12. Dwork, C.: Differential privacy. In: International colloquium on automata, languages, and programming. pp. 1–12. Springer (2006)
13. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: Privacy via distributed noise generation. In: Annual international conference on the theory and applications of cryptographic techniques. pp. 486–503. Springer (2006)
14. Ghazi, B., Golowich, N., Kumar, R., Manurangsi, P., Zhang, C.: Deep learning with label differential privacy. Advances in Neural Information Processing Systems **34** (2021)
15. Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S.: Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access **7**, 47230–47244 (2019)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. Hong, S., Chandrasekaran, V., Kaya, Y., Dumitraş, T., Papernot, N.: On the effectiveness of mitigating data poisoning attacks with gradient shaping. arXiv preprint arXiv:2002.11497 (2020)
18. Jagielski, M., Oprea, A.: Does differential privacy defeat data poisoning. In: DPML Workshop (2021)
19. Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., Li, B.: Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In: 2018 IEEE Symposium on Security and Privacy (SP). pp. 19–35. IEEE (2018)
20. Jia, J., Cao, X., Gong, N.Z.: Intrinsic certified robustness of bagging against data poisoning attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 7961–7969 (2021)
21. Koh, P.W., Steinhardt, J., Liang, P.: Stronger data poisoning attacks break data sanitization defenses. arXiv preprint arXiv:1811.00741 (2018)
22. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
23. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
24. Lu, S., Li, R., Liu, W., Chen, X.: Defense against backdoor attack in federated learning. Computers & Security **121**, 102819 (2022)

25. Ma, Y., Zhu, X., Hsu, J.: Data poisoning against differentially-private learners: Attacks and defenses. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. p. 4732–4738. IJCAI'19, AAAI Press (2019)
26. Malek Esmaeili, M., Mironov, I., Prasad, K., Shilov, I., Tramer, F.: Antipodes of label differential privacy: Pate and alibi. Advances in Neural Information Processing Systems **34** (2021)
27. Miao, L., Yang, W., Hu, R., Li, L., Huang, L.: Against backdoor attacks in federated learning with differential privacy. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2999–3003. IEEE (2022)
28. Naseri, M., Hayes, J., Cristofaro, E.D.: Local and central differential privacy for robustness and privacy in federated learning. Proceedings 2022 Network and Distributed System Security Symposium (2020)
29. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., Talwar, K.: Semi-supervised knowledge transfer for deep learning from private training data. arXiv preprint arXiv:1610.05755 (2016)
30. Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Erlingsson, Ú.: Scalable private learning with pate. arXiv preprint arXiv:1802.08908 (2018)
31. Peri, N., Gupta, N., Huang, W.R., Fowl, L., Zhu, C., Feizi, S., Goldstein, T., Dickerson, J.P.: Deep k-nn defense against clean-label data poisoning attacks. In: European Conference on Computer Vision. pp. 55–70. Springer (2020)
32. Saha, A., Subramanya, A., Pirsiavash, H.: Hidden trigger backdoor attacks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11957–11965 (2020)
33. Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J.P., Goldstein, T.: Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In: International Conference on Machine Learning. pp. 9389–9398. PMLR (2021)
34. Tang, X., Nasr, M., Mahloujifar, S., Shejwalkar, V., Song, L., Houmansadr, A., Mittal, P.: Machine learning with differentially private labels: Mechanisms and frameworks. Proceedings on Privacy Enhancing Technologies (2022)
35. Tran, B., Li, J., Madry, A.: Spectral signatures in backdoor attacks. Advances in neural information processing systems **31** (2018)
36. Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., Zhao, B.Y.: Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In: 2019 IEEE Symposium on Security and Privacy (SP). pp. 707–723. IEEE (2019)
37. Wang, L., Zheng, J., Cao, Y., Wang, H.: Enhance pate on complex tasks with knowledge transferred from non-private data. IEEE Access **7**, 50081–50094 (2019)
38. Warner, S.L.: Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association **60**(309), 63–69 (1965)
39. Weber, M., Xu, X., Karlaš, B., Zhang, C., Li, B.: Rab: Provable robustness against backdoor attacks. In: 2023 IEEE Symposium on Security and Privacy (SP). pp. 1311–1328. IEEE (2023)
40. Xu, C., Wang, J., Guzmán, F., Rubinstein, B., Cohn, T.: Mitigating data poisoning in text classification with differential privacy. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 4348–4356. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021)
41. Zhao, S., Ma, X., Zheng, X., Bailey, J., Chen, J., Jiang, Y.G.: Clean-label backdoor attacks on video recognition models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14443–14452 (2020)

A Appendix

A.1 Experimental Setup Details

Training Configuration. For MNIST and CIFAR-10 datasets, we used different architectures for neural networks. For CIFAR-10, the ResNet-50 head was followed by an average pooling layer and two linear layers of size 256 and 10. For MNIST, the neural network consisted of two convolutional layers with 16 and 32 filters, each of kernel size 8 and 4 followed by max pooling layers and two final linear layers with 32 and 10 neurons. The learning rate of all four DP and Label-DP algorithms was 0.001 and the number of epochs was fixed to 50. In contrast to Label-DP algorithms where an SGD optimizer was good enough to train the model, for DP-SGD and PATE we required to use more adaptive optimizers, i.e. RMSProp and Adam, respectively.

Table 3. Parameters of the DP and Label-DP algorithms

Method	Parameters
LP-2ST	1. **1. Data split ratio** : The portion the training dataset split between two training stages (more in the first stage helps with accurate prior but causes underfit in the second stage) 2. **2. Temperature T** : For logit z_i and calculation of prior p_i of class i, a small T in $p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$ boosts the confidence of the top classes and a large T makes the priors more uniform 3. **3. Epsilon** ε : Randomness parameter that is equivalent to the privacy budget
ALIBI	1. **1. noise of soft training labels** : Laplacian noise with $\delta = 0$ which is applied once and determines the privacy budget

B Label DP against Backdoors

In this section , we evaluate LP-2ST, and ALIBI as two Label-DP models. We investigate if their randomness or other related parameters can help to mitigate the backdoor attacks. To this end, Table 3 presents the various parameters involved in these algorithms.

B.1 LP-2ST vs. Backdoors

Since the Label-DP algorithms randomly change the labels, we found that the accuracy in high noise fluctuate among multiple runs. So for each experiment on LP-2ST and ALIBI, the accuracy and ASR are the averages of 10 trials. For each figure from left to right, we pick a parameter shown on the x-axis (which is chosen randomly) and apply it for the experiments in the succeeding figure. For the first two figures, we set $\varepsilon = 1$.

Fig. 5a demonstrates the effect of temperature with a random data split of [80/20]. Compatible to [14], sparsifying the priors helps to improve the utility, but to our surprise, it decreases ASR. We speculate the reason is that the backdoor still has a touch of the base class. Thus the first round of LP-2ST predicts target and base classes as the backdoors' top-2 classes. The sparsified prior shifts the probabilities of these two classes far away from zero, so the algorithm selects the base class more confidently.

In Fig. 5b the training data has been partitioned for two stages. [p1/p2] on the x-axis indicates the percentage of the data in stage 1 and stage 2 of LP-2ST, respectively. When 100% of data is allocated to the first stage, it means that we are using LP-1ST with RR. There is not a clear pattern between ASR and data split. But an LP-2ST model with more data in the first stage has more enhanced priors and higher accuracy. Fig. 5c compares different privacy budgets ε, which is the random factor of the RR algorithm. Naturally, more randomness helps to decrease the ASR. The results for $\varepsilon = 1$ are particularly impressive since it drops the ASR to less than 40%, while the accuracy is still roughly 80%.

Conclusion: Surprisingly, even though Label-DP only randomizes the labels, it is still successful against backdoor attacks. In this success, all parameters are involved, but noise has the major impact. LP-2ST can vividly mitigate the attack, but it is very important which ε is selected to obtain a reasonable accuracy-ASR trade-off.

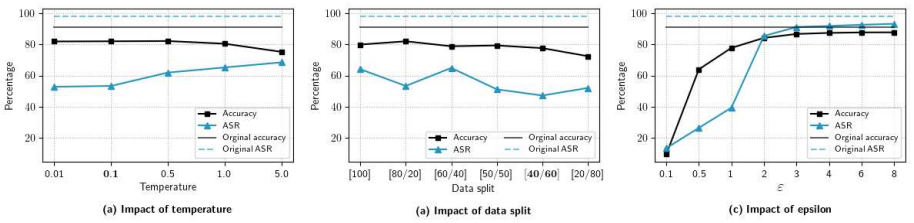

Fig. 5. The impact of temperature, data split between two stages and epsilon on LP-2ST (from left to right). Epsilon, the factor of privacy-preserving in LP-2ST, can drastically deteriorate the ASR with an acceptable utility cost (c).

B.2 ALIBI vs. Backdoors

According to Fig. 6, ALIBI with higher noise drops both accuracy and ASR proportionally. This can be justified by the fact that all the labels randomly change just once at the beginning of the training.

Conclusion: On average, ALIBI can mitigate the effect of backdoor attacks but with reduced utility costs.

B.3 Training Process

In this section, we compare the training process of DP-SGD, LP-2ST, and ALIBI on CIFAR-10. These comparisons are based on two privacy budgets $\varepsilon = \infty$

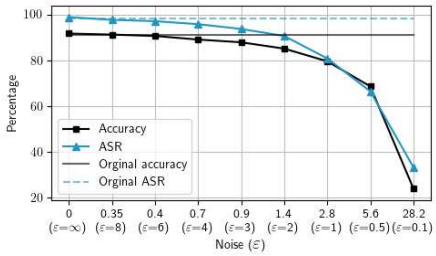

Fig. 6. Effectiveness of randomizing labels on reducing ASR in ALIBI. The noise added to one-hot labels in ALIBI impacts both accuracy and ASR proportionally.

and $\varepsilon = 1$, to provide an overview over the training process with and without randomness. For LP-2ST, we only illustrate the training of the second and final stage of the algorithm. In Fig. 7, each column demonstrates a different method, and each row indicates one of the privacy budgets. For all three differentially private methods, on the first row, with $\varepsilon = \infty$, the loss of the backdoor samples drops below the clean loss on early training epochs. It is the opposite for all three methods when $\varepsilon = 1$ on the second row. For LP-2ST the backdoor loss does not converge to the clean loss and remains higher. It is consistent with the results of LP-2ST at $\varepsilon = 1$ in Fig. 5c. For ALIBI the clean and backdoor losses change very closely. It explains the similar values for the ALIBI accuracy and ASR in Fig. 6. DP-SGD can resist the backdoor samples on early epochs. So one of the suggestions is to stop the training early to avoid backdoors from overfitting.

Conclusion: During DP training, the model underfits or suppresses the backdoor samples which results in defusing the backdoors' impact on the model. This finding confirms the results of the paper.

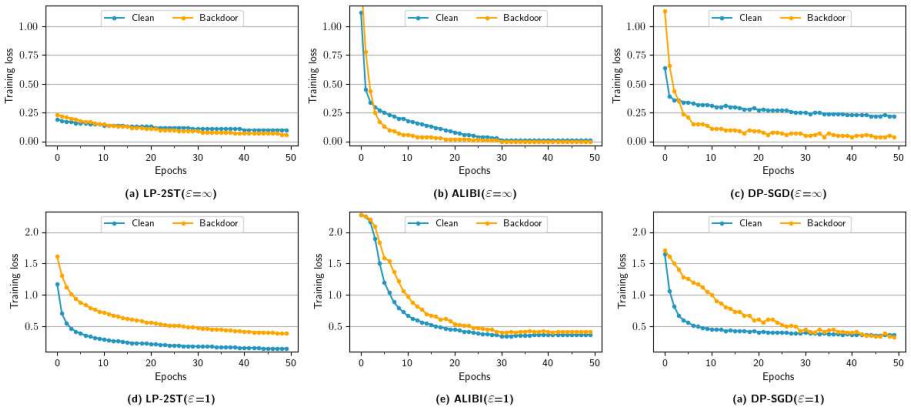

Fig. 7. An overview of the training process of LP-2ST, ALIBI and DP-SGD using $\varepsilon = \infty$ (upper) and $\varepsilon = 1$ (lower).