Inference Attack and Defense on the Distributed Private Fair Machine Learning Framework

Hui Hu, Chao Lan
Department of Computer Science, University of Wyoming
hhu1@uwyo.edu, clan@uwyo.edu

Abstract

Fairness and privacy are both significant social norms in machine learning. In (Hu et al 2019), we propose a distributed framework to learn fair prediction models while protecting the privacy of user demographics. However, we did not assume an adversary who tries to infer the hidden demographics, e.g., with a good intention of building fairer models.

In this paper, we examine vulnerability of the above framework under inference attack and two defense strategies. Under mild assumptions on the attack model, we first propose an inference strategy and formulate it as an integer programming (IP) task. We show it achieves high inference accuracy when sufficient information is exchanged across the distributed parties. Then, we present two defense strategies at one party, one perturbing its evaluation of model fairness and the other randomizing its process of selecting fair models. We show they effectively defend the inference, by preventing the IP solver from returning feasible solutions, without sacrificing a significant amount of model fairness. Theoretical properties of the proposed attack and defenses strategies are briefly discussed.

Introduction

Today, fairness and privacy are two significant social norms in machine learning. Building fair models is part of the latest national AI R&D strategy plan and has been heavily invested by NSF and companies such as Amazon. On the other hand, data privacy protection is being forced in regulations such as the Europe General Data Protection Regulation (GDPR) and the latest California Consumer Privacy Act (CCPA).

The problem is that, fairness and privacy are running into a dilemma, i.e., most existing fair learners require direct access to user demographics, while these data are increasingly restricted to use by privacy regulations. For example, to train an auto-hiring model with little discrimination against HIV carriers, most fair learners need the HIV records of all individuals in the training set. Such sensitive medical records, however, are known to be extremely restricted to access by privacy regulations. Even if a company has collected some records, the record owners can request the company to delete them (and not to use them in analysis).

To address the conflict between model fairness and demographic privacy, two technical solutions are proposed in the literature, one based on cryptography (Kilbertus et al. 2018) and the other based on randomization (Hu et al. 2019). Randomization is more efficient, and we derive theoretical guarantees on model fairness and accuracy in (Hu et al. 2019). However, none of the works assumes an adversary who tries to infer the hidden demographics, e.g., with a good intention of building fairer models, or a bad intention of conducting discrimination, or other intentions of using the inferred data in demographic-sensitive applications.

In this paper, we investigate vulnerability of the proposed framework in (Hu et al. 2019) against inference attack and present two defense strategies.

Under mild assumptions on the attack model, we propose an inference strategy that looks for a set of individual demographics that are consistent with the information exchanged between the distributed parties. The inference is formulated as an integer programming task, and we show it achieves high accuracy when sufficient information is being exchanged. We also show its performance can be further enhanced through inference ensemble.

Then, we propose two strategies to defend the above inference attack. Both strategies are operated at a trusted third party, which holds the true demographic data, uses them to evaluate fairness of random models and broadcasts fair ones. Our first defense strategy perturbs fairness evaluation and the second strategy randomizes the broadcast – by this means, the broadcast models are no longer necessarily fair, which is equivalent to randomizing constraints in the above integer programming (IP) task. We show both strategies effectively defend the attack by reducing inference accuracy; in particular, when randomization is sufficient, they prevent the standard IP solver from finding any feasible solutions, without sacrificing significant model fairness and accuracy.

We also briefly discuss the potential theoretical properties of the proposed inference attack and defense strategies.

The rest of this paper is organized as follows: we first revisit our previous framework; then we present the proposed inference attack and defense strategies respectively; experimental studies and results are presented thereafter, and discussions are presented at the end.

A Revisit of (Hu et al. 2019)

Notations and Problem Setting

Let (x, s, y) be an individual, where $s \in \mathbb{R}$ is the sensitive demographic feature, $x \in \mathbb{R}^p$ is a vector of p non-sensitive features and $y \in \mathbb{R}$ is the label. For example, when studying racial bias in auto job hiring, (x, s, y) will be an applicant, s is his race, x is a vector of his non-sensitive features such as education and experience, and y indicates if he is hired. Similar to most fairness studies, assume s and y are binary.

Let $\{(x_i, s_i, y_i)\}_{i=1,\dots,n}$ be a set of n individuals, where (x_i, s_i, y_i) is the i_{th} one. We aim to learn from this set a fair model $f : \{x\} \to \{y\}$, which takes non-sensitive feature x as input and outputs label y. Importantly, the learner and f cannot have access to s, because the latter needs to be kept private, e.g., as required by privacy regulations.

The Framework

We assume the above training set is distributed over a learner and a third party that can communicate with each other. The learner holds $\{(x_i, y_i)\}_{i=1,\dots,n}$ and is responsible for training a fair model f. The party holds $\{s_i\}_{i=1,\dots,n}$ and can assist the learner via communications that do not reveal s. A schematic diagram of the framework is in Figure 1.

Let $\mathcal{H}$ be the hypothesis set from which f will be learned. Our framework works as follows.

Algorithm 1 Distributed Fair Machine Learning Framework

```
1: The learner generates $m$ random hypotheses in $\mathcal{H}$, denoted by $h_1, \dots, h_m$.
2: The learner evaluates the predicted labels of each $h_t$ on $\{(x_i)\}_{i=1,\dots,n}$, denoted by $\hat{Y}_t = [h_t(x_1), \dots, h_t(x_n)]$, and sends all $\hat{Y}_1, \dots, \hat{Y}_m$ to the third party.
3: The party estimates $cov(\hat{Y}_t, s)$ which is the sample covariance between $\hat{Y}_t$ and $\{s_i\}_{i=1,\dots,n}$ for every $h_t$.
4: The party sends index $t$ to the learner if $cov(\hat{Y}_t, s) \leq \rho$, where $\rho$ is a preset threshold.
5: Let $r_1, \dots, r_k$ be the set of indices sent to the learner. The learner constructs the final model as
       $f = \alpha_1 h_{r_1} + \dots + \alpha_k h_{r_k},$     (1)
   where $\alpha_1, \dots, \alpha_k$ are unknown coefficients.
6: The learner optimizes $\alpha_i$'s by minimizing $f$'s prediction loss on $(\{x_i, y_i\})_{i=1,\dots,n}$.
```

In this framework, one can further specify (i) hypothesis space $\mathcal{H}$, (ii) the generative distribution in Step 1 and (iii) the optimizer in Step 6. In (Hu et al. 2019), we specify four types of f, including linear regression, logistic regression, kernel regression and PCA. All model parameters are i.i.d. drawn from normal distributions. We optimize unknown coefficients by least square (for the three regression models) and variance maximizing (for PCA). In experiment, we show each plug-in gives a new prediction model which outperforms its existing non-private counterparts (Calders et al. 2013; Kamishima et al. 2012; Pérez-Suay et al. 2017; Samadi et al. 2018; Olfat and Aswani 2018) in both fairness and accuracy across several real-world data sets.

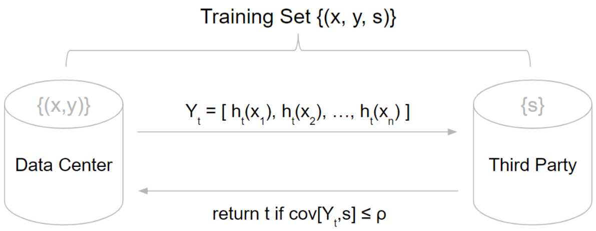
Training Set {(x, y, s)}
{(x,y)}
Data Center
Y_t = [h_t(x_1), h_t(x_2), …, h_t(x_n)]
return t if cov[Y_t,s] ≤ ρ
{s}
Third Party

Figure 1: Distributed Private Fair Learning Framework

Theoretical Properties

We have derived theoretical properties for the framework in (Hu et al. 2019). Like most studies, we will evaluate fairness of model f by a popular notion called *statistical parity* (SP) (McNamara, Ong, and Williamson 2017). It is defined as

$$\mathrm{SP}(f) = |\, p(f(x) = 1|s = 1) - p(f(x) = 1|s = 0)\, |. \quad (2)$$

Intuitively, SP measures advantaged prediction rates across different groups (divided by s). If the difference is smaller, then model prediction is considered to be more fair.

Our following fairness guarantee shows that model f has a bounded SP. Moreover, it suggests that higher model fairness can be achieved by (i) choosing smaller k or ρ, or (ii) working with a balanced demographic distribution.

Theorem 1. *In Algorithm 1, if $f(x)$ and s are positively or negatively quadrant dependent*[1]*, then*

$$SP(f) \leq \frac{\sqrt{k}||\alpha||\rho}{p(s=0)p(s=1)}, \quad (3)$$

where α is a vector of all coefficients.

Our following accuracy guarantee is derived based on a slightly modified framework which applies a *soft-threshold policy* – in Step 4, the party returns h_t with a probability proportional to $C\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(\frac{||h_t - h_*||^2}{-2\sigma_2^2}\right)$, where C and σ_2 are properly chosen constants (depending on the distribution used to generate random hypotheses in Step 1), and h_* is an ideal model satisfying $cov(h_*, s) = 0$. We then have

Theorem 2. *In Algorithm 1, suppose Step 4 adopts the soft-thresholding policy with properly chosen C and σ_2^2. Let $er(f)$ and $\hat{er}(f)$ be the expected and empirical error of f respectively. If f is linear and $||f|| = ||x|| = 1$, then with probability at least $1 - 4\delta$,*

$$er(h) \leq \hat{er}(h) + T + \frac{4}{n\delta}\sum\nolimits_{i=1}^{n} g(x_i) e^{\frac{-k\langle f, x_i\rangle^2}{8(2+||\langle f, x_i\rangle||)^2}}, \quad (4)$$

where T is constant depending on $\{k, n, \delta\}$, and $g(x_i)$ is a function of x_i depending on k and $\langle f, x_i\rangle$. Further, if

$$\left(\langle h_*, x_i\rangle^2 - 1/4\right)\left(||\langle f, x_i\rangle|| - 2\right)^2 + 1 \leq 0, \quad (5)$$

then there exist positive constants c_1 and c_2 such that

$$er(h) \leq \hat{er}(h) + c_1 + O(e^{-c_2 k}). \quad (6)$$

[1] This is a common assumption which we empirically verified.

The theorem suggests that bigger k implies smaller error, which coincides with the random projection theory (Garg, Har-Peled, and Roth 2002) we employed in analysis.

Vulnerability of the Framework

Privacy Protection in Adversarial Environment

In (Hu et al. 2019), we assume the learner obeys privacy regulation and has no malicious attempt. From Figure 1, we see the framework does protect demographic privacy, because the only information revealed to the learner is which hypothesis prediction has small covariance with demographic. The learner has no knowledge on the individual demographic.

In reality, however, the learner may be an adversary who attempts to infer the hidden demographics based on all accessible information. Her intention may be malicious, e.g., to conduct discriminatory decisions in the given or other applications. Her intention may also be kind, e.g., to simply enhance the performance of fair learning. In either case, it is crucial to examine the vulnerability of our framework under inference attack and design defense strategies when needed.

The above discussions motivate us to examine vulnerability of the distributed private fair learning framework under inference attack. In the following, we first present an inference strategy, then present two defense strategies, and at last discuss their potential theoretical properties.

An Inference Attack on Demographics

The adversary's goal is to obtain inferred demographic data

$$\hat{s} = \{\hat{s}_1, \dots, \hat{s}_n\}, \tag{7}$$

where $\hat{s}_i$ is the inferred demographic of individual x_i.

We make three assumptions on the adversary's knowledge – they can be easily satisfied in reality, e.g., the adversary is an employee at the learning center (learner).

(i) She has access to the information exchanged between the learner and the third party, including all prediction vectors $\hat{Y}_1, \dots, \hat{Y}_m$ and returned indices $r_1, \dots, r_k$.

(ii) She knows index t is returned iff $cov(\hat{Y}_t, s) \leq \rho$.

(iii) She knows the value of ρ.

Based on the assumptions, a simple way of inference is to look for an $\hat{s}$ that satisfies $cov(\hat{Y}_t, \hat{s}) \leq \rho$ for all returned indices t and $cov(\hat{Y}_{t'}, \hat{s}) > \rho$ for all unreturned indices t'. We formulate this as the following integer programming task.

$$\begin{aligned} &\max\nolimits_{\hat{s}}\ 1, \\ s.t.\ &cov(\hat{Y}_i, \hat{s}) \leq \rho, \quad && i = r_1, \dots, r_k, \\ &cov(\hat{Y}_j, \hat{s}) > \rho, && j = \{1, \dots, m\}/\{r_1, \dots, r_k\}, \\ &\hat{s}_k \in \{0, 1\}, && k = 1, \dots, n, \end{aligned} \tag{8}$$

where '/' is set difference; the first line of constraints are based on returned indices, the second based on unreturned indices, and the last based on the binary assumption of s.

To enhance performance, we further propose an ensemble inference strategy using (8) as a building block. Basically, we bootstrap $\{\hat{Y}_i\}_{i=1,\dots,m}$, apply (8) on each subset, and aggregate inference results. Details are shown in Algorithm 2.

In this algorithm, p and q are hyper-parameters.

Algorithm 2 Ensemble Demographic Inference Attack

```
1: Bootstrap $\{\hat{Y}_1, \dots, \hat{Y}_m\}$ to generate $p$ subsets of size $q$.
   Denote the $i_{th}$ subset as $\hat{Y}^{(i)} = \{\hat{Y}_{b_1}, \dots, \hat{Y}_{b_q}\}$.
2: Apply (8) on each subset $\hat{Y}^{(i)}$ to obtain an inferred demographic set, denoted by $\hat{s}^{(i)}$.
3: Aggregate $\hat{s}^{(1)}, \dots, \hat{s}^{(p)}$ in a way that, the final inferred demographic of individual $x_i$ is
       $\hat{s}_i = \sigma\left(\frac{1}{p}\sum_{j=1}^{p} \hat{s}_i^{(j)}\right)$,    (9)
   where function $\sigma$ rounds any input to its nearest integer.
```

Defense 1: Perturb Covariance Evaluation (PCE)

Our first strategy is to perturb the covariance evaluation. It is motivated by the Laplacian mechanism in differential privacy (Dwork, Roth, and others 2014). Specifically, in Algorithm 1, we replace Step 4 with the following Step 4'.

4': The party sends t to the learner if $c\tilde{o}v(\hat{Y}_t, s) \leq \rho$, where $c\tilde{o}v(\hat{Y}_t, s) = cov(\hat{Y}_t, s) + \epsilon$ is a perturbed covariance with $\epsilon \sim N(0, \sigma_2^2)$ controlled by a hyper-parameter σ_2.

The perturbation will change the set of hypotheses whose indices are sent to the learner. This will essentially change the inequality constraints in (8), and thus prevent the solver from finding accurate solutions or feasible solutions at all.

Defense 2: Soft-Threshold Policy (STP)

Our second strategy is to randomize the index selection process. It is motivated by the exponential mechanism in differential privacy (Dwork, Roth, and others 2014).

Instead of returning t when $cov(h_t, s) \leq \rho$, we now return t with higher probability if $cov(h_t, s)$ is smaller (and vice versa). Specifically, in Algorithm 1, we replace Step 4 with the following Step 4".

4": The party sends each index t to the learner with probability $p(t) = \frac{1}{1+\exp(|cov(\hat{Y}_t, s)|)}$.

The randomization will also change the set of hypotheses whose indices are sent to the learner, and essentially change the constraints in (8) and the inferred results.

Experiment

Data Preparation

We experimented on two real-world data sets: the Community Crime data set and the Credit data set.

The UCI Community Crime data set contains 1993 communities described by 101 features. Community crime rate is the label and we binarized it into 'low' and 'high'. We treated a community as minority if its fraction of African-American residents is greater than 0.5.

The UCI Credit data set contains 30000 users described by 23 features. Default payment is the label. We treated education degree as the sensitive feature and divide individuals into highly-educated and not-highly-educated.

Both our preprocessed data sets are available at[2].

Experiment Design

On Community Crime, we randomly chose 500 instances for training and 500 for testing. On Credit, we randomly chose 1000 instances for training and 1000 for testing. When comparing different methods, we will show results averaged over 20 random choices of data. When conducting sensitivity analysis, we will focus on a fixed choice.

We examined three settings of the framework: (i) standard (no defense), (ii) with PCE defense and (iii) with STP defense. We chose fair ridge regression as the base model for the framework (called DFRR), and fixed $\rho = 0.1$. We applied a popular solver (Marinescu and Dechter 2006) to solve the integer programming task (8).

For each setting, we evaluated performance based on three metrics: (a) inference attack accuracy, (b) final model prediction fairness and (c) final model prediction accuracy. We focused on non-ensemble inference attack and defense.

Experimental Results

We first present comparison results. Table 1 shows performance averaged over 20 random generations of hypothesis set H but fixed choice of training and testing data. Table 2 shows performance averaged over 20 random choices of training and testing data but fixed hypothesis set $H = \{h_1, \ldots, h_m\}$ (Step 1 in Algorithm 1). In both tables, notation '*' indicates that no feasible solution is found by the solver in any of the random trials. (Our later sensitive analysis will explain certain choice of hyper-parameters in these tables.)

We have several observations.

Observation 1. The standard framework appears vulnerable to the proposed inference attack, as the inference errors are around 21%. This further justifies the present study.

Observation 2. The proposed two defense strategies effectively prevent the attack, as they prevent the solver from finding any feasible solutions in any trials on both data sets.

Observation 3. Both defense strategies endure a trade-off between privacy and fairness, as they both have higher model disparities than the standard framework. This is partly because our framework guarantees model fairness by only using 'fair' h_t to construct f. However, both defense strategies will include certain 'unfair' h_t for the construction.

Observation 4. STP has more efficient privacy-fairness trade-off than PCE in most cases.

Observation 5. Neither defense strategy suffers a trade-off between privacy and accuracy, as all methods have similar prediction errors. This is partly because our framework guarantees model accuracy based on random projection theory, disregarding whether h_t is 'fair' or not.

Overall, we see that both defense strategies effectively enhance privacy at the cost of fairness but not accuracy, and STP seems more efficient than PCE.

Sensitivity Analysis

In this section, we analyze performance of the attack and defense strategies. We fixed the choice of training and testing data, and report results averaged over 20 random generations of the hypothesis set (same setting as in Table 1).

First, we examined the impact of m on inference error. (Recall m is the number of random hypotheses generated at the learner in Step 1.) Figure 2 shows the result with m. On the Crime Community data set, we see error decreases as m increases. This is because bigger m creates more constraints for the integer programming task (8) to find a more accurate solution. Such impact on the Credit data set seems rather limited, however, which may be because the fewer constraints are already sufficient for the solver.

Then, we examined performance of the proposed ensemble inference attack strategy. Set m = 600. Results on the two data sets are shown in Figures 3 and 4 respectively. We see ensemble improves inference accuracy by around 10% and converges at 10 bootstraps.

Next, for the PCE defense strategy, we examined the impact of additive noise on the inference error. Results are shown in Table 3 and 4 respectively. Numbers in the parenthesis are the times a feasible solution is found. We have several observations.

(i) PCE does not reduce inference error. It directly prevent the solver from finding a feasible solution if sufficient noise is added. An effective noise level for non-ensemble inference is $\sigma_2 = 0.1$, which prevents inference in most cases.

(ii) The effective noise level for ensemble inference is a bit higher than non-ensemble inference. This is because bootstrapping has a higher chance of bypassing certain noisy constraints and finding a feasible solution. This partly justifies that ensemble inference attack is more effective.

(iii) Larger m makes defense easier. This may be because larger m creates more constraints in the integer programming task, which allows PCE to create more noisy constraints to prevent inference.

Finally, for the STP defense strategy, we examined its performance versus different m. Results are shown in Table 5. We see STP effectively defends most inference.

Application on Other Private Fair Learners

We examined impact of the proposed defense strategies on other two private fair learners developed in (Hu et al. 2019): distributed fair logistic regression (DFGR) and distributed fair PCA (DFPCA). We set m = 600, fixed training and testing data and reported model fairness and error averaged over 20 random generation of the hypothesis set H. (We did not re-examine inference error as it is somewhat learner independent.) Results are shown in Tables 6, 7 and 8 respectively.

We have similar observations. For PCE, we see larger noise reduces model fairness in general, but the reduction seems not significant on DFGR and DFPCA (compared with DFRR). The defense does not sacrifice model accuracy.

[2] https://uwyomachinelearning.github.io/

		Crime			Credit	
	inference error	model disparity	model error	inference error	model disparity	model error
No Defense	.2103 ± .0347	.1680 ± .0671	.1422 ± .0158	.2126 ± .0301	.0132 ± .0043	.2365 ± .0104
PCE Defense	*	.2886 ± .0751	.1477 ± .0122	*	.0142 ± .0038	.2396 ± .0164
STP Defense	*	.1989 ± .0742	.1320 ± .0146	*	.0139 ± .0036	.2373 ± .0195

Table 1: Performance of DFRR with fixed data split and random H. (m=300, $\rho = 0.1$, PCE has $\sigma_2 = 0.1$)

		Crime			Credit	
	inference error	model disparity	model error	inference error	model disparity	model error
No Defense	.2204 ± .0555	.1849 ± .0857	.1360 ± .0124	.2258 ± .0596	.0176 ± .0059	.2476 ± .0134
PCE Defense	*	.2622 ± .0883	.1308 ± .0129	*	.0495 ± .0045	.2430 ± .0132
STP Defense	*	.2553 ± .0897	.1280 ± .0118	*	.0571 ± .0052	.2398 ± .0124

Table 2: Performance of DFRR with random data split and fixed H. (m=300, $\rho = 0.1$, PCE has $\sigma_2 = 0.1$)

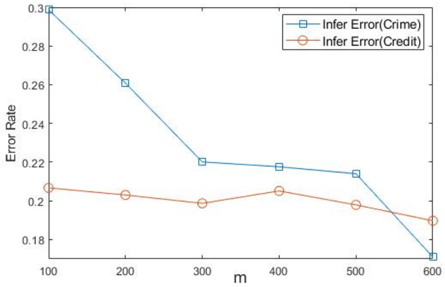

Figure 2: Inference Error versus m.

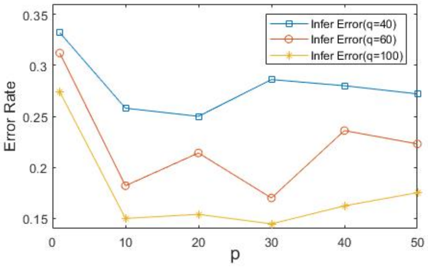

Figure 3: Error of Ensemble Inference on Crime Data Set.

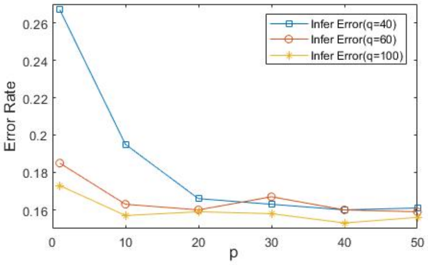

Figure 4: Error of Ensemble Inference on Credit Data Set

Related Work and Discussions

Many inference attack strategies have been proposed in the literature (Nasr, Shokri, and Houmansadr 2018; Dwork et al. 2017; Fredrikson, Jha, and Ristenpart 2015; Shokri et al. 2017; Carlini et al. 2018; Wang and Gong 2018; Wei et al. 2018). In this paper, we focus on attribute inference.

Many attribute inference strategies are developed (Salem et al. 2018; Li, Shirani-Mehr, and Yang 2007; Fredrikson et al. 2014). In this paper, we propose an adhoc inference strategy tailored for the framework. Our presented ensemble inference strategy is motivated by an existing ensemble inference (Tramèr et al. 2017).

Mechanisms to defend inference attacks have also been widely studied (Saygin, Verykios, and Clifton 2001; Cai et al. 2016; Jagielski et al. 2018; Aggarwal and Philip 2008). Our proposed defense strategies are tailored for the attack, and are largely motivated by the Laplacian and exponential mechanisms in differential privacy.

Currently, we do not have any theoretical guarantees on the presented attack and defense strategies. A naive way to derive an algorithm-independent bound for inference error may be to identify the number of feasible solutions that satisfy all constraints in (8) and simply reverse it. Defense bounds may be derived in a similar fashion as in the Laplacian and exponential mechanisms for differential privacy, although we probably need to tailor the existing theories for our special framework.

Besides the lacking of theoretical guarantees, we also realize several limitations of the present work. First, our proposed attack strategy is straightforward (albeit most natural) and based on certain assumptions on the attacker. When the assumptions are changed, e.g., if there are multiple demographics governing fairness or if the adversary is an outside with less information, one may need stronger inference strat-

Table 3: Non-Ensemble Inference Error on the Framework with PCE Defense

σ_2	m = 100		m = 200		m = 300		m = 600	
	Crime	Credit	Crime	Credit	Crime	Credit	Crime	Credit
0	0.298	0.207	0.261	0.203	0.210	0.213	0.171	0.190
0.01	0.294	0.299	0.258	0.253	0.228	*	*	*
0.1	0.289	*	*	*	*	*	*	*
1	*	*	*	*	*	*	*	*

Table 4: Ensemble Inference Error on the Framework with PCE Defense

σ_2	m = 100		m = 200		m =300		m = 600	
	Crime	Credit	Crime	Credit	Crime	Credit	Crime	Credit
0	0.272	0.189	0.252	0.183	0.212	0.181	0.142	0.176
0.01	0.284	0.193	0.256	0.185	0.276	0.174	*	0.157
0.1	0.27 (1)	*	0.242 (1)	*	*	*	*	*
1	*	*	*	*	*	*	*	*

Table 5: Inference Error on the Framework with STP Defense

Method	m = 100		m = 200		m = 300		m = 600	
	Crime	Credit	Crime	Credit	Crime	Credit	Crime	Credit
Non-Ensemble	*	*	*	*	*	*	*	*
Ensemble	0.324	*	0.298 (1)	*	*	*	*	*

Table 6: Performance with PCE Defense on Crime

Method	$\sigma_2 = 0$		$\sigma_2 = 0.01$		$\sigma_2 = 0.1$	
	SP	Err	SP	Err	SP	Err
DFGR	.056	.189	.043	.216	.057	.226
DFPCA	.025	.147	.025	.149	.029	.148

Table 7: Performance with PCE Defense on Credit

Method	$\sigma_2 = 0$		$\sigma_2 = 0.01$		$\sigma_2 = 0.1$	
	SP	Err	SP	Err	SP	Err
DFGR	.062	.263	.058	.266	.069	.268
DFPCA	.067	.243	.081	.244	.077	.246

egy. Second, although our defense strategies are effective, they seem to suffer significant trade-off between privacy and fairness (especially PCE). How to design a more efficient defense strategy with minimal trade-off is an open question.

Conclusion

In this paper, we investigated the vulnerability of our previously published distributed and private fair learning framework under inference attack. We propose an inference attack strategy which is formulated as an integer programming (IP) task. We also propose two defense strategies PCE and STP that respectively mimic the Laplacian and exponential mechanisms in differential privacy. In experiment, we show the standard framework is indeed vulnerable to the proposed attack, which achieves 10- 20% inference error. We also show the proposed defense strategies can effectively prevent the standard IP solver from finding any feasible solutions. They also suffer trade-off between privacy protection and fairness, by increasing some model disparity.

Table 8: Performance with STP on the Two Data Set

Method	Crime		Credit	
	SP	Err	SP	Err
DFGR	.099	.164	.068	.241
DFPCA	.047	.144	.081	.206

Acknowledgement

We thank the anonymous reviewers for their valuable suggestions on improving the quality of this work. We also thank the anonymous reviewer on our ICDM’19 paper for motivating this study. This work is supported in part by the US National Science Foundation under NSF III-1850418.

References

Aggarwal, C. C., and Philip, S. Y. 2008. A survey of randomization methods for privacy-preserving data mining. In *Privacy-Preserving Data Mining*. Springer. 137–156.

Cai, Z.; He, Z.; Guan, X.; and Li, Y. 2016. Collective data-sanitization for preventing sensitive information inference attacks in social networks. *IEEE Transactions on Dependable and Secure Computing* 15(4):577–590.

Calders, T.; Karim, A.; Kamiran, F.; Ali, W.; and Zhang, X. 2013. Controlling attribute effect in linear regression. In *ICDM*.

Carlini, N.; Liu, C.; Kos, J.; Erlingsson, Ú.; and Song, D. 2018. The secret sharer: Measuring unintended neural network memorization & extracting secrets. *arXiv preprint arXiv:1802.08232*.

Dwork, C.; Smith, A.; Steinke, T.; and Ullman, J. 2017. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application* 4:61–84.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9(3–4):211–407.

Fredrikson, M.; Lantz, E.; Jha, S.; Lin, S.; Page, D.; and Ristenpart, T. 2014. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, 17–32.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1322–1333. ACM.

Garg, A.; Har-Peled, S.; and Roth, D. 2002. On generalization bounds, projection profile, and margin distribution. In *ICML*.

Hu, H.; Liu, Y.; Wang, Z.; and Lan, C. 2019. A distributed fair machine learning framework with private demographic data protection. In *International Conference on Data Mining*.

Jagielski, M.; Kearns, M.; Mao, J.; Oprea, A.; Roth, A.; Sharifi-Malvajerdi, S.; and Ullman, J. 2018. Differentially private fair learning. *arXiv preprint arXiv:1812.02696*.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *ECMLPKDD*.

Kilbertus, N.; Gascon, A.; Kusner, M.; Veale, M.; Gummadi, K. P.; and Weller, A. 2018. Blind justice: Fairness with encrypted sensitive attributes. In *ICML*.

Li, C.; Shirani-Mehr, H.; and Yang, X. 2007. Protecting individual information against inference attacks in data publishing. In *International Conference on Database Systems for Advanced Applications*, 422–433. Springer.

Marinescu, R., and Dechter, R. 2006. And/or branch-and-bound search for pure 0/1 integer linear programming problems. In *International Conference on Integration of Artificial Intelligence (AI) and Operations Research (OR) Techniques in Constraint Programming*, 152–166. Springer.

McNamara, D.; Ong, C. S.; and Williamson, R. C. 2017. Provably fair representations. *CoRR*.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 634–646. ACM.

Olfat, M., and Aswani, A. 2018. Convex formulations for fair principal component analysis. *CoRR*.

Pérez-Suay, A.; Laparra, V.; Mateo-García, G.; Muñoz-Marí, J.; Gómez-Chova, L.; and Camps-Valls, G. 2017. Fair kernel learning. In *ECMLPKDD*.

Salem, A.; Zhang, Y.; Humbert, M.; Berrang, P.; Fritz, M.; and Backes, M. 2018. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*.

Samadi, S.; Tantipongpipat, U.; Morgenstern, J. H.; Singh, M.; and Vempala, S. 2018. The price of fair pca: One extra dimension. In *NIPS*.

Saygin, Y.; Verykios, V. S.; and Clifton, C. 2001. Using unknowns to prevent discovery of association rules. *ACM Sigmod Record* 30(4):45–54.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18. IEEE.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2017. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*.

Wang, B., and Gong, N. Z. 2018. Stealing hyperparameters in machine learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, 36–52. IEEE.

Wei, L.; Luo, B.; Li, Y.; Liu, Y.; and Xu, Q. 2018. I know what you see: Power side-channel attack on convolutional neural network accelerators. In *Proceedings of the 34th Annual Computer Security Applications Conference*, 393–406. ACM.