Privacy Threat and Defense for Federated Learning with Non-i.i.d. Data in AIoT

Zuobin Xiong, Zhipeng Cai *Senior Member, IEEE,* Daniel Takabi *Member, IEEE,* and Wei Li *Member, IEEE*

***Abstract*—Under the needs of processing huge amounts of data, providing high-quality service, and protecting user privacy in Artificial Intelligence of Things (AIoT), federated learning (FL) has been treated as a promising technique to facilitate distributed learning with privacy protection. Although the importance of developing privacy-preserving FL has attracted a lot of attentions, the existing research only focuses on FL with independent identically distributed (i.i.d.) data and lacks study of non-i.i.d. scenario. Whats' worse, the assumption of i.i.d. data is impractical, reducing the performance of privacy protection in real applications. In this paper, we carry out an innovative exploration of privacy protection in FL with non-i.i.d. data. First, a thorough analysis on privacy leakage in FL is conducted with proving the performance upper bound of privacy inference attack. Based on our analysis, a novel algorithm, 2DP-FL, is designed to achieve differential privacy by adding noise during training local models and when distributing global model. Especially, our 2DP-FL algorithm has a flexibility of noise addition to meet various needs and has a convergence upper bound. Finally, the real-data experiments can validate the results of our theoretical analysis and the advantages of 2DP-FL in privacy protection, learning convergence, and model accuracy.**

***Index Terms*—Privacy Protection, Federated Learning, Differential Privacy, Convergence Analysis, Artificial Intelligence of Things (AIoT)**

I. Introduction

The explosive progress and widespread deployment of Internet of Things (IoT) are being leveraged to advanced ubiquitous data sensing and collection across every corner in our life. In addition, with the growing demand for high-quality customized services by IoT users, IoT is desired to be endowed with more powerful learning capacities by Artificial Intelligence (AI) to process the enormous amount of data in data-hungry applications, such as smart city, smart transportation, and smart healthcare. During past decades, machine learning methods have been typically trained in a centralized manner via collecting all the generated data to a central server, which performs well for accuracy but fail to satisfy the needs of Artificial Intelligence of Things (AIoT) due to its essential flaws: (i) data collection for such an amount of data brings expensive cost to communications; (ii) single point failure threats the centralized storage and model easily once attackers are able to access the server; and (iii) with users' ever-increasing privacy awareness and governments' sophisticated privacy regulations, it becomes harder to encourage users to contribute their valuable private data to central storage and processing. All of these above issues raise unprecedented challenges for machine learning in AIoT – *how to effectively and efficiently learn information from massive data without unexpected privacy leakage?*

Fortunately, the advent of distributed learning technologies provide us with promising solutions, among which federated learning (FL) [1] is one of the most eye-catching paradigms. In FL, geographically distributed participants collaboratively learn a global model over their local datasets by sharing their local outputs for aggregation, which significantly reduce communication cost (*e.g.*, bandwidth and transmission time) and mitigate privacy leakage from participants' raw data. However, though FL separates data and models with privacy consideration, it is far away from perfect privacy protection. As pointed out by prior research, FL still suffers malicious attack aiming at stealing private information, such as membership privacy [2], model privacy [3], and attribute privacy [4]. On the other hand, in order to resist the threats of privacy leakage, secure multi-party computation (SMC) [5] and differential privacy (DP) [6], [7], [8], [9], [10] have been widely employed to design various privacy-preserving FL schemes. Nevertheless, the existing works on privacy threats and privacy protection in FL are limited to the impractical assumption that clients' datasets are independent identically distributed (i.i.d.). Until now, only few works on FL consider the non-i.i.d. data, and none of them is related to privacy issue. In fact, the generated data in AIoT usually has highly skewed distribution and even belongs to different data domains; that is, clients are likely to own non-i.i.d. datasets.

For the purpose of better protecting private information, this paper intends to fill the gap of investigating privacy-preserving FL under non-i.i.d. scenario. Our research endeavor starts with a comprehensive and deep analysis on the issue of privacy leakage in the original FL system by taking into account both passive and active privacy inference attack. Through our theoretical proof, the performance upper bound of privacy inference in FL is obtained, and the influence of FL scenario (including clients' data size and the difference of data distribution, *etc.*) on such a performance upper bound is clearly analyzed. Furthermore, a Dual Differentially Private FL (2DP-FL) is elaborately designed to defend privacy inference attack while guaranteeing a convergence upper bound. Particularly, with the flexible noise addition, our 2DP-FL mechanism can meet the different needs for privacy protection. In the real-data experiments, the feasibility of our considered inference attack, the effectiveness of our 2DP-FL mechanism, and the superiority of our 2DP-FL mechanism over the state-of-the-

Zuobin Xiong, Zhipeng Cai, Daniel Takabi, and Wei Li are with the Department of Computer Science, Georgia State University, Atlanta, GA, 30303 USA. E-mail: zxiong2@student.gsu.edu, {zcai, takabi, wli28}@gsu.edu.

Corresponding Author: Zhipeng Cai

art are well validated. The contributions of this paper are summarized as follows:

- To our best knowledge, this is the first work to theoretically analyze privacy leakage in FL with non-i.i.d. setting, in which the upper bound of inferring privacy is obtained.
- To defend privacy inference attack, we propose a DP-based FL mechanism, 2DP-FL.
- Our 2DP-FL mechanism is proved to be effective with a convergence upper bound.
- Intensive experiments are conducted to eveluate our theoretical analysis on privacy leakage as well as the advantages of 2DP-FL mechanism in achieving convergence, protecting privacy, and maintaining model accuracy.

This paper is organized as follows. Related works are introduced in Section II. We first analyze privacy leakage in FL in Section IV and then present our 2DP-FL mechanism in Section V. Real-data experiments are conducted in Section VI. After that, this paper is concluded in Section VII.

II. Related Works

As FL has attracted more and more attentions of research and application, various vulnerabilities of FL models have been explored to launch attacks, mainly including inference attack [11], [12], [2] and poisoning attack [13], [14], [15]. To learn local users' data privacy, Melis *et al.* [11] deveoped membership inference attack by using non-zero gradients of the embedding layer of a deep natural language processing model. A Generative Adversarial Networks (GAN)-based active inference attack was designed by Hitaj *et al.* [12] to generate targeted private samples of the victim client. In [2], authors reviewed the privacy leakage in FL and then developed an inference attack model via using each layer's gradient of the target model. Data poisoning attack of [13] modified the training data through flipping data label and changing features or small regions. In [14], model poisoning attack embedded a global backdoor trigger in FL models, which is achieved by inserting hidden backdoors into a subset of local clients before updating to the server. In [15], by modeling the interactions between training loss and attack performance as an adversarial min-max game, the authors designed model poisoning attack to bypass the poisoning detection tool of FL systems. However, most of the attack models are experiment-oriented and lack theoretical analysis on the attack factors and performance.

To protect data privacy in FL, secure multi-party computation (SMC) [5] and differential privacy (DP) [6] are commonly used solutions. Although SMC offers a strong security guarantee, the complicated computation protocols yield potentially unaffordable overheads for small devices, such as mobile devices. Existing works incorporate DP into FL from different aspects [7], [8], [9], [10]. McMahan *et al.* [7] introduced the first DP-based FL proposal for protecting the privacy of a recurrent language model. In [8], an asynchronous FL was designed for mobile edge computing in urban informatics using local differential privacy to protect the privacy of self-driving vehicles. Agarwal *et al.* [9] studied the optimal communication cost with binomial mechanism for FL under certain DP conditions. In [10], DP-based noise was added twice for data privacy in FL – the first time is after training local client models and before updating local model parameters, and the second time is during the process of parameter aggregation. But, all of these current works only focus on FL with i.i.d. scenario.

TABLE I
Notations used in this paper

Notation	Definition
X	Feature space of data
Y	Label space of data
L	Loss function of federated model
L^k	Loss function of k-th local client model
∇L	Gradient of Loss function L
∇L^k	Gradient of Loss function L^k
p	Global data distribution of X
p^k	Data distribution of k-th client
$\mathcal{F}$	Learning function of model
$\mathcal{F}_i$	i-th digit of the learning function $\mathcal{F}$
K	Number of local clients
D^k	Training dataset on k-th local client
n^k	The number of data in D^k
w_t^f	Model parameter of federated model at time t
w_t^k	Model parameter of local client k at time t
$\tilde{w}_t^f$	DP federated model at t when server noise is 0
$\tilde{w}_t^k$	DP model of client k at t when server noise is 0
$\dot{w}_t^f$	DP federated model in our method
$\dot{w}_t^k$	DP model of client k in our method
N_t^f	DP noise added by server at time t
N_t^k	DP noise added by local client at time t

III. System Model & Problem Formulation

FL is a distributed learning paradigm that allows geographically distributed participants to follow a common training procedure with the same objective and loss functions to build a federated model on a server using their local datasets. The federated model parameter is learnt by aggregating local participants' model parameters through FedAvg algorithm [1]: $w_t^f = \sum_{k=1}^{K} \frac{n^k}{|\mathcal{D}|} w_t^k$, where w_t^f and w_t^k are the federated model parameter and client k's model parameter, respectively, n^k is the size of client k's dataset, and $|\mathcal{D}| = \sum_{k=1}^{K} n^k$.

As shown in Fig. 1, in this paper, we consider that a federated learning model $\mathcal{F}$ is trained for C classes on the dataset (X, Y), where X is the feature space, and $Y = \{1, \dots C\}$ is the set of all class labels. The classifier in FL is reversible, such as liner regression and logistic regression. The goal of federated learning is to obtain an optimized model parameter w_t^f that minimizes the loss function in Eq. (1).

$$L(w_t^f) \triangleq \sum_{k=1}^{K} \frac{n^k}{|\mathcal{D}|} \sum_{i=1}^{C} p^k(y=i) \mathbb{E}_{x|y=i}[\log \mathcal{F}_i(x, w_t^f)], \tag{1}$$

where $\mathcal{F}_i$ denotes the probability of a datapoint belonging to the i-th class of Y.

It is worth noticing that in this paper, all clients' local datasets are non-i.i.d, *i.e.*, D^k is non-i.i.d. For the expected training goal, each client k optimizes his local model parameter w_t^k to minimize the loss function $L^k(w_t^k)$ on the local dataset that follows distribution p^k, *i.e.*,

$$\begin{aligned} L^k(w_t^k) =& \mathbb{E}_{(x,y)\sim p^k}[\sum_{i=1}^{C} \mathbf{1}_{y=i} \log \mathcal{F}_i(x, w_t^k)] \\ =& \sum_{i=1}^{C} p^k(y=i) \mathbb{E}_{x|y=i}[\log \mathcal{F}_i(x, w_t^k)]. \end{aligned} \tag{2}$$

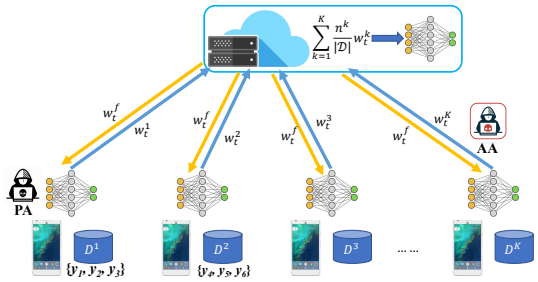


Fig. 1. The framework of federated learning (FL).

To obtain the optimal parameter w_t^k, gradient descent-based method is used to solve the optimization problem iteratively with the following equation:

$$w_t^k = w_{t-1}^k - \eta \nabla_w L^k(w_{t-1}^k) \tag{3}$$
$$= w_t^k - \eta \sum_{i=1}^{C} p^k(y=i) \nabla_w \mathbb{E}_{x|y=i}[\log \mathcal{F}_i(x, w_{t-1}^k)],$$

where η is the selected learning rate.

Theoretically, four common assumptions are considered to facilitate performance analysis on FL [16], [10], [17].

1) **Bounded and Unbiased Gradient.** The gradient of each local loss function $\nabla L^k(w)$ is bounded, and the estimator of global loss function's gradient $\nabla L(w)$ is unbiased:

$$\|\nabla L^k(w)\| \leq G; \nabla L(w) = \mathbb{E}\{\nabla L^k(w)\}. \tag{4}$$

2) **Lipschitz Continuity.** The global loss function $L(w)$ is Lipschitz continuous:

$$\|L(w) - L(\bar{w})\| \leq \lambda \|w - \bar{w}\|, \tag{5}$$

where λ is Lipschitz constant.

3) **Lipschitz Continuous Gradient.** The gradient of each local loss function $L^k(w)$ is Lipschitz continuous:

$$\|\nabla L^k(w) - \nabla L^k(\bar{w})\| \leq \mu \|w - \bar{w}\|, \tag{6}$$

where μ is Lipschitz constant.

4) **Polyak-Lojasiewicz (PL) inequality.** The global loss function $L(w)$ has strong convexity and satisfies Polyak-Lojasiewicz (PL) inequality:

$$\tau(L(w) - L(w^*)) \leq \frac{1}{2}\|\nabla L(w)\|^2, \tag{7}$$

where w^* is the optimal model parameter.

IV. Privacy Leakage of Federated Learning

A. Privacy Leakage Analysis

Although every client and the server in FL cannot directly access others' local data, private information can still be inferred from the shared model parameters. Especially, at the end of FL, every client holds w_t^f that contains the information about other clients and can be used to infer other clients' privacy via passive attack and/or active attack.

Typically, a classifier $\mathcal{C}$ is represented as a parametric function: $\mathcal{C}(x, w) = y$. If there exists an inverse function, we can compute $x = \mathcal{C}^{-1}(y, w)$. As a result, given the model parameter w and output label y of $\mathcal{C}$, the corresponding input x can be inferred. Under this situation, any client k in FL is able to learn other clients' private information when knowing $\mathcal{C}^{-1}$ and w_t^f. For example, as a type of preimage privacy attack, model inversion attack [18] can recover the input data in FL.

Theorem 1: Given a classifier $\mathcal{C}$ and an output label y, if $\mathcal{C}$ is reversible and *Lipschitz* continuous, the distance between the real input x and the inferred input x' has an upper bound: $\|x - x'\| \leq \lambda \|w - w'\|$, where λ is the *Lipschitz* constant, w is the real model parameter, w' is the parameter used for inference attack.

Proof: For a reversible discriminative model $\mathcal{C}$, there are $\mathcal{C}(x, w) = y$ and $x = \mathcal{C}^{-1}(y, w)$. Accordingly, the inference result is $x' = \mathcal{C}^{-1}(y, w')$. From the *Lipschitz* continuity of $\mathcal{C}$, we have $\|x - x'\| = \|\mathcal{C}^{-1}(y, w) - \mathcal{C}^{-1}(y, w')\| \leq \lambda \|w - w'\|$. ■

Furthermore, Theorem 1 can be extended to a generic attack scenario. Any honest-but-curious client k can infer other clients' data by analyzing w_t^f and/or w_t^k. What's worse, the inference performance of any client k has an upper bound, which is demonstrated in Theorem 2 and Theorem 3.

Theorem 2: Given K clients in federated learning, each client k's local dataset has a size n^k and a distribution p^k. If $\nabla_w \mathbb{E}_{x|y=i}[\log \mathcal{F}_i(x, w)]$ is $\alpha_{x|y=i}$-Lipschitz for $\forall i \in Y$, and each local model parameter w_{mT}^k is updated to the server every m local iterations, then the distance between the federated model parameter w_{mT}^f and any target local model parameter w_{mT}^j after T updates is upper bounded by Eq. (8):

$$\|w_{mT}^f - w_{mT}^j\| \leq \sum_{k=1}^{K} \frac{n^k}{|\mathcal{D}|}[(b^k)^m \|w_{m(T-1)}^j - w_{m(T-1)}^k\| \tag{8}$$
$$+ \eta \sum_{i=1}^{C} \|p^j(y=i) - p^k(y=i)\|(\sum_{l=0}^{m-1} (b^k)^l g_{max}(w_{mT-1-l}^j))],$$

where $g_{max}(\cdot)$ is the maximal gradient of model parameter.

Proof: To prove this theorem, there are two cases for consideration: $m = 1$ and $m > 1$.

When $m = 1$, Eq. (9) can be obtained.

$$\|w_{mT}^f - w_{mT}^j\| = \|\sum_{k=1}^{K} \frac{n^k}{|\mathcal{D}|} w_{mT}^k - w_{mT}^j\| \tag{9}$$
$$= \|\sum_{k=1}^{K} \frac{n^k}{|\mathcal{D}|}(w_{mT-1}^k - \eta \sum_{i=1}^{C} p^k(y=i) \nabla_w \mathbb{E}_{x|y=i}[\log \mathcal{F}_i(x, w_{mT-1}^k)])$$
$$- w_{mT-1}^j + \eta \sum_{i=1}^{C} p^j(y=i) \nabla_w \mathbb{E}_{x|y=i}[\log \mathcal{F}_i(x, w_{mT-1}^j)]\|$$
$$\leq \|\sum_{k=1}^{K} \frac{n^k}{|\mathcal{D}|} w_{mT-1}^k - w_{mT-1}^j\|$$
$$+ \|\eta \sum_{i=1}^{C} p^j(y=i) \nabla_w \mathbb{E}_{x|y=i}[\log \mathcal{F}_i(x, w_{mT-1}^j)]$$
$$- \sum_{k=1}^{K} \frac{n^k}{|\mathcal{D}|} \eta \sum_{i=1}^{C} p^k(y=i) \nabla_w \mathbb{E}_{x|y=i}[\log \mathcal{F}_i(x, w_{mT-1}^k)]\|.$$

For the first term at the right side of the inequality in Eq. (9), we have $\|\sum_{k=1}^{K} \frac{n^k}{|\mathcal{D}|} w_{mT-1}^k - w_{mT-1}^j\| = \sum_{k=1}^{K} \frac{n^k}{|\mathcal{D}|} \|w_{mT-1}^k - w_{mT-1}^j\|$. The second term at the right side of the inequality in Eq. (9) can be rewritten as:

$$\|\eta \sum_{i=1}^{C} p^j(y=i) \nabla_w \mathbb{E}_{x|y=i}[\log \mathcal{F}_i(x, w_{mT-1}^j)] \tag{10}$$
$$- \sum_{k=1}^{K} \frac{n^k}{|\mathcal{D}|} \eta \sum_{i=1}^{C} p^k(y=i) \nabla_w \mathbb{E}_{x|y=i}[\log \mathcal{F}_i(x, w_{mT-1}^k)]\|$$

$$
\begin{aligned}
=&\eta\sum_{k=1}^{K}\frac{n^k}{|\mathcal{D}|}\|\sum_{i=1}^{C}p^j(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^j_{mT-1})]\\
&-\sum_{i=1}^{C}p^k(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^k_{mT-1})]\|\\
=&\eta\sum_{k=1}^{K}\frac{n^k}{|\mathcal{D}|}\|\sum_{i=1}^{C}p^j(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^j_{mT-1})]\\
&-\sum_{i=1}^{C}p^k(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^j_{mT-1})]\\
&+\sum_{i=1}^{C}p^k(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^j_{mT-1})]\\
&-\sum_{i=1}^{C}p^k(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^k_{mT-1})]\|\\
\leq&\eta\sum_{k=1}^{K}\frac{n^k}{|\mathcal{D}|}[\sum_{i=1}^{C}\|p^j(y=i)-p^k(y=i)\|\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^j_{mT-1})]\\
&+\sum_{i=1}^{C}p^k(y=i)(\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^j_{mT-1})]\\
&-\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^k_{mT-1})])].
\end{aligned}
$$

Since $\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w)]$ is $\alpha_{x|y=i}$-Lipschitz for $\forall i\in Y$, Eq. (10) can be written as:

$$
\begin{aligned}
&\eta\sum_{k=1}^{K}\frac{n^k}{|\mathcal{D}|}[\sum_{i=1}^{C}\|p^j(y=i)-p^k(y=i)\|\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^j_{mT-1})]\\
&+\sum_{i=1}^{C}p^k(y=i)\alpha_{x|y=i}\|w^j_{mT-1}-w^k_{mT-1}\|] \qquad (11)\\
\leq&\eta\sum_{k=1}^{K}\frac{n^k}{|\mathcal{D}|}[\sum_{i=1}^{C}\|p^j(y=i)-p^k(y=i)\|g_{max}(w^j_{mT-1})\\
&+\sum_{i=1}^{C}p^k(y=i)\alpha_{x|y=i}\|w^j_{mT-1}-w^k_{mT-1}\|],
\end{aligned}
$$

where $g_{max}(\cdot)$ is the largest gradient of weight matrix w^j_{mT-1}.

By combining Eq. (10) and Eq. (11), Eq. (9) can be equivalently simplified as Eq. (8), *i.e.*,

$$
\begin{aligned}
&\|w^f_{mT}-w^j_{mT}\|\\
\leq&\sum_{k=1}^{K}\frac{n^k}{|\mathcal{D}|}[(1+\eta\sum_{i=1}^{C}p^k(y=i)\alpha_{x|y=i})\|w^j_{mT-1}-w^k_{mT-1}\|\\
&+\eta\sum_{i=1}^{C}\|p^j(y=i)-p^k(y=i)\|g_{max}(w^j_{mT-1})].
\end{aligned}
$$

That is, this theorem holds when m=1.

When $m>1$, additional analysis is addressed below.

$$
\begin{aligned}
&\|w^j_{mT-1}-w^k_{mT-1}\|\\
=&\|w^j_{mT-2}-\eta\sum_{i=1}^{C}p^j(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^j_{mT-2})]\\
&-w^k_{mT-2}+\eta\sum_{i=1}^{C}p^k(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^k_{mT-2})]\|\\
\leq&\|w^j_{mT-2}-w^k_{mT-2}\|\\
&+\eta\|\sum_{i=1}^{C}p^k(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^k_{mT-2})]\\
&-\sum_{i=1}^{C}p^j(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^j_{mT-2})]\|.
\end{aligned}
$$

Following the same analysis in Eq. (11), we get Eq. (12).

$$
\begin{aligned}
&\|w^j_{mT-1}-w^k_{mT-1}\|\\
\leq&(1+\eta\sum_{i=1}^{C}p^k(y=i)\alpha_{x|y=i})\|w^j_{mT-2}-w^k_{mT-2}\|\\
&+\eta\sum_{i=1}^{C}\|p^j(y=i)-p^k(y=i)\|g_{max}(w^j_{mT-2}). \qquad (12)
\end{aligned}
$$

Let $b^k=(1+\eta\sum_{i=1}^{C}p^k(y=i)\alpha_{x|y=i})$. Since b^k is a constant greater than 1, Eq. (12) is rewritten as:

$$
\begin{aligned}
&\|w^j_{mT-1}-w^k_{mT-1}\| \qquad (13)\\
\leq&b^k\|w^j_{mT-2}-w^k_{mT-2}\|+\eta\sum_{i=1}^{C}\|p^j(y=i)-p^k(y=i)\|g_{max}(w^j_{mT-2})\\
\leq&(b^k)^2\|w^j_{mT-3}-w^k_{mT-3}\|+\\
&\eta\sum_{i=1}^{C}\|p^j(y=i)-p^k(y=i)\|(g_{max}(w^j_{mT-2})+b^kg_{max}(w^j_{mT-3}))\\
&...\\
\leq&(b^k)^{m-1}\|w^j_{m(T-1)}-w^k_{m(T-1)}\|+\\
&\eta\sum_{i=1}^{C}\|p^j(y=i)-p^k(y=i)\|(\sum_{l=0}^{m-2}(b^k)^lg_{max}(w^j_{mT-2-l}))
\end{aligned}
$$

By substituting Eq. (13) into Eq. (8), the theorem holds when $m>1$. Therefore, the theorem is proved. ■

Theorem 2 states that the distance between the federated model parameter w^f_{mT} and the target client j's model parameter w^j_{mT} is upper bounded by two factors: (i) the difference of data distribution between client j and other clients; and (ii) the maximum gradient value of client j during training. When an honest-but-curious client intends to attack client j using w^f_t, the inference error $\|x-x'\|$ should also be restricted by the above two factors according to Theorem 1. Moreover, this finding can be used to improve privacy protection against attackers' inference when the data distribution is known.

Theorem 3: Given K clients in federated learning, each client k's local dataset has a size n^k and a distribution p^k. If $\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w)]$ is $\alpha_{x|y=i}$-Lipschitz for $\forall i\in Y$, and each local model parameter w^k_{mT} is updated every m local iterations, then the distance between any two local model parameters, w^u_{mT} and w^v_{mT}, after T updates is upper bounded by Eq. (14):

$$
\begin{aligned}
&\|w^u_{mT}-w^v_{mT}\|\leq(b^v)^m\|w^v_{m(T-1)}-w^u_{m(T-1)}\| \qquad (14)\\
&+\eta\sum_{i=1}^{C}\|p^v(y=i)-p^u(y=i)\|(\sum_{l=0}^{m-1}(b^v)^lg_{max}(w^u_{mT-1-l})).
\end{aligned}
$$

Proof:

$$
\begin{aligned}
&\|w^u_{mT}-w^v_{mT}\| \qquad (15)\\
=&\|w^u_{mT-1}-\eta\sum_{i=1}^{C}p^u(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^u_{mT-1})]\\
&-w^v_{mT-1}+\eta\sum_{i=1}^{C}p^v(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^v_{mT-1})]\|\\
\leq&\|w^u_{mT-1}-w^v_{mT-1}\|\\
&+\|\eta\sum_{i=1}^{C}p^v(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^v_{mT-1})]\\
&-\eta\sum_{i=1}^{C}p^u(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^u_{mT-1})]\|.
\end{aligned}
$$

For the second term at the right side of inequality in Eq. (15), we have

$$\begin{aligned}&\|\eta\sum_{i=1}^{C}p^v(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^v_{mT-1})] \\ &-\eta\sum_{i=1}^{C}p^u(y=i)\nabla_w\mathbb{E}_{x|y=i}[\log\mathcal{F}_i(x,w^u_{mT-1})]\| \\ \leq&\eta\|\sum_{i=1}^{C}p^v(y=i)\alpha_{x|y=i}\|w^v_{mT-1}-w^u_{mT-1}\| \\ &+\eta\sum_{i=1}^{C}\|p^v(y=i)-p^u(y=i)\|g_{max}(w^u_{mT-1}).\end{aligned} \tag{16}$$

Combining Eq. (15) and Eq. (16), we get a new inequality:

$$\begin{aligned}&\|w^u_{mT}-w^v_{mT}\| \\ \leq&(1+\eta\sum_{i=1}^{C}p^v(y=i)\alpha_{x|y=i})\|w^v_{mT-1}-w^u_{mT-1}\| \\ &+\eta\sum_{i=1}^{C}\|p^v(y=i)-p^u(y=i)\|g_{max}(w^u_{mT-1}).\end{aligned} \tag{17}$$

For simplicity, we denote $b^v=(1+\eta\sum_{i=1}^{C}p^v(y=i)\alpha_{x|y=i})$ and obtain the following result:

$$\begin{aligned}&\|w^u_{mT}-w^v_{mT}\| \\ \leq&b^v\|w^v_{mT-1}-w^u_{mT-1}\|+\eta\sum_{i=1}^{C}\|p^v(y=i)-p^u(y=i)\|g_{max}(w^u_{mT-1}) \\ \leq&(b^v)^2\|w^v_{mT-2}-w^u_{mT-2}\|+ \\ &\eta\sum_{i=1}^{C}\|p^v(y=i)-p^u(y=i)\|(g_{max}(w^u_{mT-1})+b^kg_{max}(w^u_{mT-2})) \\ &... \\ \leq&(b^v)^m\|w^v_{m(T-1)}-w^u_{m(T-1)}\|+ \\ &\eta\sum_{i=1}^{C}\|p^v(y=i)-p^u(y=i)\|(\sum_{l=0}^{m-1}(b^v)^lg_{max}(w^u_{mT-1-l})).\end{aligned}$$

■

Theorem 3 implies that the distance between two local clients' model parameters, w^u_{mT} and w^v_{mT}, is upper bounded by two factors: (i) the difference of data distribution between clients u and v; (ii) the maximum gradient of the honest-but-curious client u during training process. Accordingly, when client u acts as an attacker using his local model to perform inference attack towards client v, the inference error $\|x-x'\|$ is determined by the above two factors.

To sum up, from Theorem 2 and Theorem 3, inference attack can be implemented to learn privacy in FL under the non-i.i.d. setting, and the attack performance is influenced by the difference of data distribution. Moreover, passive attackers (*e.g.* an honest-but-curious client k holding w^f_t and w^k_t) can steal preimage privacy with easy implementation, while active attackers (*e.g.*, a malicious client k and an external malicious attacker) can reveal both preimage privacy and membership privacy but requires a more powerful capacity to collect prior knowledge (*e.g.,* a victim's model for white-box attack). More details about the attack scenarios are addressed in the following two subsections.

B. Passive Attack

To collaboratively train a global model in FL, all clients should achieve some consensus, such as the same model structure, loss function and similar data domain, which can be used as the prior knowledge for an honest-but-curious client to perform passive inference attack. According to Theorem 2 and Theorem 3, the honest-but-curious client only needs to analyze the received global model parameter and/or his local model parameter without tampering training rules or bringing negative impact on learning accuracy.

In the example of passive attack (PA) in Fig. 1, client 1 wants to infer the features of $\{y_4, y_5, y_6\}$ of client 2. Ideally, the best way is using w^2_t to get $x=\mathcal{C}^{-1}(y,w^2_t)$. Unfortunately, it is hard or impossible for an honest-but-curious client in FL to obtain w^2_t. Instead, client 1 uses w^1_t and/or w^f_t for inference. That is, client 1 learns client 2's private information via $x'=\mathcal{C}^{-1}(y,w^1_t)$ and/or $x'=\mathcal{C}^{-1}(y,w^f_t)$. The attack performance $\|x-x'\|$ is upper bounded by either $\|w^1_t-w^2_t\|$ or $\|w^f_t-w^2_t\|$ as analyzed in Theorem 1, Theorem 2 and Theorem 3.

C. Active Attack

Besides the honest-but-curious clients in passive attack, there may be active attackers aiming at stealing privacy from benign clients of FL. An active attacker could be either an external adversary or a participant of FL. Unlike the passive attackers who only hold their own model parameters, the active attackers usually have stronger power to acquire more prior knowledge and resources (*e.g.*, hijacking transmitted parameters, eavesdropping information exchange, and compromising local clients), leading to severe privacy leakage in FL.

As shown in Fig. 1, the active attacker (AA) has the ability to access a victim client k's model parameter w^k_t and/or the aggregated model parameter w^f_t. With w^k_t and w^f_t in hand as a white box, the active attacker can launch three kinds of privacy inference attack. (i) **Instance-level membership attack** on D^k with w^k_t and $\mathcal{D}$ with w^f_t, in which the attacker can easily use the target model as a white box to learn whether a specific datapoint x is in the target model's training dataset [19]. (ii) **Model inversion attack** on client k with w^k_t and the entire system with w^f_t, which is a white box attack and is similar to the passive attack in Section IV-B. (iii) **Client-level membership attack** on a target client by consistently analyzing w^f_t to infer whether the target client joins the training process of FL or not, which causes serious consequences when the target client holds identity-related data in FL. For example, in a FL system that is trained on mobile phone trajectory datasets, the trajectory data is user-dependent and can be used to infer other private information like sex, address, and occupation, *etc.* [20], [21].

V. Dual Differentially Private Federated Learning (2DP-FL)

For passive attack, the root cause of successfully inferring any victim client j' parameter w^j_t from w^k_t and/or w^f_t by any honest-but-curious client k is that the learning model overfits the training dataset. The stronger overfitting, the more accurate inference results. Since DP can introduce randomized noise

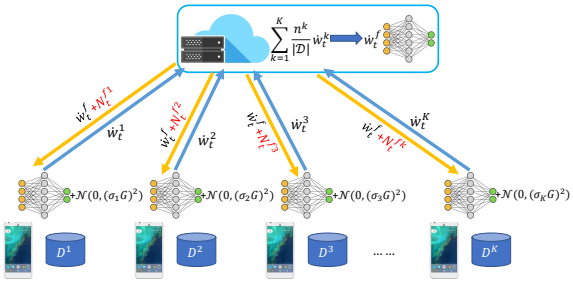


Fig. 2. The framework of our proposed 2DP-FL mechanism.

into training process, extend the generalization capability, and reduce the overfitting [22], it is an effective solution to relieve privacy inference attack. Besides, DP is applicable to defend active inference attack. As active attack is essentially a white-box attack on the victim client's parameter w_t^j and/or the federated model parameter w_t^f, injecting randomized noise into the training process can conceal private information.

In light of the above analysis, we propose a novel mechanism, named "Dual Differentially Private Federated Learning (2DP-FL)", in which DP-based noise is added when training $\dot{w}_t^k$ and before downloading $\dot{w}_t^f$ as illustrated in Fig. 2. Adding noise into $\dot{w}_t^k$ can perturb the model parameters to resist data-level membership privacy attack and model inversion attack, and adding noise into $\dot{w}_t^f$ can defend client-level membership privacy attack and model inversion attack [7], [23].

At the beginning, the initialized global model is distributed to all local clients by setting $w_t^k = \dot{w}_0^f$. Within time slot t, each client k independently trains his own local dataset D^k by minimizing the loss function L^k, in which a random batch data B is picked and the gradient $g(x)$ is calculated based on each data point $x \in D^k$. To bound the gradient contribution of each x, we clip the gradient with a predefined upper bound G, average all gradients in B, and add a scaled gaussian noise, $N_t^k \sim \mathcal{N}(0,(\sigma_k G)^2)$, to achieve DP at the client side. Then, the local model $\dot{w}_t^k$ with DP protection is updated by gradient descent method for each client k. After receiving the local model parameter $\dot{w}_{t+1}^k$ from the selected clients, the server performs FedAvg algorithm to get $\dot{w}_{t+1}^f$. Then, the federated model $\dot{w}_{t+1}^f$ is updated and distributed to all local clients, in which a noise $N_t^{fk} \sim \mathcal{N}(0, \frac{(\sigma S)^2}{U})$ is added into $\dot{w}_t^f$ for each client k. With the perturbed model parameter $\dot{w}_{t+1}^f$, each client k can update $\dot{w}_{t+2}^k$ in time slot $(t+1)$. The pseudo-code of the operations at the clients and the server is described in Algorithm 1.

In the original FL system, $\|w_{mT}^j - w_{mT}^k\|$ is gradually reduced with the increase of T because both w_{mT}^j and w_{mT}^j are trained based on the commonly shared federated model parameter w_{mT}^f, leading to an improved performance of inferring client j's privacy at the side of client k. However, when the server distributed noise N_t^{fk} is equal for all clients, which is similar as most existing works [6], [7], the privacy still can not be protected. On the contrary, in our 2DP-FL system, every client k receives a different perturbed model $(\dot{w}_t^f + N_t^{fk})$ from the server, which is helpful to relieve the reduction of $\|w_{mT}^j - w_{mT}^k\|$ and thus enhance the difficulty of

Algorithm 1: Twice Noised Differentially Private Federated Learning (2DP-FL)

Input: Total iteration T for FL, number of clients K, selected client U, initialized model $\dot{w}_{t=0}^f$,
Output: 2DP-FL model $\dot{w}_T^f$

t=0;
for $k \in \{1,2,\dots K\}$ **do**
  $w_t^k = \dot{w}_0^f$;
end
while $t < T$ **do**
  for $k \in \{1,2,\dots K\}$ **do**
    take a random batch B from D^k with probability $p = \frac{|B|}{|D^k|}$;
    compute gradient for each $x \in B$, $g(x) \leftarrow \nabla_w L^k(w_t^k, x)$;
    clip gradient, $\bar{g}(x) \leftarrow g(x)/\max(1, \frac{\|g(x)\|_2}{G})$;
    add noise, $\tilde{g} \leftarrow \frac{1}{|B|}(\sum_{x\in B} \bar{g}(x) + \mathcal{N}(0,(\sigma_k G)^2)$;
    update local model, $\dot{w}_{t+1}^k \leftarrow w_t^k - \eta\tilde{g}$;
  end
  $\dot{w}_{t+1}^f \leftarrow \sum_{k=1}^K \frac{|D^k|}{|\mathcal{D}|}\dot{w}_{t+1}^k$;
  for $k \in \{1,2,\dots K\}$ **do**
    $w_{t+1}^k \leftarrow \dot{w}_{t+1}^f + \mathcal{N}(0, \frac{(\sigma S)^2}{U})$;
  end
end

privacy inference. Moreover, the noise addition at the server in 2DP-FL is flexible and can be pre-determined according to the different application requirements. For examples, when $N_t^{fj} = N_t^{fk} \neq 0$ for $j,k \in [1,K]$, a same DP noise is added into the federated model for distribution, which is a common method used in the current works; when $N_t^{fk} = 0$ for $k \in [1,K]$, the DP noise is only added into the clients' local model parameters, and the corresponding federated model is denoted by $\tilde{w}_t^f$ for presentation in the following analysis.

For privacy-preserving FL, the model accuracy is another important concern as adding too much noise into a model would inevitably reduce learning performance. The elegant design of 2DP-FL lies in the flexible setting of N_t^{fk} that can meet various privacy protection needs with negligible impact on model accuracy, which is analyzed in Theorem 4.

Theorem 4: In our 2DP-FL mechanism, the difference between $\dot{w}_t^f$ and $\tilde{w}_t^f$ is negligible, *i.e.*, $\|\tilde{w}_t^f - \dot{w}_t^f\| = 0$.

Proof: From the setting of N_t^k at each client k and N_t^{fk} at the server, we have the following equation:

$$
\begin{aligned}
\|\tilde{w}_t^f - \dot{w}_t^f\| &= \sum_{k=1}^{K} \frac{n^k}{|\mathcal{D}|}(\dot{w}_{t-1}^f - \eta[\nabla_w L^k(\dot{w}_{t-1}^f) + N_t^k]) \\
&\quad - \sum_{k=1}^{K} \frac{n^k}{|\mathcal{D}|}(\dot{w}_{t-1}^f + N_t^{fk} - \eta[\nabla_w L^k(\dot{w}_{t-1}^f + N_t^{fk}) + N_t^k]) \\
&= \mathbb{E}\{\|(\dot{w}_{t-1}^f - \eta[\nabla_w L^k(\dot{w}_{t-1}^f) + N_t^k]) \\
&\quad - (\dot{w}_{t-1}^f + N_t^{fk} - \eta[\nabla_w L^k(\dot{w}_{t-1}^f + N_t^{fk}) + N_t^k])\|\} \\
&= \mathbb{E}\{\|N_t^{fk} + \eta[\nabla_w L^k(\dot{w}_{t-1}^f + N_t^{fk}) - \nabla_w L^k(\dot{w}_{t-1}^f)]\} \\
&\overset{1}{\leq} 0 + \eta\mathbb{E}\{\|\mu \cdot N_t^{fk}\|\} \\
&= 0
\end{aligned}
$$

The inequality 1 holds because N_t^{fk} and N_t^k are both normal distribution noise with mean value 0. ■

Theorem 4 tells that with various noise N_t^{fk}, the performance of model $\dot{w}_t^f$ is still nearly the same as that of model $\tilde{w}_t^f$.

Before analyzing the convergence of 2DP-FL, an important conclusion is introduced in Theorem 5.

Theorem 5: For the federated model $\tilde{w}^f$, $\mathbb{E}\{\|L(\tilde{w}_{t+1}^f) - L(\tilde{w}_t^f)\|\}$ is upper bounded by the following inequality:

$$\begin{aligned}&\mathbb{E}\{\|L(\tilde{w}_{t+1}^f) - L(\tilde{w}_t^f)\|\} \le \gamma_1 \mathbb{E}\{\|\nabla L(\tilde{w}_t^f)\|^2\} \\ &+ \gamma_2 \mathbb{E}\{\|N_{t+1}\|\}\|\nabla L(\tilde{w}_t^f)\| + \gamma_3 \mathbb{E}\{\|N_{t+1}\|^2\}.\end{aligned} \tag{18}$$

Proof: First, the federated model $\tilde{w}_t^f$ at t-th step can be represented as $\tilde{w}_t^f = \sum_1^K P^k(w_t^k + N_t^k)$, where $P^k = \frac{n^k}{|\mathcal{D}|}$ is the weight of client k, w_t^k is client k's local model parameter without noise, and N_t^k is the injected noise in client k at time t-th step. According to the property of Lipschitz continuous gradient, we have

$$L^k(\tilde{w}_{t+1}^f) \le L^k(\tilde{w}_t^f) + \nabla L^k(\tilde{w}_t^f)(\tilde{w}_{t+1}^f - \tilde{w}_t^f) + \frac{\mu}{2}\|\tilde{w}_{t+1}^f - \tilde{w}_t^f\|^2.$$

By taking the expectation of both sides, Eq. (19) is obtained.

$$\begin{aligned}&\mathbb{E}\{\|L(\tilde{w}_{t+1}^f) - L(\tilde{w}_t^f)\|\} \\ &\le \mathbb{E}\{\nabla L(\tilde{w}_t^f)(\tilde{w}_{t+1}^f - \tilde{w}_t^f)\} + \frac{\mu}{2}\mathbb{E}\{\|\tilde{w}_{t+1}^f - \tilde{w}_t^f\|^2\},\end{aligned} \tag{19}$$

Particularly, $\tilde{w}_{t+1}^f - \tilde{w}_t^f$ can be expressed by:

$$\begin{aligned}\tilde{w}_{t+1}^f - \tilde{w}_t^f &= \sum_1^K P^k(w_{t+1}^k + N_{t+1}^k) - \tilde{w}_t^f \\ &= \sum_1^K P^k(-\eta \nabla L^k(\tilde{w}_t^f)) + N_{t+1},\end{aligned} \tag{20}$$

where $N_{t+1} = \sum_{k=1}^K P^k N_{t+1}^k$. Similarly, for $\|\tilde{w}_{t+1}^f - \tilde{w}_t^f\|$, we have

$$\begin{aligned}\|\tilde{w}_{t+1}^f - \tilde{w}_t^f\| &= \|\sum_{k=1}^K P^k(w_{t+1}^k + N_{t+1}^k) - \tilde{w}_t^f\| \\ &= \|\sum_{k=1}^K P^k(w_{t+1}^k - \tilde{w}_t^f) + N_{t+1}\| \\ &\le \mathbb{E}\{\|w_{t+1}^k - \tilde{w}_t^f\|\} + \|N_{t+1}\| \\ &= \|\eta \nabla L(\tilde{w}_t^f)\| + \|N_{t+1}\|.\end{aligned} \tag{21}$$

Then we substitute Eq. (20) and Eq. (21) into Eq. (19) to get the following inequality:

$$\begin{aligned}&\mathbb{E}\{\|L(\tilde{w}_{t+1}^f) - L(\tilde{w}_t^f)\|\} \\ &\le \mathbb{E}\{\nabla L(\tilde{w}_t^f)(\sum_1^K P^k(-\eta \nabla L^k(\tilde{w}_t^f)) + N_{t+1})\} \\ &\quad + \frac{\mu}{2}\mathbb{E}\{[\|\eta \nabla L(\tilde{w}_t^f)\| + \|N_{t+1}\|]^2\} \\ &= (-\eta + \frac{\mu\eta^2}{2})\mathbb{E}\{\|\nabla L(\tilde{w}_t^f)\|^2\} + (1 + \mu\eta)\mathbb{E}\{\|N_{t+1}\|\}\|\nabla L(\tilde{w}_t^f)\| \\ &\quad + \frac{\mu}{2}\mathbb{E}\{\|N_{t+1}\|^2\}.\end{aligned} \tag{22}$$

For simplicity, let $\gamma_1 = -\eta + \frac{\mu\eta^2}{2}$, $\gamma_2 = 1 + \mu\eta$ and $\gamma_3 = \frac{\mu}{2}$. Thus, Eq. (22) can be equivalently rewritten as Eq. (18):

$$\begin{aligned}&\mathbb{E}\{\|L(\tilde{w}_{t+1}^f) - L(\tilde{w}_t^f)\|\} \le \gamma_1 \mathbb{E}\{\|\nabla L(\tilde{w}_t^f)\|^2\} \\ &+ \gamma_2 \mathbb{E}\{\|N_{t+1}\|\}\|\nabla L(\tilde{w}_t^f)\| + \gamma_3 \mathbb{E}\{\|N_{t+1}\|^2\}.\end{aligned}$$

■

Theorem 6: The convergence upper bound of our proposed 2DP-FL method after T iterations is given by Eq. (23) when $\eta \in (0, \frac{2}{\mu}]$, or Eq. (24) when $\eta \in (\frac{2}{\mu}, +\infty)$.

$$\mathbb{E}\{\|L(\dot{w}_T^f) - L(w^{f*})\|\} \le (1 + 2\tau\gamma_1)^T C_0 + \frac{\gamma_3 \omega^2 T \log\frac{1}{\delta}}{2\tau\gamma_1 \epsilon^2}, \tag{23}$$

$$\mathbb{E}\{\|L(\dot{w}_T^f) - L(w^{f*})\|\} \le (\frac{1}{2\tau} + \gamma_1)G^2 + \frac{\gamma_3 \omega^2 T \log\frac{1}{\delta}}{\epsilon^2}, \tag{24}$$

where $C_0 = \|L(\tilde{w}_0^f) - L(w^{f*})\|$ represents the initialization quality of federated model.

Proof: From Theorem 4, there is

$$\begin{aligned}&\mathbb{E}\{\|L(\dot{w}_T^f) - L(w^{f*})\|\} \\ &= \mathbb{E}\{\|L(\dot{w}_T^f) - L(\tilde{w}_T^f) + L(\tilde{w}_T^f) - L(w^{f*})\|\} \\ &\le \mathbb{E}\{\|L(\tilde{w}_T^f) - L(w^{f*})\|\}.\end{aligned} \tag{25}$$

With Eq. (25) and Theorem 5, we can get the following inequality:

$$\begin{aligned}&\mathbb{E}\{\|L(\tilde{w}_{t+1}^f) - L(w^{f*})\|\} \le \mathbb{E}\{\|L(\tilde{w}_t^f) - L(w^{f*})\|\} \\ &+ \gamma_1 \mathbb{E}\{\|\nabla L(\tilde{w}_t^f)\|^2\} + \gamma_2 \mathbb{E}\{\|N_{t+1}\|\}\|\nabla L(\tilde{w}_t^f)\| + \gamma_3 \mathbb{E}\{\|N_{t+1}\|^2\}.\end{aligned} \tag{26}$$

When $\eta \in (0, \frac{2}{\mu}]$, $\gamma_1 \le 0$. According to Polyak-Lojasiewicz inequality, we have

$$\gamma_1 \mathbb{E}\{\|\nabla L(w)\|^2\} \le 2\tau\gamma_1 \mathbb{E}\{\|L(w) - L(w^*)\|\}. \tag{27}$$

The result of substituting Eq. (27) into Eq. (26) is

$$\begin{aligned}&\mathbb{E}\{\|L(\tilde{w}_{t+1}^f) - L(w^{f*})\|\} \\ &\le \mathbb{E}\{\|L(\tilde{w}_t^f) - L(w^{f*})\|\} + 2\tau\gamma_1 \mathbb{E}\{\|L(\tilde{w}_t^f) - L(w^{f*})\|\} \\ &\quad + \gamma_2 \mathbb{E}\{\|N_{t+1}\|\}\|\nabla L(\tilde{w}_t^f)\| + \gamma_3 \mathbb{E}\{\|N_{t+1}\|^2\} \\ &= (1 + 2\tau\gamma_1)\mathbb{E}\{\|L(\tilde{w}_t^f) - L(w^{f*})\|\} \\ &\quad + \gamma_2 \mathbb{E}\{\|N_{t+1}\|\}\|\nabla L(\tilde{w}_t^f)\| + \gamma_3 \mathbb{E}\{\|N_{t+1}\|^2\} \\ &\le (1 + 2\tau\gamma_1)^2 \mathbb{E}\{\|L(\tilde{w}_t^f) - L(w^{f*})\|\} \\ &\quad + (1 + 2\tau\gamma_1)[\gamma_2 \mathbb{E}\{\|N_t\|\}\|\nabla L(\tilde{w}_{t-1}^f)\| + \gamma_3 \mathbb{E}\{\|N_t\|^2\}] \\ &\quad + \gamma_2 \mathbb{E}\{\|N_{t+1}\|\}\|\nabla L(\tilde{w}_t^f)\| + \gamma_3 \mathbb{E}\{\|N_{t+1}\|^2\},\end{aligned} \tag{28}$$

where $N_{t+1} = \sum_{k=1}^K P^k N_{t+1}^k$. Since the noise N_t follows normal distribution, $\mathbb{E}\{\|N_t\|\} = 0$ and $\mathbb{E}\{\|N_t\|^2\} = \sigma^2$. Therefore, we can rewrite Eq. (28) at time T to be

$$\begin{aligned}&\mathbb{E}\{\|L(\tilde{w}_T^f) - L(w^{f*})\|\} \\ &\le (1 + 2\tau\gamma_1)^T \mathbb{E}\{\|L(\tilde{w}_0^f) - L(w^{f*})\|\} + \gamma_3 \sigma^2 \sum_0^{T-1}(1 + 2\tau\gamma_1)^T \\ &\le (1 + 2\tau\gamma_1)^T C_0 + \frac{\gamma_3 \sigma^2}{2\tau\gamma_1}.\end{aligned} \tag{29}$$

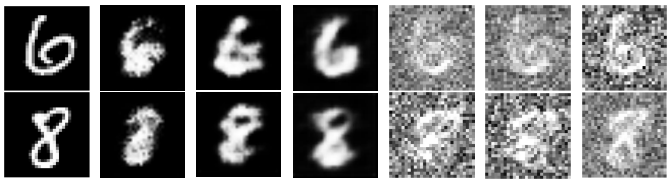
where $\sigma = \omega\sqrt{T \log\frac{1}{\delta}}/\epsilon$ is the noise scale used in moments accountant [24]. Thus, by combining Eq. (25) and Eq. (29) the convergence upper bound in Eq. (23) is proved.

When $\eta \in (\frac{2}{\mu}, +\infty)$, we obtain $\gamma_1 > 0$ and Eq. (30).

$$\gamma_1 \mathbb{E}\{\|L(w) - L(w^*)\|\} \le \frac{\gamma_1}{2\tau}\mathbb{E}\{\|\nabla L(w)\|^2\}. \tag{30}$$

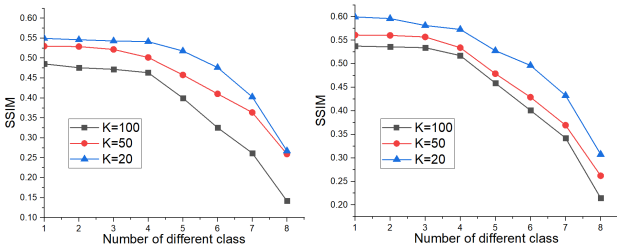
Substituting Eq. (30) and Eq. (26) into Eq. (25), we can prove the upper bound in Eq. (24). ■

Notice that in real data experiments, the learning rate η is always set to be a small scalar around 10^{-4} [24], which yields a negative γ_1 and leads to the convergence bound in Eq. (23).



(a) (b) (c) (d) (e) (f) (g)

Fig. 3. Visual results of attack on MNIST dataset under different scenarios.



(a) Passive attack of local model (b) Passive attack on global model

Fig. 4. Attack performance vs. data distribution difference.

VI. Experiments

In this section, intensive real-data experiments are carried out to validate our analysis on privacy leakage in FL and our 2DP-FL mechanism.

MNIST contains 10 classes (*i.e.*, $0-9$) for classification problem and is adopted in our experiments. Considering the FL system under non-i.i.d. setting, we distribute the dataset to different clients according to their class labels, ensuring the assigned local datasets follow different types of distribution. The whole dataset is divided into 15 non-overlapping data buckets and 5 overlapping data buckets, each of which contains data associated with a pre-determined label group. Specifically, for the non-overlapping buckets, there are 15 label groups including $\{0,1\}$, $\{2,3\}$, $\{4,5\}$, $\{6,7\}$, $\{8,9\}$, $\{0\}$, $\{1\}$, ..., $\{9\}$; and for the overlapping buckets, there are 5 label groups including $\{0,1,2,3\}$, $\{2,3,4,5\}$, $\{4,5,6,7\}$, $\{6,7,8,9\}$, and $\{8,9,0,1\}$. Each client can get one or multiple data buckets for local distribution configuration.

The experiments consist of two parts, including privacy leakage analysis and defense performance evaluation. In the analysis of privacy leakage, we show the feasibility of privacy inference attack towards the original FL system by visualizing the data recovery results. To investigate the defense performance of our 2DP-FL mechanism, the convergence, attack accuracy, and classification accuracy are evaluated. Moreover, to the best of our knowledge, the state-of-the-art scheme termed "NbAFL" [10] realizes privacy-preserving FL with the idea similar to our 2DP-FL and is taken as a baseline for performance comparison. All experiments are performed on a Linux server with Intel(R) Xeon CPU E5-1607, 16 GB memory, and the NVIDIA GeForce RTX 2080 GPU with 11 GB memory, and the common used machine learning library Pytorch, Pysyft and OpenCV are adopted.

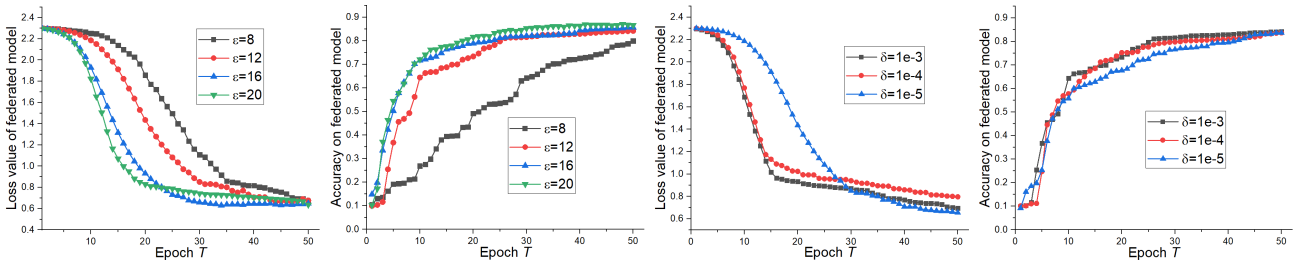
A. Analysis of Privacy Leakage in FL

From the analysis in Section IV, we know that an honest-but-curious client in FL can work as a passive attacker to infer the privacy of a victim client's data class by using his local model and/or federated model. Taking model inversion attack as a case of attack scenario, such a passive attacker aims to recover an unseen class's common features of the victim client whose dataset has a different distribution. In our experiments, to infer the features of unseen classes in the victim's dataset that holds class labels $\{5,6,7,8,9\}$, the passive attacker trains his own dataset with class labels $\{0,1,2,3,4\}$ and implements model inversion attack using his local model parameter w_t^k and/or the federated model parameter w_t^f.

The results of privacy inference attack are visualized in Fig. 3. Fig. 3(a) shows the original images with labels 6 and 8 from MNIST dataset. Fig. 3(b) and Fig. 3(c) display the results of passive inference attack with the attacker's local model parameter and the federated model parameter in the original FL, respectively, from which we can see that both attack results expose some feature information of the target classes. Compared with Fig. 3(b), the recovered images in Fig. 3(c) is closer to the original one. This is because the difference between global data distribution and victim data distribution is more similar, which is in line with our analysis in Section IV – a smaller difference of data distribution leads to a higher upper-bound of attack performance. When the victim's model is used in active inference attack (*i.e.*, the victim's model is captured and used as a white box), as shown in Fig. 3(d), the reconstructed image is more clear, demonstrating the feasibility of active attack. Fig. 3(e), Fig. 3(f) and Fig. 3(g) present the recovered images of passive attack with local model, passive attack with federated model, and active attack under our 2DP-FL mechanism, in which it is hard to perceive that the digit labels of recovered images are 6 and 8. With the protection of our 2DP-FL mechanism, the visual quality of images recovered by model inversion attack is significantly reduced. Moreover, we can see that the results of Fig. 3(e) and Fig. 3(f) are worse than the result of Fig. 3(g), because active attack uses white-box to perform inference that tends to have smaller model error.

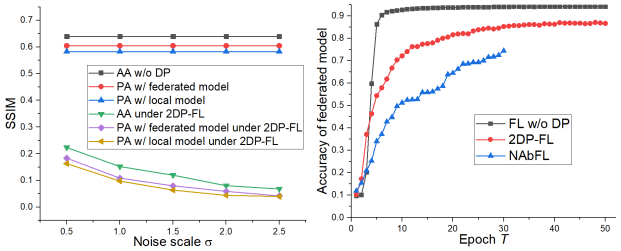
To evaluate the attack performance varying with data distribution, we plot the attack results in Fig. 4 by changing the difference of data distribution between the attacker and the victim, where x-axis represents the number of different classes between the attacker's dataset and the victim's dataset, and the y-axis denotes the similarity between the original and the recovered images. More specifically, more different classes between the attacker's dataset and the victim's dataset results in a larger difference between data distribution. The image similarity is measured by Structure Similarity Index Metric (SSIM) with a range of $[0,1]$, where 0 means totally different and 1 means exactly the same.

Fig. 4(a) depicts the attack performance when the attacker's local model parameter w_t^k is used. According to the Fig. 4(a), the value of SSIM decreases as the number of different classes is increasing, which is consistent with our theoretical analysis in Section IV. Meanwhile, Fig. 4(a) shows the impact of total number of client, K, on privacy leakage in FL. As K



(a) Loss of federated model vs. ϵ (b) Accu. of federated model vs. ϵ (c) Loss of federated model vs. δ (d) Accu. of federated model vs. δ

Fig. 5. Evaluation on convergence of 2DP-FL mechanism.



(a) Attack performance vs. σ (b) Baseline comparison

Fig. 6. Evaluation on privacy and utility of 2DP-FL mechanism.

becomes larger, the amount of each client's private information included in the federated model is reduced as the contribution of each client' local model to the federated model is reduced, mitigating privacy leakage in FL. When the global model parameter w_t^f by the attacker for privacy inference, similar observations can be obtained from Fig. 4(b).

B. 2DP-FL Evaluation

According to the analysis in Section V, our proposed 2DP-FL mechanism can defend inference attack while benefiting a good data utility. We design different experiments in this subsection to deeply investigate the performance of 2DP-FL mechanism from the aspects of learning convergence, privacy protection, and data utility.

First, to evaluate the convergence of our 2DP-FL mechanism, we make the following settings for Fig. 5: the number of clients is K=50, the number of selected clients is U=10, and the number of training epochs is T=50. Additionally, ϵ and δ are the privacy parameters of (ϵ, δ)-DP, where larger ϵ and δ mean less privacy protection.

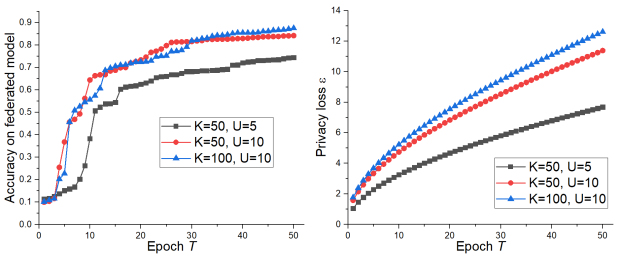
As shown in Fig. 5(a), with the increase of T (*i.e.*, the number of training epoch), the decrease of loss value becomes smaller and smaller, gradually reaching a stable loss value, which reflects the convergence trend of 2DP-FL scheme from experimental perspective. Thus we can conclude that our 2DP-FL mechanism converges when T is large enough, which is consistent with the convergence analysis in Theorem 6. On the other hand, a larger ϵ (*i.e.*, a higher privacy budget) leads to a faster convergence, because higher privacy budget implies weaker privacy protection with less injected noise during training process. Fig. 5(b) also confirms the convergence from the viewpoint of model accuracy, in which the accuracy increases with epoch T and stays stable after a certain threshold. In addition, the impact of δ on convergence is evaluated in Fig. 5(c) and Fig. 5(d). Similar to ϵ, larger δ results in faster convergence, less privacy protection, and higher accuracy. For example, the loss value of $\delta = 10^{-3}$ shows the fastest decrease and is the first to convergence in Fig. 5(c) as well as reaches the highest accuracy in Fig. 5(d), because it sacrifices more privacy for maintaining utility. From the results of Fig. 5, it also demonstrates that the maintenance of data utility is achieved at the price of privacy protection.

Then, we investigate attack performance in terms of SSIM through changing the noise scale σ of DP. The adopted FL setting is: K=50, U=10, T=50, and δ=0.001. In Fig. 6(a), AA represents active attack, and PA represents passive attack. When our 2DP-FL is implemented for privacy protection, the attack performance is drastically reduced. The noise scale σ along x-axis denotes the privacy protection level. With the increase of σ, the attack performance is gradually reducing, which shows that our privacy protection works as expected.

Furthermore, we compare the classification accuracy of the original FL without DP, the baseline NbAFL mechanism, and our 2DP-FL mechanism and show the experimental results in Fig. 6(b). Especially, we fix $\epsilon = 20$ for NbAFL and 2DP-FL when DP is taken into account. Obviously, the accuracy of the original FL without DP is the best as no noise is added for privacy protection. The accuracy of our 2DP-FL is better than that of NbAFL. In particular, our 2DP-FL becomes convergent within the given ϵ range and achieves an accuracy of 85%. However, for NbAFL, when $T = 31$, the exhausted value of ϵ exceeds the given budget 20 and only gets an accuracy of 74%. From the comparison, we obtain two critical findings: (i) when the number of epochs is the same, 2DP-FL costs a smaller ϵ for better privacy protection; and (ii) when privacy budget is fixed, 2DP-FL can run more epochs for better accuracy. Thus, we can conclude that our 2DP-FL mechanism outperforms NbAFL in terms of classification accuracy and privacy protection.

C. Impact of K and U

In a FL system, besides privacy related parameters, hyper-parameters such as the number of client K and the number of selected client U also play important roles in fine-tuning the systems. Under our 2DP-FL mechanism, we evaluate the influence of K and U for further investigating hyper-parameter



(a) Impact of K and U on accuracy (b) Impact of K and U on privacy

Fig. 7. Impact of K and U.

strategies in the FL system. In Fig. 7(a), when the number of client is fixed at K=50, the federated model with a larger value of U reaches a faster convergence and a higher accuracy as a higher participant radio (represented by U/K) is helpful to enhance the training performance of FL systems; while for the same U (*i.e.,* U=10), a larger of K implies more clients' local datasets are available to the FL system, improving training performance.

The overall privacy loss of the FL system, indicated by ϵ, is exhibited in Fig. 7(b). When K is equal, a smaller value of U can help reduce privacy loss, because with a smaller value of U, a larger noise is added to the clients' datasets (see line 15 of Algorithm 1). If U remains the same, a larger value of K causes an increased privacy loss. As more clients' local datasets are available for federated learning, it is more possible to reveal privacy of the FL system via data correlation, increasing the risk of privacy loss. In a summary, the learning accuracy and privacy loss of our privacy-preserving FL mechanism 2DP-FL are dependent on the values of K and U, which can be exploited to balance the trade-off between learning accuracy and privacy loss in real applications.

VII. Conclusion

For the first time in literature, this paper rigorously analyzes the issue of privacy leakage and proves the performance upper bound of privacy inference attack in FL with non-i.i.d. data. This analysis motivates us to develop a novel mechanism, 2DP-FL, for preserving private information with ensuring differential privacy. Besides, the noise addition in 2DP-FL can be flexibly set according to different application requirements, and the upper-bounded convergence of 2DP-FL can guarantee its learning performance. Through extensive experiments, the results of our theoretical analysis and the effectiveness of our 2DP-FL mechanism can be confirmed.

Acknowledgment

This work was partly supported by the National Science Foundation of U.S. (1741277, 1829674, 1704287, 1912753, and 2011845) and the Microsoft Investigator Fellowship.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[2] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE Symposium on Security and Privacy*. IEEE, 2019, pp. 739–753.

[3] M. Hao, H. Li, X. Luo, G. Xu, H. Yang, and S. Liu, "Efficient and privacy-enhanced federated learning for industrial artificial intelligence," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 10, pp. 6532–6542, 2019.

[4] L. Zhu and S. Han, "Deep leakage from gradients," in *Federated Learning*. Springer, 2020, pp. 17–31.

[5] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 1175–1191.

[6] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.

[7] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," *arXiv preprint arXiv:1710.06963*, 2017.

[8] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Differentially private asynchronous federated learning for mobile edge computing in urban informatics," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 2134–2143, 2019.

[9] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan, "cpsgd: Communication-efficient and differentially-private distributed sgd," in *Advances in Neural Information Processing Systems*. NIPS, 2018, pp. 7564–7575.

[10] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, pp. 3454–3469, 2020.

[11] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE Symposium on Security and Privacy*. IEEE, 2019, pp. 691–706.

[12] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 603–618.

[13] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "Dba: Distributed backdoor attacks against federated learning," in *International Conference on Learning Representations*. ICLR, 2019.

[14] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2938–2948.

[15] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *International Conference on Machine Learning*. PMLR, 2019, pp. 634–643.

[16] R. Hu, Y. Guo, E. P. Ratazzi, and Y. Gong, "Differentially private federated learning for resource-constrained internet of things," *arXiv preprint arXiv:2003.12705*, 2020.

[17] A. Sonee and S. Rini, "Efficient federated learning over multiple access channel with differential privacy constraints," *arXiv preprint arXiv:2005.07776*, 2020.

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

[19] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy*. IEEE, 2017, pp. 3–18.

[20] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 766–775, 2018.

[21] Q. Han, Z. Xiong, and K. Zhang, "Research on trajectory data releasing method via differential privacy based on spatial partition," *Security and Communication Networks*, vol. 2018, 2018.

[22] B. Wu, S. Zhao, C. Chen, H. Xu, L. Wang, X. Zhang, G. Sun, and J. Zhou, "Generalization in generative adversarial networks: A novel perspective from privacy protection," in *Advances in Neural Information Processing Systems*. NIPS, 2019, pp. 307–317.

[23] J. Zhao, Y. Chen, and W. Zhang, "Differential privacy preservation in deep learning: Challenges, opportunities and solutions," *IEEE Access*, vol. 7, pp. 48 901–48 911, 2019.

[24] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in

Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, 2016, pp. 308–318.

Zuobin Xiong received the B.S. degree from the Department of Mathematics, Northeast Forestry University, Harbin, Heilongjiang, China in 2016, and the M.S. degree from College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang, China in 2019. Currently, he is pursuing the Ph.D. degree at the Department of Computer Science, Georgia State University, Atlanta, GA, USA. His research interests lie in a broad area of cybersecurity and privacy, including Privacy-preserving Machine Learning, Privacy-preserving Data Mining, Internet of Things (IoT), Differential Privacy and other privacy and/or security issues of related topics.

Zhipeng Cai (SM'06) is currently an Associate Professor at Department of Computer Science, Georgia State University, USA. He received his PhD and M.S. degrees in the Department of Computing Science at University of Alberta, and B.S. degree from Beijing Institute of Technology. Prior to joining GSU, Dr. Cai was a research faculty in the School of Electrical and Computer Engineering at Georgia Institute of Technology. Dr. Cai's research areas focus on Internet of Things, Machine Learning, Cyber-Security, Privacy, Networking and Big data. Dr. Cai is the recipient of an NSF CAREER Award. He served as a Steering Committee Co-Chair and a Steering Committee Member for WASA and IPCCC. Dr. Cai also served as a Technical Program Committee Member for more than 20 conferences, including INFOCOM, ICDE, ICDCS. Dr. Cai has been serving as an Associate Editor-in-Chief for Elsevier High-Confidence Computing Journal (HCC), and an Associate Editor for more than 10 international journals, including IEEE Internet of Things Journal (IoT-J), IEEE Transactions on Knowledge and Data Engineering (TKDE), IEEE Transactions on Vehicular Technology (TVT).

Daniel Takabi (M'10) received the B.S. degree in Computer Engineering from Amirkabir University of Technology in 2004, M.S. degree in Information Technology, Sharif University of Technology in 2007, and PhD degree in Information Science and Technology from University of Pittsburgh in 2013. He is currently an Associate Professor of computer science and the Next Generation Scholar with Georgia State University (GSU), Atlanta, GA, USA. He is also a Founding Director of the Information Security and Privacy: Interdisciplinary Research and Education (INSPIRE) Center which is designated as the National Center of Academic Excellence in Cyber Defense Research (CAE-R). His research interests include various aspects of cybersecurity and privacy, including privacy preserving machine learning, adversarial machine learning, advanced access control models, insider threats, and usable security and privacy. He is a member of IEEE and ACM.

Wei Li (M'16) is currently an Assistant Professor in the Department of Computer Science at Georgia State University. Dr. Li received her Ph.D. degree in computer science, from The George Washington University, in 2016 and M.S. degree in Computer Science from Beijing University of Posts and Telecommunications, in 2011. She won the Best Paper Awards in ACM MobiCom Workshop CRAB 2013 and international conference WASA 2011, respectively. Her current research spans the areas of blockchain technology, security and privacy for the Internet of Things and Cyber-Physical Systems, secure and privacy-aware computing, Big Data, game theory, and algorithm design and analysis. She is a member of IEEE and a member of ACM.