2023 IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC) | 978-1-6654-9626-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/SPAWC53906.2023.10304534

Attri-Fed: A GIB Framework for Attribute-Based Privacy and Communication-Efficient Federated Learning

Ahmet Faruk Saz
Georgia Institute of Technology
Atlanta, USA
asaz3@gatech.edu

Yashas Malur Saidutta
Georgia Institute of Technology
Atlanta, USA
ysaidutta3@gatech.edu

Faramarz Fekri
Georgia Institute of Technology
Atlanta, USA
faramarz.fekri@ece.gatech.edu

Mustafa Riza Akdeniz
Intel Corporation
Santa Clara, USA
mustafa.akdeniz@intel.com

Brandon Edwards
Intel Corporation
Santa Clara, USA
brandon.edwards@intel.com

Nageen Himayat
Intel Corporation
Santa Clara, USA
nageen.himayat@intel.com

***Abstract*—We present an attribute-based privacy framework for Federated Learning. Our framework utilizes the Generalized Information Bottleneck (GIB) principle to create functionally compressed representations that obscure designated sensitive attributes from potential inferential adversaries at the server. These functional representations are generated following a minimax adversarial optimization of the privacy and utility bounds on the optimization function. We show that our proposed framework inherently provides attribute-based differential privacy (DP) guarantees, reduces communication overhead and improves utility (e.g., image classification performance). We presented both theoretical and experimental comparisons of our less restrictive attribute-based DP approach with conventional DP.**

***Index Terms*—information bottleneck, differential privacy, federated learning, minimax optimization, variational bounds**

I. Introduction

In recent years, the advancement of complex machine learning (ML) models has significantly increased the demand for data. Alongside this, the proliferation of interconnected devices has resulted in a massive accumulation of user-generated data. However, due to privacy concerns, this data is not immediately available for use. The data access problems are further amplified by the distributed nature of the data and communication, computation and storage burden it would impose if they were to be transmitted to a central location. As a solution, 'Federated Learning' (FL) was proposed, shifting model training to the edge and addressing these issues simultaneously [1]. Shortly after the introduction of FL, secure aggregation framework [2] was developed to ensure cryptographic security whereas 'Differential Privacy' (DP) [3] is selected as semantic privacy mechanism. Cryptographic techniques, however, impose significant communication and computation burden, and in the case of DP, injected noise decreases the utility.

In this work, we claim that attempting to protect all user data with its entire set of input features can be needlessly strict in certain cases, with disadvantages outweighing the benefits from perspectives of utility, communication and computation. In other words, we suggest that not all features of a user dataset may be privacy sensitive, and it may be sufficient to protect only a subset of sensitive features. In return, improved utility and reduced communication can be achieved in comparison to that of secure aggregation and DP combination. Hence, we propose a privacy framework for FL that offers protection only for a subset of input features that are deemed private.

Our contributions are as follows:

- We propose a novel 3-stage training algorithm based on Generalized Information Bottleneck (GIB) principle that assures attribute-based privacy in the FL.
- We develop novel variational bounds, namely enforcement and encouragement, for GIB-based objective.
- Privatization occurs entirely locally, incurring no additional communication.
- Our proposed system can be a considered as a lighter alternative to differential privacy, in the sense that a version that satisfies DP only for certain subset of input attributes.
- We establish theoretical connections between our framework of privacy, attribute-based DP, and regular DP. Further, we compare our method with DP experimentally.
- We experimentally demonstrate that our system is successful against curious inferential adversaries in FL setting even under the scarcity of the data.

The following section motivates our system through poten-

This work is supported by National Science Foundation under award ID MLWiNS-2003002 and a Gift from Intel Co.

tial use cases.

II. Motivation

Property inference attacks are performed to infer a property of the input dataset that does not necessarily related to the target task. The core principle behind these attacks lies in exploiting the vast function approximation capacity of deep models where model captures unintended patterns as well as intended ones. The literature includes various works demonstrating the potency of such attacks. Examples include estimating whether a person is wearing glasses from a gender classifier, determining a doctor's specialty and identity from healthcare review sentiment analysis, and inferring the authorship of comments from healthcare review sentiment analysis [4]. It is worth noting that simply discarding 'sensitive' input features may not suffice, as correlations between sensitive features and others can still lead to information leakage. This issue could pose problems for legal documents that are redacted before release or for facial recognition-based payment systems, as they may unintentionally disclose sentiment/demographic information related to shopping preferences. Furthermore, in image-based applications, private features are often entangled with non-private ones, making it impossible to discard the private features. In such scenarios, differential privacy becomes an option. However, because DP does not differentiate between features, it can disproportionately impair utility. Consequently, in such cases, we believe that our system can serve as a lightweight alternative to differential privacy. In the next section, we compare our work with existing studies.

III. Related Works

[5] and [6] use adversarial training for information obfuscation, but only within a centralized setting. Conversely, [7] and [8] consider distributed setups, with [8] focusing specifically on the context of fairness. Notably, none of these prior studies consider the possibility of data scarcity or provide comparisons to the de-facto privacy standard, Differential Privacy (DP). This limits their practical applications.

IV. Differential Privacy

Due to space constraints, we include a very brief treatment of regular notion of DP in this manuscript. Interested reader can refer to [3]. Differential privacy is defined as:

Definition 1 ((ϵ, δ)-DP): A randomized mechanism $\mathcal{K}$ with domain $\mathcal{Z}$ and $\mathcal{A} \subseteq Range(\mathcal{K})$ is (ϵ, δ) differentially private if for any two neighboring datasets $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$:

$$\frac{P(\mathcal{K}(\mathbf{z}) \in \mathcal{A})}{P(\mathcal{K}(\mathbf{z}') \in \mathcal{A})} \leq e^{\epsilon} + \delta \tag{1}$$

Definition 1 implies that outcomes of the randomized mechanism $\mathcal{K}$ have bounded dissimilarity across neighboring datasets. When $\delta = 0$, the randomized mechanism is said to be ϵ-DP, which, obviously, provides stricter privacy guarantees. It is straightforward to extend this definition of differential privacy to attribute-based differential privacy (ADP).

Definition 2 ((ϵ, δ)-ADP): Given input dataset $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$ where $\mathbf{x}_i \in \mathcal{X}^p$ for $\forall i \in \{1, ..., M\}$, define a new dataset $\mathbf{u} = \{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M\}$ where each $\mathbf{u}_i$ is the vector set of sensitive attributes corresponding to dataset entry $\mathbf{x}_i$, i.e., $\mathbf{u}_i \in \mathcal{U}^t$ for $\forall i \in \{1, ..., M\}$. Then, given two neighboring datasets **u** and **u'**, (ϵ, δ)-Attribute-based DP is defined as:

$$\frac{P(\mathcal{K}(\mathbf{x}|\mathbf{u}) \in \mathcal{A})}{P(\mathcal{K}(\mathbf{x'}|\mathbf{u'}) \in \mathcal{A})} \leq e^{\epsilon} + \delta \tag{2}$$

Intuitively, an (ϵ, δ)-ADP mechanism $\mathcal{K}$ assures that the generated outputs will not permit differentiation between different sets of sensitive attributes, thereby effectively obscuring them.

V. Theory

In this section, theoretical backbone of the proposed privatization system is introduced.

A. GIB Principle

The Generalized Information Bottleneck (GIB) principle is the fusion of two existing frameworks in the literature [9], [10], namely Information Bottleneck and Privacy Funnel. GIB framework provides a disciplined way of finding a compressed representation $\hat{Y}$ of an input observation X that is maximally informative about a target attribute V -subject to a rate constraint- and minimally informative about private attributes U. Graphical representation of the GIB-based data generation process is given in Fig. 1.

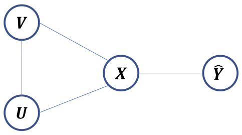

Fig. 1: Graphical model representing the dependencies among variables.

B. Optimization Objective

Moving from the GIB framework, let's first consider the problem of finding 'privatized' representations. Using Fig. 1, the optimization objective can be formulated as:

$$min - I(V; \hat{\mathbf{Y}}) + \beta I(\mathbf{U}; \hat{\mathbf{Y}}) \tag{3}$$

where $\hat{\mathbf{Y}} \in \mathcal{Y}^d$, $\mathbf{U} \in \mathcal{U}^t$, and β determines the trade-off between privacy and utility. Intuitively, the first term encourages $\hat{\mathbf{Y}}$ to preserve information about V while the second term supports the removal of information about **U**. Given that mutual information is hard to compute and minimize for high dimensional variables, we employ variational approximations to transform the objective in eqn. (3) into a form that can be optimized in practice.

C. Variational Bounds on Optimization Objective

1) Utility Upper Bound on 1st Term of Eqn. (3):

$$
\begin{aligned}
T_1 &= -I(V; \hat{\mathbf{Y}}) && (4)\\
&= -H(V) + H(V|\hat{\mathbf{Y}}) && (5)\\
&= -H(V) - E_{\hat{\mathbf{Y}}}[KL(p_{V|\hat{\mathbf{Y}}}(v|\hat{\mathbf{y}})||q_{V|\hat{\mathbf{Y}}}(v|\hat{\mathbf{y}}))]\\
&\quad - E_{V,\hat{\mathbf{Y}}}[log(q_{V|\hat{\mathbf{Y}}}(v|\hat{\mathbf{y}}))] && (6)\\
&\leq -H(V) - E_{V,\hat{\mathbf{Y}}}[log(q_{V|\hat{\mathbf{Y}}}(v|\hat{\mathbf{y}}))] && (7)
\end{aligned}
$$

where $q(v|\hat{\mathbf{y}})$ is the variational approximation to $p(v|\hat{\mathbf{y}})$. The upper bound in eqn. (7) follows from the non-negativity of KL Divergence and its minimization serves to maximize the utility. This upper bound is realized via a local encoder neural network (NN) (with params. θ_ϵ) that generates privatized representations $\hat{Y}$ from input and a FL model NN (with params. θ_ζ) that is responsible for prediction of digit class ($\hat{V}$) from $\hat{Y}$.

2) Encouragement Upper Bound on 2nd Term of Eqn. (3):

$$
\begin{aligned}
T_2 &= I(\mathbf{U}; \hat{\mathbf{Y}}) && (8)\\
&= I(\hat{\mathbf{Y}}; \mathbf{X}) - I(\hat{\mathbf{Y}}; \mathbf{X}|\mathbf{U}) + I(\hat{\mathbf{Y}}; \mathbf{U}|\mathbf{X}) && (9)\\
&= I(\hat{\mathbf{Y}}; \mathbf{X}) - I(\hat{\mathbf{Y}}; \mathbf{X}|\mathbf{U}) && (10)\\
&= I(\hat{\mathbf{Y}}; \mathbf{X}) - E_{\mathbf{U},\hat{\mathbf{Y}}}[KL(p_{\mathbf{X}|\mathbf{U},\hat{\mathbf{Y}}}(\mathbf{x}|\mathbf{u},\hat{\mathbf{y}})||q_{\mathbf{X}|\mathbf{U},\hat{\mathbf{Y}}}(\mathbf{x}|\mathbf{u},\hat{\mathbf{y}}))]\\
&\quad - E_{\mathbf{X},\mathbf{U},\hat{\mathbf{Y}}}[log(q_{\mathbf{X}|\mathbf{U},\hat{\mathbf{Y}}}(\mathbf{x}|\mathbf{u},\hat{\mathbf{y}}))] + H(\hat{\mathbf{Y}}|\mathbf{X}) && (11)\\
&\leq I(\hat{\mathbf{Y}}; \mathbf{X}) - E_{\mathbf{X},\mathbf{U},\hat{\mathbf{Y}}}[log(q_{\mathbf{X}|\mathbf{U},\hat{\mathbf{Y}}}(\mathbf{x}|\mathbf{u},\hat{\mathbf{y}}))] + C_{\hat{\mathbf{Y}}|\mathbf{X}} && (12)\\
&= H(\hat{\mathbf{Y}}|\mathbf{X}) - KL(p_{\hat{\mathbf{y}}}(\hat{\mathbf{y}}))||q_{\hat{\mathbf{y}}}(\hat{\mathbf{y}}))) - E_{\hat{\mathbf{Y}}}[log(q_{\hat{\mathbf{y}}}(\hat{\mathbf{y}}))]\\
&\quad - E_{\mathbf{X},\mathbf{U},\hat{\mathbf{Y}}}[log(q_{\mathbf{X}|\mathbf{U},\hat{Y}}(\mathbf{x}|\mathbf{u},\hat{\mathbf{y}}))] + C_{\hat{\mathbf{Y}}|\mathbf{X}} && (13)\\
&\leq E_{\hat{\mathbf{Y}}}[log(q_{\hat{\mathbf{y}}}(\hat{\mathbf{y}}))] - E_{\mathbf{X},\mathbf{U},\hat{\mathbf{Y}}}[log(q_{\mathbf{X}|\mathbf{U},\hat{Y}}(\mathbf{x}|\mathbf{u},\hat{\mathbf{y}}))] + 2C_{\hat{\mathbf{Y}}|\mathbf{X}} && (14)\\
&\leq E_{\hat{\mathbf{Y}}}[log(q_{\hat{\mathbf{y}}}(\hat{\mathbf{y}}))] - E_{\mathbf{X},\mathbf{U},\hat{\mathbf{Y}}}[log(q_{\mathbf{X}|\mathbf{U},\hat{Y}}(\mathbf{x}|\mathbf{u},\hat{\mathbf{y}}))] && (15)
\end{aligned}
$$

where transition from equation (10) to (11) follows from the chain rule of mutual information and independence of $\hat{\mathbf{Y}}$ and $\mathbf{U}$ conditioned on $\mathbf{X}$. Upper bounds on eqn. (12) and (14) follow from non-negativity of KL Divergence. $C_{\hat{\mathbf{Y}}|\mathbf{X}} = H(\hat{\mathbf{Y}}|\mathbf{X})$ is constant since $\hat{\mathbf{Y}}$ is a deterministic function of $\mathbf{X}$. $q_{\mathbf{X}|\mathbf{U},\hat{\mathbf{Y}}}(\mathbf{x}|\mathbf{u},\hat{\mathbf{y}})$ is the variational approximation to $p_{\hat{\mathbf{Y}};\mathbf{X}|\mathbf{U}}(\hat{\mathbf{y}};\mathbf{x}|\mathbf{u}) = p_{\hat{\mathbf{Y}};V|\mathbf{U}}(\hat{\mathbf{Y}}; f(\mathbf{X}) = V|\mathbf{U})$ and $q_{\hat{\mathbf{y}}}(\hat{\mathbf{y}})$ is the variational approximation to $p_{\hat{\mathbf{Y}};\mathbf{X}}(\hat{\mathbf{y}};\mathbf{x})$. The first term in equation (13) can be interpreted as a rate constraint and hence is upper bounded with a non-parametric rate estimator. This estimator is realized with a encouragement helper NN (with params. θ_η) that takes $\hat{Y}$ as input and outputs an estimate of rate $E_{\hat{\mathbf{Y}}}[log(q_{\hat{\mathbf{y}}})]$. Intuitively, the second term encourages the system *encouraged* to contain information about target attribute V orthogonal to private attribute $\mathbf{U}$. This bound is implemented with a encouragement helper NN (with params. θ_η) that takes $\hat{Y}$ and U as its input and outputs $\hat{V}$. However, as experimentally demonstrated, when the dimension of $\hat{\mathbf{Y}}$ is large, this bound may not by itself suffice and $\hat{\mathbf{Y}}$ can still leak information about private attribute. Henceforth, we need a term that explicitly enforces the removal of information, bringing us to the lower bound presented in the next section.

3) Enforcement Lower Bound on 2nd Term:

$$
\begin{aligned}
T_2 &= I(\mathbf{U}; \hat{\mathbf{Y}}) && (16)\\
&= H(\mathbf{U}) - H(\mathbf{U}|\hat{\mathbf{Y}}) && (17)\\
&= H(\mathbf{U}) + E_{\hat{\mathbf{Y}}}[KL(p_{\mathbf{U}|\hat{\mathbf{Y}}}(\mathbf{u}|\hat{\mathbf{y}})||q_{\mathbf{U}|\hat{\mathbf{Y}}}(\mathbf{u}|\hat{\mathbf{y}}))]\\
&\quad + E_{\mathbf{U},\hat{Y}}[log(q_{\mathbf{U}|\hat{Y}}(\mathbf{u}|\hat{\mathbf{y}}))] && (18)\\
&\geq H(\mathbf{U}) + E_{\mathbf{U},\hat{Y}}[log(q_{\mathbf{U}|\hat{Y}}(\mathbf{u}|\hat{\mathbf{y}}))] && (19)
\end{aligned}
$$

where $q(u|\hat{\mathbf{y}})$ is the variational approximation to $p_{\mathbf{U}|\hat{\mathbf{Y}}}(\mathbf{u}|\hat{\mathbf{y}})$ and inequality follows from non-negativity of KL Divergence. This bound explicitly *enforces* removal of private information from $\hat{Y}$. One concern is that, however, minimizing lower bounds may not be as effective as upper bounds if they are loose. To address the issue, we make use of an *adversarial* training framework to first tighten the lower bound, and then perform minimization. The minimax training can be accomplished with the help of a local privatizer NN (θ_ψ) that takes $\hat{Y}$ as its input and outputs U. Putting all the bounds together, the optimization problem reduces to:

$$
\begin{aligned}
&\min_{\theta_\epsilon,\theta_\zeta,\theta_\eta,\theta_\phi} \max_{\theta_\psi} \; E_{V,\hat{\mathbf{Y}}}[D_v(v; g_d(\hat{\mathbf{y}};\theta_\zeta))] + (\lambda+\beta)E_{\hat{\mathbf{Y}}}[log(q_{\hat{\mathbf{y}}}(\hat{\mathbf{y}};\theta_\phi))]\\
&+ \beta E_{V,\mathbf{U},\hat{\mathbf{Y}}}[D_v(v; (g_d(\hat{\mathbf{y}};\theta_\zeta),\mathbf{u}))] - \gamma E_{\mathbf{U},\hat{\mathbf{Y}}}[D_u(\mathbf{u}; g_u(\hat{\mathbf{y}};\theta_\psi))] \qquad (20)
\end{aligned}
$$

where D_v is a distortion metric for V, and D_u is a distortion metric for U. The minimax objective in eqn. (20) can be optimized via neural networks as presented in the next section.

D. System Architecture and Implementation

In the following subsections, we elaborate on the implementation details of the proposed 3-stage training process.

1) Stage 1: Initialization: This pre-training phase begins with the joint training of the encoder and the local Federated Learning (FL) model for the target task, while the local privatizer remains frozen. Then, local privatizer is trained for private attribute prediction while local encoder and the FL model are frozen. This stage, in turn, allows local nodes to have networks that are capable of predicting meaningful results such that they can better optimize minimax objective. Algorithm 1 details the process.

2) Stage 2: Privatization: After pretraining, block gradient descent-based adversarial minimax optimization starts. Maximization phase trains local privatizer for private attribute prediction whereas competing minimization phase trains local encoder, with the help of encouragement helper, rate estimator and adversary, to generate representations that confuse the adversary but still (along with FL model) useful for target task. Steps of this stage is provided in Algorithm 2.

3) Stage 3: Federated Learning: In this final stage, local nodes possess privatized compressed inputs, which are utilized for Federated Learning. According to the data processing inequality, private representations retain their privacy guarantees, irrespective of the FL model architecture.

Algorithm 1 *Initialization.* The set of clients is denoted by $\mathcal{C}$; E_j are respective number of training epochs; D_j are appropriate distortion metrics.

```
Input: $\bigcup_{i \in \mathcal{C}} \mathcal{D}_{private,i}$
Initialize: $g_e(\mathbf{x}; \theta_{\epsilon_0})), g_d(\hat{\mathbf{y}}; \theta_{\zeta_0})), g_h((\mathbf{u}, \hat{\mathbf{y}}), g_u(\hat{\mathbf{y}}; \theta_{\psi_0}))$
for $i \in \mathcal{C}$ do
  repeat
    $\mathcal{L}_{g_{e,d}}(t) = D_v(v; g_d(g_e(\mathbf{x}; \theta_{\epsilon_i}(t)); \theta_{\zeta_i}(t)))$
    $\mathcal{L}_{g_u}(t) = D_u(u; D_u((\mathbf{u}, g_u(g_e(\mathbf{x}; \theta_{\epsilon_i})); \theta_{\psi_i}(t))))$
    $\mathcal{L}_{g_h}(t) = D_v((v; g_h((\mathbf{u}, g_e(\mathbf{x}; \theta_{\epsilon_i}(t)); \theta_{\eta_i}(t)))), \theta_{\eta_i}(t)))$
    $\theta_{\epsilon_i}(t+1) = \theta_{\epsilon_i}(t) - \nabla \mathcal{L}_{g_{e,d}}(t)$
    $\theta_{\zeta_i}(t+1) = \theta_{\zeta_i}(t) - \nabla \mathcal{L}_{g_{e,d}}(t)$
    $\theta_{\psi_i}(t+1) = \theta_{\psi_i i}(t) - \nabla \mathcal{L}_{g_u}(t)$
    $\theta_{\psi_i}(t+1) = \theta_{\psi_i i}(t) - \nabla \mathcal{L}_{g_h}(t)$
  until convergence
end for
```

Algorithm 2 *Local Privatization.* The set of clients is denoted by $\mathcal{C}$; E_j are respective number of training epochs; D_j are appropriate distortion metrics; λ, β and γ are hyperparameters.

```
Input: $\bigcup_{i \in \mathcal{C}} \mathcal{D}_{private,i}$
repeat
  for $i \in \mathcal{C}$ do
    while $t \in 0,\ \ldots,\ E_{minimization}$ do
      $\mathcal{L}_{g_{e,d,h,u}}(t) = D_v(v; g_d(\hat{\mathbf{y}}; \theta_{\zeta_i}(t))) + (\lambda + \beta) log(q_{\hat{\mathbf{y}}}(\hat{\mathbf{y}}; \theta_{\phi_i}(t))) + \beta D_v(v; (g_d(\hat{\mathbf{y}}; \theta_{\zeta_i}(t)), \mathbf{u})) - \gamma D_u(\mathbf{u}; g_u(\hat{\mathbf{y}}; \theta_{\psi_i}(t)))$
      $\theta_{\epsilon_i}(t+1) = \theta_{\epsilon_i}(t) - \nabla \mathcal{L}_{g_{e,d,h,u}}(t)$ for $\Delta = \epsilon_i, \zeta_i, \phi_i, \psi_i$
      while $t \in 0,\ \ldots,\ E_{maximization}$ do
        $\mathcal{L}_{g_u}(t)\ \ D_u((\mathbf{u}, g_u(\hat{\mathbf{y}}; \theta_{\psi_i}(t))))$
        $\theta_{\psi_i}(t+1) = \theta_{\psi_i}(t) + \nabla \mathcal{L}_{g_{e,d,h,u}}(t)$
      end while
    end while
  end for
until convergence
```

E. Connection to Differential Privacy

Differential privacy provides its privacy guarantee through a randomization mechanism such as the addition of Gaussian or Laplacian noise. In situations where data is distributed and a trusted aggregator does not exist, local differential privacy becomes a viable option. However, it can severely compromise utility [11]. In light of this, we propose a privacy measure that doesn't overly sacrifice utility, unlike differential privacy. It is important to note that our work doesn't intend to replace differential privacy; rather, our goal is to provide a lightweight alternative for use cases where the privacy guarantees of differential privacy may be overly stringent. We can establish the theoretical relationship between our mechanism and differential privacy through statistical privacy schema presented in [12]. For that matter, let's first define another privacy metric, namely information privacy:

Definition 3 [12]: (ϵ-Information Privacy): A privacy preserving mapping $p_{\hat{\mathbf{Y}}|\mathbf{U}}(.|.)$ provides ϵ-information privacy if, for all sensitive attributes $\mathbf{u} \subseteq \mathcal{U}$:

$$exp(-\epsilon) \leq \frac{p_{\mathbf{U}|\hat{\mathbf{Y}}}(\mathbf{u}|\hat{\mathbf{y}})}{p_{\mathbf{U}}(\mathbf{u})} \leq exp(\epsilon) \tag{21}$$

Then, connection between information privacy and differential privacy is given by a straightforward extension of Theorem 3 in [12] such that ϵ-information privacy implies (2ϵ, 0)-differential privacy. In its attack, the adversary tries to determine a revised distribution $q \in \mathcal{P}_{\mathcal{U}}$ where $\mathcal{P}_{\mathcal{U}}$ is the all possible probability distributions over $\mathcal{U}$ such that a cost function, i.e. $\mathcal{C}(\mathbf{U}, q) = -loq(q(\mathbf{U}))$ (self-information loss) is minimized. Then, the maximum gain an adversary can achieve is $\Delta C^* = c_0^* - \min_{\hat{\mathbf{y}} \in \hat{\mathbf{Y}}}[c_{\hat{\mathbf{y}}}^*]$ where $c_0^* = \min_{q \in \mathcal{P}_{\mathcal{U}}} E_{\mathbf{U}}[-loq(q(\mathbf{U}))]$ is the initial guess of the adversary, i.e. prior distribution $p(\mathbf{u})$, and $c_{\hat{\mathbf{y}}}^* = H(\mathbf{U}|\hat{\mathbf{Y}} = \hat{\mathbf{y}})$ is what adversary can infer after observing a particular sample $\hat{\mathbf{y}}$ during the training process. It is possible to determine an upper bound on this maximum information leakage *per any sample* with the help of the following theorem and lemma:

Theorem 2: Given a randomization mechanism κ with ϵ-IP guarantee over the distribution of private attributes $\mathcal{P}_{\mathcal{U}}$, the upper bound on the maximum information leakage of any sample is given as (proof of which is an extension of Corollary 2 in [12]):

$$\Delta C^* \leq (1 - e^{\epsilon}) H(U) + \frac{e^{\epsilon}\epsilon}{ln2} \tag{22}$$

Lemma 1: Assume that $\forall \mathbf{u}_i \in \mathcal{U}^t$, $T \leq p(\mathbf{u}_i)$, where $T \in (0, 1]$ is a positively lower bounded distribution and U is a discrete random variable. Then, $\frac{p(\mathbf{u}|\mathbf{y'})}{p(\mathbf{u})} \leq \frac{p(\mathbf{u}|\mathbf{y'})}{T} \leq \frac{1}{T}$.

The importance of Lemma 1 lies in the observation that it offers a practical way of computing the ϵ term in information privacy, as the prior $p(\mathbf{u})$ is already assumed to be available to adversary. Hence:

Corollary 1: Given the positively lower bounded prior distribution $p(\mathbf{U})$ of discrete random variable $\mathbf{U}$, a privatization

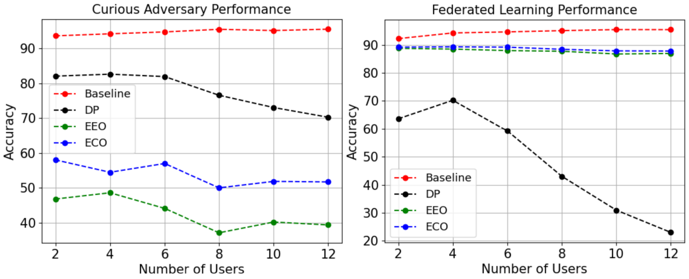

Fig. 2: Global Adv. Acc. (Left), FL Performance (Right)

scheme $p_{\mathbf{U},\hat{\mathbf{Y}}}(\mathbf{u}|\hat{\mathbf{y}})$ is ϵ-IP for $\epsilon = ln\big(\frac{1}{T}\big)$.

Then, one can use Corollary 1 to determine an attribute-based DP guarantee. In the next section, we validate our theoretical findings through experimental results.

VI. Simulation Results and Conclusion

In our experiments, we adopted the procedure outlined in [13] to create a colored MNIST dataset (with 10 colors non-uniformly and randomly distributed among different digit classes) where digit color is treated as private attribute and prediction of the digit is target task. The training and test sets are evenly split into 100 shards, which was done deliberately to restrict the number of samples available to each participant. Experiments are run for $N = 2, 4, 6, 8, 10, 12$ participants. Local encoder, FL model, local privatizer, encouragement helper, and rate estimator are all selected to be feedforward neural networks. Specifically, the local encoder comprises two hidden layers of 1024 neurons each with ReLU activation, while the FL model features a single hidden layer of 32 neurons with ReLU activation and an intermediate 50-neuron bottleneck layer. Moreover, two separate feedforward neural networks with 1024 and 32 neurons in the hidden layers with ReLU activation were employed as the encouragement helper and local privatizer. As for the rate estimator, we used the tensorflow-compression toolbox. We then subjected our system to testing against a category of non-linear curious inferential adversaries situated at the server. To ensure be able to compare our notion of (ϵ, 0)-ADP to regular (ϵ, δ)-DP, we use the Proposition 3 of [14] with a large α=100 value and implement (ϵ, δ)-DP via Opacus AI toolbox.

We simulate three setups where in the baseline, only employ local encoder and FL model are employed for digit classification. Then, privacy protecting systems, where one of them using only encouragement helper and other one using both encouragement helper and local privatizer are simulated where the former only uses encouragement bound (ECO) whereas the latter uses both enforcement and encouragement bounds (EEO). Finally, (ϵ, δ)-DP is simulated and results are presented in fig. (2). The tests aimed to assess the performance of curious adversaries' predictions as well as the performance of FL.

As demonstrated in Table 1, the global adversary achieved the highest success rate in the absence of privacy protection. Interestingly, (ϵ, δ)-DP is not very effective in preventing private attribute leak as it attempts to protect entire input rather than a particular property. EO protection managed to eliminate information to a certain extent, while EEO delivered the most robust protection. Furthermore, both EO and EEO privacy mechanisms successfully remove private information even under limited data availability with relatively low effect on FL performance. The effect of (ϵ, δ) local DP becomes detrimental to FL performance with increasing number of participants. Hence, we have shown that our proposed 3-stage privacy protection system based on GIB principle is successful at locally removing private attribute information in FL setting with limited number of data points and relatively low degredation to utility compared to DP.

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*, 2017.

[2] 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*, ser. CCS '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1175–1191. [Online]. Available: https://doi.org/10.1145/3133956.3133982

[3] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

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

[5] M. Bertran, N. Martinez, A. Papadaki, Q. Qiu, M. Rodrigues, G. Reeves, and G. Sapiro, "Adversarially learned representations for information obfuscation and inference," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 614–623. [Online]. Available: https://proceedings.mlr.press/v97/bertran19a.html

[6] H. Zhao, J. Chi, Y. Tian, and G. J. Gordon, "Trade-offs and guarantees of adversarial representation learning for information obfuscation," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9485–9496. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/6b8b8e3bd6ad94b985c1b1f1b7a94cb2-Paper.pdf

[7] Z. Alsulaimawi and H. Liu, "Distributed variational information bottleneck for iot environments," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, 2021, pp. 1–6.

[8] P. P. Liang, T. Liu, L. Ziyin, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," *ArXiv*, vol. abs/2001.01523, 2020.

[9] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *ArXiv*, vol. physics/0004057, 2000.

[10] S. Salamatian, F. du Pin Calmon, N. Fawaz, A. Makhdoumi, and M. Médard, "Privacy-utility tradeoff and privacy funnel," 2020.

[11] E. Lomurno and M. Matteucci, "On the utility and protection of optimization with differential privacy and classic regularization techniques," *ArXiv*, vol. abs/2209.03175, 2022.

[12] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2012, pp. 1401–1408.

[13] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *ArXiv*, vol. abs/1907.02893, 2019.

[14] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 2017, pp. 263–275.