Do Backdoors Assist Membership Inference Attacks?

Yumeki Goto[1(✉)], Nami Ashizawa[2], Toshiki Shibahara[2], and Naoto Yanai[1]

[1] Osaka University, 1-5 Yamadaoka, Suita-shi, Osaka 565-0871, Japan
{y-goto,yanai}@ist.osaka-u.ac.jp
[2] NTT Social Informatics Laboratories, 3-9-11 Midori-cho, Musashino-shi, Tokyo 180-8585, Japan
{nami.ashizawa,toshiki.shibahara}@ntt.com

Abstract. When an adversary provides poison samples to a machine learning model, privacy leakage, such as membership inference attacks that infer whether a sample was included in the training of the model, becomes effective by moving the sample to an outlier. However, the attacks can be detected because inference accuracy deteriorates due to poison samples. In this paper, we discuss a *backdoor-assisted membership inference attack*, a novel membership inference attack based on backdoors that return the adversary's expected output for a triggered sample. We found three key insights through experiments with an academic benchmark dataset. We first demonstrate that the backdoor-assisted membership inference attack is unsuccessful when backdoors are trivially used. Second, when we analyzed latent representations to understand the unsuccessful results, we found that backdoor attacks make any clean sample an inlier in contrast to poisoning attacks which make it an outlier. Finally, our promising results also show that backdoor-assisted membership inference attacks may still be possible only when backdoors whose triggers are imperceptible are used in some specific setting.

Keywords: Backdoor-assisted membership inference attack · backdoor attack · poisoning attack · membership inference attack

1 Introduction

Membership inference attacks [31] are attacks where an adversary infers whether a sample was utilized for training a machine learning model. They are currently used for evaluating privacy leakage in various machine learning models [2,8,23,41]. Then, many researchers focus on enhancing privacy violations through other attacks, such as poisoning attacks, for the best evaluation of privacy leakage. In recent years, Tramer et al. [36] proposed an advanced attack, called poisoning-assisted membership inference attack, for amplifying privacy leakage by injecting poison samples into a dataset. The drawback of the attack is to deteriorate the inference accuracy of the victim model injected poison samples. Consequently,

H. Duan et al. (Eds.): SecureComm 2023, LNICST 568, pp. 251–265, 2025.
https://doi.org/10.1007/978-3-031-64954-7_13

the owner of the victim model can detect the underlying poison samples in any kind of poisoning attack [35]. Namely, the poisoning-assisted membership inference attack can be prevented by detecting poison samples; thus, it may be less severe than expected.

The above limitation leads us to a membership inference attack utilizing a backdoor attack, i.e., *a backdoor-assisted membership inference attack*. Backdoor attacks [14] are stealthier than poisoning attacks because they manipulate the output of only triggered samples and maintain test accuracy [35]. There are also advanced attacks, called imperceptible backdoors [9,21,27,33,34,43–45], that bypass existing backdoor detection tools [5,7,13,24,37,38].

In this paper, we take the first step for answering the following key question on backdoor-assisted membership attacks: *Is a backdoor-assisted membership inference attack feasible?* This is non-trivial. In particular, it is unclear whether a backdoor-assisted membership inference attack works because the backdoor attacks maintain inference accuracy. The key idea of the existing poisoning-assisted membership inference attack [36] is to make the target sample an outlier by deteriorating accuracy with poison samples. In contrast, backdoors may not make the target sample an outlier because they maintain accuracy. For this reason, the backdoor-assisted membership inference attack is significantly different from the existing attack.

We found three key insights through the experiments with a typical academic benchmark. As the first insight, the backdoor-assisted membership inference attack is *unsuccessful* as long as typical backdoor attacks [14,34] are trivially used, as opposed to the poisoning-assisted membership inference attack. We evaluate the attack success rate (ASR) as a metric of membership inference attacks. When ASR of the original membership inference attack [2] is 58.3%, that of the backdoor-assisted membership inference attack is <57%, and that of the poisoning-assisted membership inference [36] is 95%, respectively. Specifically, a backdoor attack amplifies only a few attack success rates of the membership inference. Next, we analyze latent representations of the victim models to understand the above phenomenon deeply. Then, as the second insight, we found the fact that the backdoor-assisted membership inference attack makes a target sample an *inlier* in the distribution of latent representations, while the existing poisoning-assisted membership inference attack [36] makes it an outlier. As the third insight, when we use an imperceptible backdoor attack for triggers, called LIRA [9], we found the fact that scores of evaluation metrics for the backdoor-assisted membership inference attack are improved with only a few points in some specific setting called *untargeted* setting [36], although it is quite lower than the results by Tramer et al. [36]. In our experimental setting, in contrast to 63% ASR for the original membership inference attack, that for imperceptible backdoor attack is 69%. Here, that of the poisoning-assisted membership inference [36] is 87% ASR. Our observation indicates that backdoors may still have a chance to assist in membership inference if an imperceptible backdoor attack for triggers is used. (See Sect. 5 for detail.)

To sum up, we found the following key insights in this paper:

(1) Backdoor-assisted membership inference attacks are unsuccessful in a trivial way.
(2) We analyze latent representations to understand the reason for the unsuccessful results. We then demonstrate that backdoor-assisted membership inference attacks make a target sample an inlier, while poisoning-assisted membership inference attacks make it an outlier.
(3) Backdoor-assisted membership inference attacks may be possible if an imperceptible backdoor attack for triggers is used in the untargeted setting.

2 Related Work

In this section, we describe related works of backdoor attacks and privacy violations assisted by poisoning attacks.

2.1 Backdoor Attacks

Backdoor attacks [14,25] are a kind of attack whereby an adversary trains a model such that he/she obtains the expected output for only triggers. Recent backdoor attacks [20,21,27,29,40,44] can bypass detection tools [5,7,13,24,37,38], and thus existence of backdoors is imperceptible. (We call them imperceptible backdoor attacks for the sake of convenience.)

There are two approaches for constructing imperceptible backdoor attacks. The first approach [9,21,27,44] is based on trigger generation that is visually imperceptible for humans, referred to as the trigger method. The second approach [33,34] is based on latent representations whose distributions are close between clean inputs and triggers, referred to as the latent-representation method. We evaluate backdoor-assisted membership inference attacks based on the above two imperceptible backdoor attacks as well as the original backdoor attack [14]. Although there are several works [43,45] that unify the above two attacks, we believe that our work also implies the results of the works described above.

In recent years, backdoor attacks have been discussed in real-world applications such as natural language processing [20,28] and face authentication [40]. Combining with these works is the next step to reveal real-world threats.

2.2 Privacy Violations Assisted by Poisoning

As mentioned in Sect. 1, enhancing privacy violations through other attacks, such as poisoning attacks, is an important research direction in the field of privacy for machine learning models. The existing works on privacy violations [6,17,26,36,39] are based on poisoning attacks. The first work [17] was with simple models such as support vector machines. Whereas several papers [26,39]

discussed property inference attacks [12] that infer properties of a training dataset, Tramer et al. [36] discussed membership inference, attribute inference [10,11,42], and data extraction [3,4,15]. Nevertheless, the above works did not discuss backdoor attacks. Namely, the above works succeeded in membership inference attacks by sacrificing accuracy [36].

The closest work to ours is by Chen et al. [6]. They evaluated a membership inference attack with clean-label poisoning [30,46], whose labels remain unchanged and samples are visually indistinguishable from clean samples. However, the drawback of their work is to be detected with the latest detection methods on latent representation [5,19,34]. In their attack, the distance between clean and poison samples in latent representations still becomes far from each other to maximize the influence of the target. We discuss not only the original backdoor attack [14], which is regardless of the distance between clean and poison samples but also the imperceptible backdoor attacks whose distance between clean and poison samples in latent representations is close to each other. Consequently, our work is able to bypass these detection methods, and thus the use of imperceptible backdoor attacks extremely differs from Chen et al.'s work.

Although several works [18,22] combine backdoors with membership inference, they are close to watermarking [1] to check if a model is backdoored. Namely, these works are quite different from privacy violations, i.e., our main problem, because they infer backdoors embedded by a model owner. Meanwhile, Redactor [16] is a tool to prevent membership inference attacks by generating samples to dilute specific data. Although the motivation of Redactor is to design a tool for preventing membership inference attacks, our results seem to be consistent with Redactor in the context that the attacks are unsuccessful.

3 Backdoor-Assisted Membership Inference Attack

We describe a backdoor-assisted membership inference attack as the problem setting of this paper below. We first define an attack formally and its metrics. We then describe the key questions in detail.

3.1 Formalization

The attacks in this paper are defined as a game between an adversary $\mathcal{A}$ and a challenger $\mathcal{C}$. We first denote by $\mathcal{X}$ a set of data samples, by $\mathcal{Y}$ a set of labels, by $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$ a set of datasets, and by $\mathcal{M}$ a set of machine learning models. Then, a machine learning model $M \in \mathcal{M}$ is defined as a mapping function $M : \mathcal{X} \to \mathcal{Y}$ and a training algorithm is defined as a mapping function $L_M : \mathcal{D} \to \mathcal{M}$.

The game is defined below. $\mathcal{A}$ then interacts with the final trained model. $\mathcal{A}$ is in the black-box setting that he/she sends queries to the model M and obtains only the outputs. Here, $\mathcal{A}$ can only provide statically poison samples.

1. $\mathcal{C}$ chooses a clean dataset $D \subseteq \mathcal{D}$.
2. $\mathcal{C}$ chooses a bit $b \leftarrow \{0, 1\}$. If $b = 1$, $\mathcal{C}$ chooses a sample $z \in D$; otherwise, $\mathcal{C}$ chooses $z \in \mathcal{D} \backslash D$.

3. Given z, $\mathcal{A}$ chooses a poisoning dataset $D_p \subset \mathcal{D}$ consisting of n samples, and send it to $\mathcal{C}$.
4. $\mathcal{C}$ trains M by $M = L_M(D^*)$ with the entire dataset $D^* = D \cup D_p$. In doing so, assuming $M^* = L_{M^*}(D)$ for any model $M^* \in \mathcal{M}$, M^* and M achieve the following relations: (1) for any $(x, y) \in D$, $M(x) = M^*(x)$ holds; and, (2) for any $(x^*, y^*) \in D_p$, $M(x^*) = y^*$.
5. $\mathcal{A}$ sends samples $x_1, \cdots, x_q \in \mathcal{X}$ to M and gets $y_1 = M(x_1), \cdots, y_q = M(x_q)$.
6. $\mathcal{A}$ returns a bit $b' \in \{0, 1\}$. If $b = b'$ holds, $\mathcal{A}$ wins the game.

In the above game-based definition, the sentences with red color differ from the existing attack by Tramer et al. [36]. While an adversary in the existing attack does not require a model M for anything other than learning a dataset D_p, our adversary requires a model M to misinfer only a sample x^* in D_p as his/her expected output y^*. It is the requirement of backdoor attacks [14].

3.2 Evaluation Metrics

We adopt the following evaluation metrics in this paper.

Membership-Inference-Attack Success Rates (ASR) [31]: For the number n of execution times of the game described in the previous section and the number a of times that $\mathcal{A}$ wins the game, it is defined as a/n.

Membership-Inference-Attack AUC (AUC) [32]: AUC is defined as an area under the ROC curve. The ROC curve is a two-dimensional curve defined by true positive rates (TPR) and false positive rates (FPR). It means that positive and negative values are accurately estimated as members and non-members.

In addition to the above metrics, we introduce metrics for backdoor attacks [14], i.e., test accuracy and backdoor identification rates[1]. For a model $M = L_M(D^*)$ with the entire dataset $D^* = D \cup D_p$ and another model $M^* = L_{M^*}(D)$ with only a clean dataset D, they are defined as follows:

Test Accuracy (TA): For any pair $(x, y) \in D \subseteq \mathcal{D} \backslash D_p$ of a clean sample and its label, it is defined as a ratio such that $y = M(x)$ holds, where $M(x) = M^*(x)$ holds for the original inference by D possibly. Intuitively, accuracy is necessary for stealthy compared to conventional poisoning attacks.

Backdoor Identification Rates (BIR): For any pair $(x^*, y^*) \in D_p$ of a poison sample and its label, it is defined as a ratio such that $M(x^*) = y^*$ holds, where $M(x^*) \neq M^*(x^*)$ may hold. It means that an adversary $\mathcal{A}$ can certainly exploit backdoors embedded in M.

3.3 Key Questions

As described in Sect. 1, our primary motivation is to identify whether backdoor-assisted membership inference attacks are feasible. To this end, we have three

[1] It is originally defined as the attack success rate in [14], but we say backdoor identification rate for convenience.

Table 1. Property of Poisoning-/Backdoor-Assisted Attacks: We investigate and discuss the following attack methods. For the column of "Accuracy," the checkmark means that backdoor attacks keep the accuracy. For the columns of "Trigger" and "Latent Representation," the checkmarks mean imperceptible backdoors in each context.

	Method	Accuracy	Trigger	Latent Representation
	Truth Serum [36]			
	Chen et al. [6]		✓	
Our	BadNets [14]	✓		
Our	TaCT [34]	✓		✓
Our	LIRA [9]	✓	✓	

key questions about backdoor-assisted membership inference attacks. First, we discuss the impact of the difference from the existing works [6,36], i.e., the sentences with red color, on the attacks. We then evaluate ASR and AUC with respect to TA and BIR. Table 1 summarizes the primary differences from the existing works [6,36].

Second, we analyze latent representations, which is an activation of a victim model when it takes samples as input. Latent representations are used for the analysis of backdoors [5,33] because they are often embedded into neurons that are otherwise dormant in the presence of clean inputs [14]. We will be able to understand why the results of backdoor-assisted membership inference attacks are obtained through the analysis of latent representations.

Third, motivated by Tramer et al. [36], we discuss the attacks in two settings, i.e., targeted setting and untargeted setting. In the targeted setting, an adversary focuses on data points of specific samples for backdoor attacks. It is considered that the target setting is a standard way for backdoor attacks in the context of specifying samples for the attacks. On the other hand, in the untargeted setting [36], an adversary controls a larger fraction of the training data in order to increase privacy leakage of all other data points. One might think that the untargeted setting still gives an adversary the chance to improve ASR and AUC even if the adversary fails in the targeted setting.

4 Experiment

We conduct experiments with our backdoor-assisted membership inference attacks. As described in the previous section, our goal is to discuss the impact of backdoors on ASR and AUC by comparing them with the existing poisoning-assisted membership inference attack by Tramer et al. [36]. We follow the setting in Tramer et al.'s work [36], including the targeted and untargeted settings. Our source code is publicly available for reproducibility and subsequent works[2].

[2] https://github.com/fseclab-osaka/backdoor-assisted-MIA.

4.1 Setting

We describe the targeted setting and the untargeted setting in terms of dataset, model, and baseline below. Our experiment is conducted on a machine equipped with Intel Core i7-13700KF CPU, 32GB memory, and Geforce RTX4090 GPU. Then, we implemented the code of our work with several frameworks, i.e., PyTorch and the scikit-learn library.

Dataset. We utilize the CIFAR-10 dataset for the experiments. Then, with reference to the experimental setting in Tramer et al.'s work [36], the CIFAR-10 dataset is split in each setting as follows:

Targeted Setting: In this setting, 250 samples are extracted as poison samples in D_p from 50,000 training samples of the CIFAR-10 dataset. Also, the $50,000$ training samples are divided into two groups: the training dataset D and the test dataset $\mathcal{D}\backslash D$ of the victim model. The victim model learns the entire dataset $D^* = D \cup D_p$.

Untargeted Setting: In the untargeted settings, $12,500$ samples are extracted as poison samples in D_p from $50,000$ training samples of the CIFAR-10 dataset. Then, the remaining $37,500$ training samples are divided into three groups: $12,500$ samples in the training dataset D, $12,500$ samples in the test dataset $\mathcal{D}\backslash D$ of a victim model, and $12,500$ samples that are never used. The victim model learns the entire dataset $D^* = D \cup D_p$.

Model and Baseline. We train five ResNet18 models in each setting as follows:

Targeted Setting: (1) model attacked with neither poisoning nor backdoor attacks, referred to as Clean-Only model; (2) model referred to as Truth Serum [36] as a baseline attacked with $250 \times r$ poison samples in D_p for $r \in \{1, 2, 4, 8, 16\}$; and (3)–(5) models backdoor-attacked with $250 \times r$ poison samples in D_p for backdoors based on BadNets [14], TaCT [34], and LIRA [9], respectively. Each model learns $25,000$ samples randomly chosen from 50,000 training samples as D. In addition, the models of (2)–(5) further learn poison samples in D_p in the above manner.

Untargeted Setting: The above six models described in the targeted setting are also used in the untargeted setting, where only the number of training samples is different as follows: (1) the Clean-Only model is trained with $12,500$ training samples as D; (2) Truth Serum with $12,500$ poison samples in D_p in addition to D; and (3) BadNets, (4) Tact, and (5) LIRA with $12,500$ poison samples in D_p for backdoors in addition to D.

Membership Inference Attack Method. We implemented the membership inference attack by Carlini et al. [2]. Their attack needs shadow models to mimic the data distribution of a victim model M, where the number of shadow models in each setting is as follows:

Table 2. Results of membership inference attacks on the Clean-Only model. The Clean-Only model infers classes of non-training data with high test accuracy. When ASR and AUC are close to 50%, it means that they are random.

	TA	ASR [%]	AUC
Targeted setting	90.94	58.30	0.60
Untargeted setting	86.45	63.05	0.65

Targeted Setting: We prepare 20 models and then choose the victim model M among the models. The remaining models, i.e., 19 models, are used for the shadow models to conduct a full leave-one-out cross-validation. We measure ASR and AUC on these five models and then evaluate their results.

Untargeted Setting: We prepare 40 models, and the other setting is common with the targeted setting.

4.2 Results

We show the results in each setting. Here, Table 2 shows the result of the membership inference attack against the Clean-Only model as the baseline, and Table 3 shows the results of each attack. We discuss the results in detail below.

Targeted Setting: According to Table 3, BadNets and TaCT, which need a few triggers for backdoors, keep high test accuracy but decrease both ASR and AUC, unlike Truth Serum. More specifically, whereas Truth Serum can increase ASR with greater than or equal to 27.7 points and ASR with greater than or equal to 0.32 compared to the Clean-Only model shown in Table 2, those for BadNets and TaCT are decreased with a few points. We also note that, when we compare BadNets with TaCT, both ASR and AUC for TaCT deteriorate by a few points for any number of samples. It indicates that ASR and AUC deteriorate due to triggers generated in imperceptible approaches.

We also explain why TA and BIR for LIRA are low below. LIRA needs the same number of triggers as clean samples for backdoors. In this experiment, we used at most 4000 poison samples for LIRA, despite training with 25000 clean samples. Therefore, LIRA could not achieve as high BIR as other backdoors.

Untargeted Setting: According to Table 3, we obtained quite different results from those in the targeted setting. In particular, compared with the Clean-Only model, while Truth Serum improves ASR and AUC instead of significantly decreasing TA, LIRA also improves them despite decreasing TA by a few points. Interestingly, both TA and BIR for LIRA are also higher than those in the targeted setting. It means that LIRA was trained well because we could provide enough number of training samples. We also note that ASR and AUC for BadNets and TaCT still deteriorate compared with the Clean-Only model.

Table 3. Results of each attack. ASR becomes higher for Truth Serum but lower with each backdoor attack. TA, BIR, and ASR are the percentages of leave-one-out cross-validation with 20 shadow models.

	Targeted $r = 1$				Targeted $r = 2$				Targeted $r = 4$			
	TA	BIR	ASR	AUC	TA	BIR	ASR	AUC	TA	BIR	ASR	AUC
Truth Serum	85.94	-	**86.08**	**0.92**	85.85	-	**91.34**	**0.96**	85.52	-	**93.48**	**0.97**
BadNets	90.85	94.72	56.68	0.57	90.83	94.45	55.28	0.57	90.82	94.57	54.40	0.55
TaCT	90.84	80.48	51.86	0.52	90.86	80.73	51.48	0.52	90.87	81.06	51.04	0.51
LIRA	43.49	16.45	50.12	0.50	49.62	18.99	49.68	0.49	53.06	39.20	50.98	0.51
	Targeted $r = 8$				Targeted $r = 16$				Untargeted			
	TA	BIR	ASR	AUC	TA	BIR	ASR	AUC	TA	BIR	ASR	AUC
Truth Serum	85.20	-	**93.89**	**0.98**	84.66	-	**95.09**	**0.98**	40.55	-	**81.47**	**0.87**
BadNets	90.83	94.41	54.04	0.56	90.82	94.34	53.68	0.55	84.01	99.44	61.74	0.64
TaCT	90.86	81.72	50.62	0.50	90.84	82.21	51.28	0.51	89.64	89.79	59.38	0.61
LIRA	61.98	62.16	49.92	0.50	68.13	70.29	50.12	0.50	82.00	99.99	**66.50**	**0.69**

5 Discussion

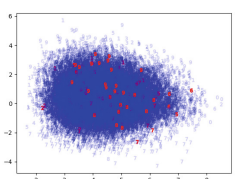
In this section, we discuss why ASR and AUC of the backdoor-assisted membership inference attack entirely deteriorate from the standpoint of latent representations in order to understand the results obtained in the previous section.

We first built a hypothesis that backdoors do not make training samples an outlier, and therefore ASR and AUC were not improved for BadNets, TaCT, and LIRA in the previous section. We observe latent representations in each attack to confirm the above hypothesis and then find the results shown in Fig. 1 and Fig. 2. We describe these results below, including a detailed analysis of LIRA and its open problem.

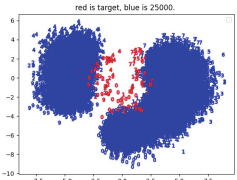
Targeted Setting. In the targeted setting, according to Fig. 1, Truth Serum makes a target sample an outlier over the distribution of training samples, which is consistent with the original work [36]. By contrast, all the backdoor attacks make target samples inliers over the distributions of training samples. It means that triggers for backdoors have independent distributions of training samples, and hence ASR and AUC were not improved as shown in Table 3. Thus, no evidence of assisting membership inference by backdoor attacks was not found in the targeted setting of this paper.

Untargeted Setting. What we were surprised is the results in the untargeted setting shown in Fig. 2. The latent representation of the Clean-Only model is close to those of BadNets and TaCT: that is, we can see clusters for each class in their latent representations. On the other hand, the latent representation of Truth Serum is close to that of LIRA, and clusters are blurred.

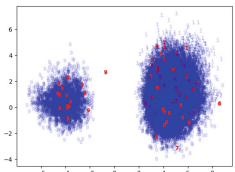
We recall that, as shown in Table 3 in the previous section, only Truth Serum and LIRA could improve ASR and AUC compared with the Clean-Only model. It



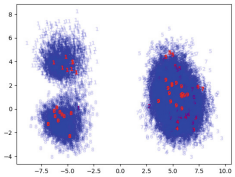
(a) Clean-Only



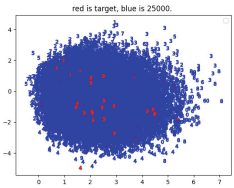
(b) Truth Serum



(c) BadNets

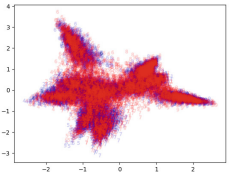


(d) TaCT

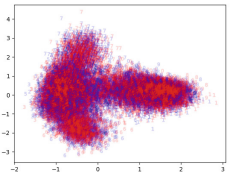


(e) LIRA

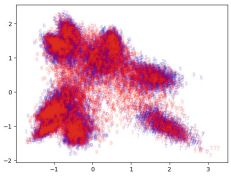
Fig. 1. Latent representations in the targeted setting: This figure visualizes latent representations of training data and non-training data of the victim model for each attack. 4000 poison samples in D_p for $r = 16$ are plotted as red points, and every 25000 samples in training data D and non-training data $\mathcal{D}\backslash D$ are plotted as blue points (Color figure online)



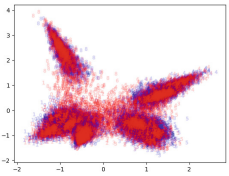
(a) Clean-Only



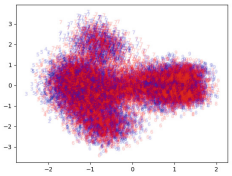
(b) Truth Serum



(c) BadNets



(d) TaCT



(e) LIRA

Fig. 2. Latent representation in the untargeted setting: The setting is common with Fig. 1. 12,500 poison samples in D_p are plotted as red points, and every 12,500 samples in training data D and non-training data $\mathcal{D}\backslash D$ are plotted as blue points. (Color figure online)

is considered that only Truth Serum and LIRA can control all other data points differently from the given poison samples, and hence their latent representations became quite different from those of the Clean-Only model (and BadNets and TaCT). Although improvements in ASR and AUC by LIRA are lower than the values we expected, it is considered that ASR and AUC are improved in comparison with the Clean-Only model if the latent representations become different from the Clean-Only model.

Detailed Analysis of LIRA. As ASR and AUC are improved by virtue of a different latent representation from the Clean-Only model for LIRA, we discuss why improvements in ASR and AUC by LIRA are lower than the values we expected, i.e., compared with Truth Serum. Although we found no strong evidence, it is considered that this phenomenon is caused by the generation of poison samples in LIRA. We describe the reason by taking the formalization in Sect. 3.1 into account as well as Truth Serum and the poisoning-assisted membership inference attack based on clean-label poisoning [6], referred to as Clean-Label Poisoning attack for the sake of convenience.

Recall that a clean sample is denoted by a tuple of (x, y) and a poison sample is denoted by a tuple of (x^*, y^*). Then, the backdoor-assisted membership inference attack, including LIRA, provides a poison sample (x^*, y^*) such that $x \neq x^*$ and $y \neq y^*$. On the other hand, Truth Serum provides (x^*, y^*) such that $x = x^*$ and $y \neq y^*$ while the Clean-Label Poisoning attack provides (x^*, y^*) such that $x \neq x^*$ and $y = y^*$.

It is considered that the differences between the poison samples in each attack significantly affect ASR and AUC. Namely, to make any clean sample an outlier, $x = x^*$ or $y = y^*$ is necessary because it will change the decision boundary related to x or y inside a victim model. In contrast, the backdoor-assisted membership inference attack provides poison samples such that $x \neq x^*$ and $y \neq y^*$, and hence it may make a different decision boundary from a victim model. Making a different decision boundary seems to be crucial to maintaining test accuracy, which is an important metric for backdoor attacks. It is thus considered that the improvement of ASR and AUC by LIRA is limited instead of maintaining the test accuracy by the differences between the poison samples described above.

Open Problem. We still have a plausible chance for the backdoor-assisted membership inference attack because LIRA could slightly improve ASR and AUC with a few points compared to the Clean-Only model. Although we leave it as an open problem to improve ASR and AUC, we describe the reason why LIRA could slightly improve ASR and AUC with a few points.

Indeed, similar phenomena were shown in the Clean-Label Poisoning attack [6]. Specifically, for the Clean-Label Poisoning attack, a poison sample x^* is generated, so that x^* is close to a clean sample x with a correct label y in the input space and close to a different sample x' with a different label $y'(\neq y)$, where the label y^* of x^* itself is equal to y. As described in the previous discussion, LIRA also makes the distance with the clean sample x in the feature space

far from its correct label y although labeling $y^*(\neq y)$ for x^* in LIRA is different from the Clean-Label Poisoning attack.

As described in the previous discussion, the reason why the improvement of ASR and AUC by LIRA is limited is that test accuracy is maintained. In other words, although the improvement of ASR and AUC was not ideal in the current experiments, we nevertheless believe that they will be improved by more increasing the size of D_p for LIRA, i.e., trigger-based methods for the imperceptible backdoor attacks. Further studies, which take the possibility of backdoor-assisted membership inference attacks into account, will need to be undertaken.

6 Conclusion

We discussed backdoor-assisted membership inference attacks, which do not deteriorate the accuracy. We first evaluated whether backdoor-assisted membership inference attacks with the original backdoors [14] and the imperceptible backdoors [9,34,45] are successful in comparison with the existing poisoning-assisted membership inference attack [36]. In contrast to the existing poisoning-assisted membership inference attack by Tramer et al. [36], we showed that backdoor-assisted membership inference attacks are unsuccessful if typical backdoors are used in a trivial way. When we analyzed their resultant models with respect to latent representation to deeply understand the reason for the unsuccessful results, we confirmed that any clean sample becomes an inlier while the existing attack makes it an outlier. We also found an interesting phenomenon where ASR and AUC of backdoor-assisted membership inference attacks were slightly improved if a trigger-based method for imperceptible backdoor attacks, i.e., LIRA [9], is used in the untargeted setting. We are in the process of improving ASR and AUC of backdoor-assisted membership inference attacks based on the trigger-based method. We also plan to evaluate backdoor-assisted membership inference attacks based on the unified attacks [43,45].

Code Availability. Our source code is publicly available for reproducibility and subsequent works (https://github.com/fseclab-osaka/backdoor-assisted-MIA).

References

1. Adi, Y., Baum, C., Cisse, M., Pinkas, B., Keshet, J.: Turning your weakness into a strength: watermarking deep neural networks by backdooring. In: Proceedings of USENIX Security 2018, pp. 1615–1631. USENIX Association (2018)
2. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramèr, F.: Membership inference attacks from first principles. In: Proceedings of IEEE S&P 2022, pp. 1897–1914. IEEE (2022)
3. Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., Song, D.: The secret sharer: evaluating and testing unintended memorization in neural networks. In: Proceedings of USENIX Security, pp. 267–284. USENIX Association (2019)
4. Carlini, N., et al: Extracting training data from large language models. In: Proceedings of USENIX Security 2021, pp. 2633–2650. USENIX Association (2021)

5. Chen, B., et al.: Detecting backdoor attacks on deep neural networks by activation clustering. In: Proceedings of SafeAI 2019 (2019)
6. Chen, Y., Shen, C., Shen, Y., Wang, C., Zhang, Y.: Amplifying membership exposure via data poisoning. CoRR abs/2211.00463 (2022). https://doi.org/10.48550/arXiv.2211.00463
7. Chou, E., Tramèr, F., Pellegrino, G.: Sentinet: detecting localized universal attacks against deep learning systems. In: Proceedings of IEEE SPW 2020, pp. 48–54. IEEE (2020)
8. Conti, M., Li, J., Picek, S., Xu, J.: Label-only membership inference attack against node-level graph neural networks. In: Proceedings of AISec 2022, pp. 1–12 (2022)
9. Doan, K., Lao, Y., Zhao, W., Li, P.: Lira: learnable, imperceptible and robust backdoor attacks. In: Proceedings of ICCV 2021, pp. 11966–11976 (2021)
10. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of CCS 2015, pp. 1322–1333. ACM (2015)
11. 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: Proceedings of USENIX Security 2014, pp. 17–32. USENIX Association (2014)
12. Ganju, K., Wang, Q., Yang, W., Gunter, C.A., Borisov, N.: Property inference attacks on fully connected neural networks using permutation invariant representations. In: Proceedings of CCS 2018, pp. 619–633. ACM (2018)
13. Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D.C., Nepal, S.: Strip: a defence against trojan attacks on deep neural networks. In: Proceedings of ACSAC 2019, pp. 113–125. ACM (2019)
14. Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S.: Badnets: evaluating backdooring attacks on deep neural networks. IEEE Access **7**, 47230–47244 (2019)
15. Henderson, P., et al.: Ethical challenges in data-driven dialogue systems. In: Proceedings of AIES 2018, pp. 123–129. ACM (2018)
16. Heo, G., Whang, S.E.: Redactor: targeted disinformation generation using probabilistic decision boundaries. CoRR abs/2202.02902 (2022). https://arxiv.org/abs/2202.02902
17. Hidano, S., Murakami, T., Katsumata, S., Kiyomoto, S., Hnnaoka, G.: Model inversion attacks for online prediction systems: without knowledge of non-sensitive attributes. IEICE Trans. Inf. Syst. **E101.D**(11), 2665–2676 (2018)
18. Hu, H., Salcic, Z., Dobbie, G., Chen, J., Sun, L., Zhang, X.: Membership inference via backdooring. In: Proceedings of IJCAI 2022, pp. 3832–3838. International Joint Conferences on Artificial Intelligence Organization (2022)
19. Jebreel, N.M., Domingo-Ferrer, J., Li, Y.: Defending against backdoor attacks by layer-wise feature analysis. In: Kashima, H., Ide, T., Peng, W.C. (eds.) PAKDD 2023. LNCS, vol. 13936, pp. 428–440. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-33377-4_33
20. Li, S., et al.: Hidden backdoors in human-centric language models. In: Proceedings of CCS 2021, pp. 3123–3140. ACM (2021)
21. Li, S., Xue, M., Zhao, B., Zhu, H., Zhang, X.: Invisible backdoor attacks on deep neural networks via steganography and regularization. IEEE Trans. Dependable Secure Comput. **18**(05), 2088–2105 (2021)
22. Li, Y., Bai, Y., Jiang, Y., Yang, Y., Xia, S., Li, B.: Untargeted backdoor watermark: towards harmless and stealthy dataset copyright protection. CoRR abs/2210.00875 (2022)

23. Li, Z., Liu, Y., He, X., Yu, N., Backes, M., Zhang, Y.: Auditing membership leakages of multi-exit networks. In: Proceedings of CCS 2022, pp. 1917–1931. ACM (2022)
24. Liu, Y., Lee, W.C., Tao, G., Ma, S., Aafer, Y., Zhang, X.: ABS: scanning neural networks for back-doors by artificial brain stimulation. In: Proceedings of CCS 2019, pp. 1265–1282. ACM (2019)
25. Liu, Y., et al.: Trojaning attack on neural networks. In: Proceedings of NDSS 2018. The Internet Society (2018)
26. Mahloujifar, S., Ghosh, E., Chase, M.: Property inference from poisoning. In: Proceedings of IEEE S&P 2022, pp. 1569–1569. IEEE (2022)
27. Ning, R., Li, J., Xin, C., Wu, H.: Invisible poison: a blackbox clean label backdoor attack to deep neural networks. In: Proceedings of INFOCOM 2021. IEEE (2021)
28. Qi, F., et al.: Hidden killer: invisible textual backdoor attacks with syntactic trigger. In: Proceedings of ACL—IJCNLP, vol. 1, pp. 443–453. ACL (2021)
29. Saha, A., Subramanya, A., Pirsiavash, H.: Hidden trigger backdoor attacks. In: Proceedings of AAAI 2020, vol. 34, pp. 11957–11965. AAAI (2020)
30. Shafahi, A., et al.: Poison frogs! targeted clean-label poisoning attacks on neural networks. In: Proceedings of NeurIPS 2018, vol. 31, pp. 6106–6116. Curran Associates, Inc. (2018)
31. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: Proceedings of IEEE S&P 2018, pp. 3–18. IEEE Computer Society (2017)
32. Song, C., Shmatikov, V.: Auditing data provenance in text-generation models. In: Proceedings of CCS 2019, pp. 196–206. ACM (2019)
33. Tan, T.J.L., Shokri, R.: Bypassing backdoor detection algorithms in deep learning. In: Proceedings of EuroS&P, pp. 175–183. IEEE (2020)
34. Tang, D., Wang, X., Tang, H., Zhang, K.: Demon in the variant: statistical analysis of dnns for robust backdoor contamination detection. In: Proceedings of Usenix Security 2021, pp. 1541–1558. USENIX Association (2021)
35. Tian, Z., Cui, L., Liang, J., Yu, S.: A comprehensive survey on poisoning attacks and countermeasures in machine learning. ACM Comput. Surv. **55**(8) (2022)
36. Tramèr, F., et al.: Truth serum: poisoning machine learning models to reveal their secrets. In: Proceedings of CCS 2022, pp. 2779–2792. ACM (2022)
37. Tran, B., Li, J., Madry, A.: Spectral signatures in backdoor attacks. In: Proceedings of NIPS 2018, pp. 8011–8021. ACM (2018)
38. Wang, B., et al.: Neural cleanse: identifying and mitigating backdoor attacks in neural networks. In: IEEE S&P 2019, pp. 707–723. IEEE (2019)
39. Wang, Z., Huang, Y., Song, M., Wu, L., Xue, F., Ren, K.: Poisoning-assisted property inference attack against federated learning. IEEE Trans. Dependable Secure Comput. 1–13 (2022)
40. Xue, M., He, C., Wang, J., Liu, W.: Backdoors hidden in facial features: a novel invisible backdoor attack against face recognition systems. Peer Peer Netw. Appl. **14**(3), 1458–1474 (2021)
41. Ye, J., Maddi, A., Murakonda, S.K., Bindschaedler, V., Shokri, R.: Enhanced membership inference attacks against machine learning models. In: Proceedings of CCS 2022, pp. 3093–3106. ACM (2022)
42. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy risk in machine learning: analyzing the connection to overfitting. In: Proceedings of CSF 2018, pp. 268–282. IEEE (2018)

43. Zhao, Z., Chen, X., Xuan, Y., Dong, Y., Wang, D., Liang, K.: Defeat: deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints. In: Proceedings of CVPR 2022, pp. 15213–15222 (2022)
44. Zhong, H., Liao, C., Squicciarini, A.C., Zhu, S., Miller, D.: Backdoor embedding in convolutional neural network models via invisible perturbation. In: Proceedings of CODASPY 2020, pp. 97–108. ACM (2020)
45. Zhong, N., Qian, Z., Zhang, X.: Imperceptible backdoor attack: from input space to feature representation. In: Raedt, L.D. (ed.) Proceedings of IJCAI 2022, pp. 1736–1742. IJCAI Organization (2022)
46. Zhu, C., Huang, W.R., Li, H., Taylor, G., Studer, C., Goldstein, T.: Transferable clean-label poisoning attacks on deep neural nets. In: Proceedings of ICML 2019, vol. 97, pp. 7614–7623. PMLR (2019)