Anonymous Submission

Subject Membership Inference Attacks in Federated Learning

Abstract: Privacy in Federated Learning (FL) is studied at two different granularities - item-level, which protects individual data points, and user-level, which protects each user (participant) in the federation. Nearly all of the private FL literature is dedicated to the study of privacy attacks and defenses alike at these two granularities. More recently, subject-level privacy has emerged as an alternative privacy granularity to protect the privacy of individuals whose data is spread across multiple (organizational) users in cross-silo FL settings. However, the research community lacks a good understanding of the practicality of this threat, as well as various factors that may influence subject-level privacy. A systematic study of these patterns requires complete control over the federation, which is impossible with real-world datasets. We design a simulator for generating various synthetic federation configurations, enabling us to study how properties of the data, model design and training, and the federation itself impact subject privacy risk. We propose three inference attacks for subject-level privacy and examine the interplay between all factors within a federation. Our takeaways generalize to real-world datasets like FEMNIST, giving credence to our findings.

Keywords: Subject level Privacy, Federated Learning, Distribution Inference, Subject Membership

AS: Add a sentence to the caption of each relevant figure/table to summarize the findings and any take-home messages.

1 Introduction

Federated Learning (FL) [1] allows multiple parties to collaboratively train a Machine Learning model while keeping the training data decentralized. FL was originally introduced for mobile devices, with a core motivation of protecting data privacy. In the *cross-device* setting, privacy is usually defined at two *granularities*: first, *item-level privacy*, which describes the protection of individual data items [2, 3] and *user-level privacy*, which describes the protection of the entire data distribution of the device user [3, 4].

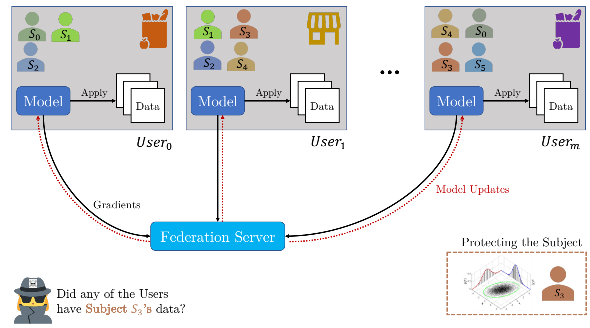


Fig. 1. Subject-level Privacy in cross-silo Federated Learning. Data subjects can appear in multiple federation users, setting the problem apart from User-level or Item-level Privacy.

Federated Learning is now also employed in collaborations between large organizations or data centers across geographies, the so called *cross-silo* setting [5]. The 'users' of the federation in this setting are the organizations, such as a group of retailers or hospitals, who in turn might have collected data about individuals. These individuals are often referred to as *data subjects* [6]. Interestingly, data about one subject might be spread across multiple users of the federation. For example, a consumer shopping at multiple retailers or a patient going to multiple hospitals. Both item-level and user-level privacy definitions are insufficient to address the need to protect an individual's data in such a setting. A third privacy granularity, called *subject-level* privacy was recently introduced [7], which precisely describes the protection of the data distribution of a data subject in cross-silo FL settings (Figure 1).

Note that subject-level privacy may or may not be distinct from item-level or user-level privacy, depending on how the data is setup. For example, datasets in which one row of data corresponds to one person, item-level privacy is sufficient to protect the individual's identity, thus enforcing subject-level privacy. Similarly, in cross-devise FL setting, the distinction between user-level and subject-level privacy is somewhat blurred, because there is roughly a one-to-one correspondence between a data subject and a device, which acts as a user in the federation: each device typically holds the data from just one individual, and each individual's data is typically stored in just one (or few) devices. However, in the cross-silo setting, in which users are large organizations collecting data from a large number of individuals, and a data subject can easily be associated with a number of different users in the federation, this distinction becomes much more significant. Subject-level privacy formulation is important because ultimately, we are interested in preserving the privacy of an individual, not just that of a data item or a data silo.

Even though FL offers a certain degree of privacy by restraining each user's training data at its local site, the model trained using this data is prone to a variety of inference attacks [8, 9] that aim to reveal some part of the private information in the original training data. The literature on inference attacks is extensive [10–16] and continues to grow at a rapid pace. However, the research has largely focused on attacks such as membership inference [10, 11], model inversion [12, 13], and property inference [14–16]. To the best of our knowledge, no existing work in the FL setting has explored inference attacks that leak privacy of subjects whose data is scattered across multiple users in the federation. In this paper, we introduce *subject membership inference attacks*, which aim to infer the presence of an individual's data spanning across multiple federation users. By measuring effectiveness of such an attack, one can assess the vulnerability of the FL model and estimate the risk of privacy leakage for data subjects.

We assume an *honest-but-curious* threat model in which either one of the federation users or the federation server could be adversarial (Section 4). We introduce three different subject membership privacy attacks. Consistent with existing literature on membership inference attack, our attacks assume that a model tends to perform better on data similar to its training data compared to the data it has not seen during training. However, unlike typical membership inference attacks, which check for membership of specific data points, we *sample* from target subject's data distribution to ascertain the subject's presence in the training data. In this sense, we can also view these attacks as *distribution inference attacks* [16]. The first of these attacks checks the loss values on data points sampled from target subjects' data distribution against a threshold (Section 5.1). The second attack tracks the changes to the loss across training rounds (Section 5.3).

Success of privacy attacks on ML models depend on both the nature of the training data as well as the type of modeling technique. A FL system with multiple users and data subjects can be quite complex and effectiveness of privacy attacks can greatly be influenced by a variety of factors. For a systematic and thorough evaluation of the proposed attacks, we first build a synthetic data simulator, capable of simulating different federation configurations (Section 6.1). Each configuration consists of multiple user datasets, which in turn are composed of data from multiple data subjects.

We focus on studying effects of the structure of the federation, the data distribution, and the model architecture on the attack accuracy. We generate several hundred configurations, changing variables such as number of users, number of data subjects, number of data items per user, data dimensionality and data generating distributions. In addition, we also experiment with different ML model complexities and training rounds. After setting up each of these federations, we train FL models using FedAvg [17] algorithm. Next, we carry out our subject-level attacks and measure their effectiveness against known ground truth about subject membership (Section 6.2).

We find that the proposed attacks are surprisingly effective in inferring subject membership in a large fraction of the configurations. This grid-based experimental protocol also helps us uncover some important trends, which can be used to provide practical guidelines to ML practitioners about the vulnerability of their FL setup or model architectures (Section 6.3). We also test the effectiveness of these attacks on the FEMNIST dataset [18] (Section 6.4) and show attack accuracy of 62%, similar to membership inference attack accuracy on FEMNIST by prior work []AS: Need to add this citation- I think Daniel mentioned he knew which one it was?.

Finally, we assess effectiveness of a popular mitigation strategy prescribed for ML privacy – Differential Privacy (DP). We retrain one of the most vulnerable configurations of our synthetic dataset using DP at the granularities of data items, federation users, and data subjects respectively [2, 7, 19], and repeat the attacks (Section 6.5). We empirically demonstrate that DP is indeed effective in reducing privacy attack efficacy. We report results for a similar experiment on the FEMNIST dataset that also demonstrate that DP based mitigation strategies make FL models robust from subject membership privacy attacks.

2 Background

Our work is concerned with what adversaries can learn about training data given access to machine learning models, which may constitute a breach of privacy. There are many attack surfaces for a machine learning model across its life-cycle, including after deployment [20–22].

We divide relevant works into three broad categories- studies on membership inference and its utility in scenarios like ours (Section 2.1), privacy in the context of federated learning (Section 2.2), and other relevant attacks and notions of privacy that can be relevant to subject level membership inference (Section 2.3).

2.1 Membership Inference

Privacy attacks are a common approach used to assess privacy risks in machine learning. The key benefit of this approach is that it grounds the privacy discussion concretely in terms of the training data whose privacy can be compromised by the model. Membership inference [10, 11] is a popular type of privacy attack that is highly relevant to our work: The membership inference task is to determine whether a particular data item was part of the training dataset. A successful membership inference attack concretely demonstrates privacy risks to individual data items from the attacked model's training dataset.

We particularly follow the line of prior work on membership inference attacks [23–25], which deliver an empirical lower bound to the risk of data leakage through a machine learning model. The adversary's accuracy in determining whether a data point is part of the training set gives a very real picture of whether the data was leaked.

Several advancements in this line of work improve this lower bound by making the attack models more accurate, and applicable in more realistic scenarios [26]. In particular, white-box attacks which rely on the gradients sent during training show that these gradients reveal a lot of detail about the training data [27–29], although the attacks may be defended against with appropriate strategies [30]. White-box attacks are quite plausible if the adversary is posing as a legitimate user in federated learning, and this opens up new avenues of risk [31–33] that are highly relevant to our work.

While these advanced attacks are worth investigating to improve subject-level membership inference, they are overwhelmingly focused on membership inference of particular data points. User-level privacy leakage [34] is more closely tied to subject-level privacy leakage that we study here, and indeed is equivalent under a one-to-one correspondence between subjects and federation users.

One of the main contributions of our paper is that we empirically test the success of membership inference attacks as properties of the data distribution change. This is related to prior work that examines patterns of membership inference success as the architecture of the model changes [35] across an even wider array of model architectures, but which focuses on item-level membership inference and a smaller number of datasets.

2.2 Privacy in Federated Learning

On a meta level, federated learning operates on data just as any other machine learning algorithm: extracting and learning features from observations that can be helpful in predictions on unseen data. However, the changes in the training environment as well as distribution of training data across clients can significantly influence properties of the learned model(s). Factors like the number of federation users and number of training rounds are known to directly affect convergence performance and privacy protection [36].

The majority of work on data privacy in machine learning focuses on item-level privacy [2, 5]: measuring and protecting the privacy of individual training examples. However, in federated learning, each user of the system sends back parameter updates corresponding to batches of examples. Even if no single data point is leaked in this process, the evolution of the FL model gives information about the batches of training data - since a user has multiple data points, their privacy may be compromised beyond what the item-level privacy guarantee would suggest. Measuring and bounding the privacy loss to users leads naturally to user-level privacy [3]. However, in this work, we focus on the even more general subject-level privacy [7]: there may be multiple data points about a particular individual in the dataset, but there is not a 1-to-1 mapping between individuals and federation users. This situation occurs commonly in real-world datasets, because a federation user may have data about multiple individuals in its dataset, or the same individual may have records scattered across several federation users.

To illustrate the differences using a real-world analogy, consider a dataset of grocery store market baskets, collected over time, and with each basket having a corresponding member ID. If each grocery location aggregates its purchases to train a model, we will certainly find the majority of individuals shop multiple times over the year, and that any individual may sometimes shop at different stores. Item-level privacy tries to protect information about individual market baskets, so that no single checkout can be identified definitively. User-level privacy will guarantee the privacy of individual stores, ensuring that no single neighborhood can be identified within the dataset. Subject-level privacy will make sure that no individual's data is compromised, despite making multiple purchases across multiple stores. Figure 1 depicts how subject-level privacy manifests itself in cross-silo FL.

2.3 Other Attacks

This paper is focused on evaluating the efficacy of subject membership inference attacks in FL, but there are taxonomies of other attacks against ML models [21] that are worth describing in order to put our work in context.

Attacks may be grouped by whether or not the adversary has knowledge (or perhaps partial knowledge) of the model, into black-box, white-box, and grey-box attacks [9, 35]. During the training of a FL model, the model structure and parameters are shared across all participants, so if the adversary is part of the federation they are free to launch white-box attacks. In general, though, we do not assume the attacker is part of the federation, and our attacks use only knowledge of the data points, the labels assigned by the model, and the loss function the model is optimizing; this is essentially a black-box attack.

From the data perspective, inferring properties about a particular subject in FL can be reduced to inferring the presence of the subject's sub-distribution within the FL's full data distribution. This corresponds to the task of distribution inference [16] (also known as property inference). The general case consists of an adversary that wishes to distinguish between two possible distributions from which a given model's training data was sampled. The go-to approach uses meta-classifiers, with the Permutation Invariant Network [15] for most neural network architectures. Although there have been some attempts to extend and evaluate these attacks in multi-party settings [37], it is unclear if they can extend to more than two clients, especially in a volatile FL environment. In our attack model, the adversary has access to the data distribution for the subject of interest, and is simply trying to ascertain whether or not the subject was part of the training set.

3 Subject Membership Inference

AS: Too short a section- perhaps we can consider merging it with some other Section, or shift/add content to this one?

Let $\mathcal{S}_0$ and $\mathcal{S}_1$ be two sets of subjects, and s_{interest} the subject whose membership the adversary wants to infer, such that $s_{\text{interest}} \notin \mathcal{S}_0, \mathcal{S}_1$. Let $\mathcal{D}_s$ be the distribution corresponding to a subject s. Then, using the definitions of distribution inference in [16] we can formulate our subject membership inference task as differentiating between models trained on datasets sampled from either of the distributions $\mathcal{D}_0$ and $\mathcal{D}_1$, defined as:

$$\mathcal{D}_0 \text{ s.t. } D \sim \mathcal{D}_0 \iff D = \bigcup_{s \in \mathcal{S}_0} d \sim \mathcal{D}_s \qquad (1)$$

$$\mathcal{D}_1 \text{ s.t. } D \sim \mathcal{D}_1 \iff D = \bigcup_{s \in \hat{\mathcal{S}}_1} d \sim \mathcal{D}_s \qquad (2)$$

where $\hat{\mathcal{S}_1} = \mathcal{S}_1 \cup \{s_{\text{interest}}\}$. This is equivalent to stating that a data sample from either of $\mathcal{D}_{\{0.1\}}$ is equivalent to taking a union of samples from the individual subjects' distributions. The first distribution $\mathcal{D}_0$ corresponds to the absence of subject of interest in the federation, while $\mathcal{D}_1$ includes it.

Note that subject membership inference is orthogonal to the FL setting, and is indeed more broadly applicable to machine learning models beyond FL. For subject membership inference in FL, it is important to note that it does not matter how a subject's data is divided across different users of the federation. Even if only one user has the subject's data, or if an individual subject's data is divided across all users, the subject's data is ultimately used in the overall training process and thus the subject should be inferred as being present. The adversary only cares about the subject's presence in the overall federation and using a formulation like the one above is apt for the given threat model. This, of course, is barring scenarios where a subject's data never gets sampled during any federated training rounds by any of the participating users. In such a case the subject's data has technically not been used in the training, and thus should not be inferred as being present.

4 Attacks Setting

4.1 Threat Model

We assume a passive adversary that wants to infer membership of a particular subject in the federation. This attacker can exist as a honest-but-curious federation server or user in the federation. In either case, by design the attacker has access to the global model's weights after each training round. For all of our attacks, we assume the adversary has access to the following:

Samples (finite) from the distribution of subjects. If the adversary wishes to launch an attack against a particular subject, it must have the capability to quantify and differentiate subjects and identify the one it is interested in. This can be done by either knowing (or estimating) a subject's distribution or possessing samples to estimate it. Having access to finite samples is the weaker assumption of these two. Note that in theory, it is not necessary to have estimates for distributions for *all* of the subjects – just the subject of interest and some samples from subjects that are not relevant to the adversary's inference task.

API access to the global model after each federation round. We assume access to prediction probabilities from the global model M_i after each training round i. Both the central server and individual participants have access to the global model after each training round, making it easy to satisfy this requirement. This may be further weakened to limit access to just the last round-the final global model that may be released to the world. For two out of our three attack methods (Section 5), we assume API access to only the final global model.

In one way or another, all of our attacks are based on a common underlying assumption: Given the objective of training machine learning models, it is natural to expect that the model's performance on data similar to that seen during training would be better than that not seen during training. This can be quantified in many ways: from raw loss values, to robustness in predictions. The flow of information for the proposed subject membership inference attack is described in Figure 2. As described above, the adversary has a belief over subjects present and absent in a federation, and wishes to infer the membership of some subject s_{interest} using some algorithm $\mathcal{H}$, given access to models M_i per round i, for all r training rounds.

4.2 Attack Setup

For any attack, it is common for the adversary to have some form of hyper-parameters that are specific to a target subject (Since we are stating this as something commonly known, don't we need a reference for it?). These hyper-parameters are usually computed using additional information that may be available through side-channel attacks or just by the adversary participating in the federation's training process.

All of our attacks involve computing some form of tunable thresholds that are used to execute the attacks. This threshold computation be done using this additional information; however, this forms a cyclic problem since inferring a subject's membership is itself the adversary's objective. One of the primary motives of this research is to study the impact of different configuration parameters on this inference risk. Thus, we choose an extremely strong adversary that already knows which subjects did and did not participate in the federation (**Scenario A**). Our results then help us study this empirical upper bound on how much the adversary could learn from the given model(s), even if it somehow computed its threshold(s) using actual ground truth. We consider two possible scenarios with respect to the knowledge possessed by the adversary before it launches its attack(s)

Scenario A: Membership known for many subjects. For the scenario where the adversary is the federation server (which is what we assume for the rest of the paper, unless specified otherwise), subjects used in the federation are already known. If the adversary is not the federation server, it can make an educated guess for a prior of the target subject's participation in the federation, and then launch its attack to then confirm the guess.

Scenario B: Membership known for few subjects. For the scenario where the adversary is a participant in the federation, it can use its split of data to know for sure which subjects were used in the federation. (AS: Just one sentence- any more information we want to/can add to make this more clear?)

The adversary can then guess subjects not used in the federation by randomly sampling/generating other subjects and their data. Once data for both subjects used and (probably) not used during training is available, the adversary can launch an appropriate attack to generate some score/loss values and find appropriate thresholds on top of these scores in order to maximize its metric of choice to identify subjects used in training effectively. Then, using the derived threshold and data sampled from a query subject's distribution, the adversary can predict or confirm whether a subject's data was used in the federation or not.

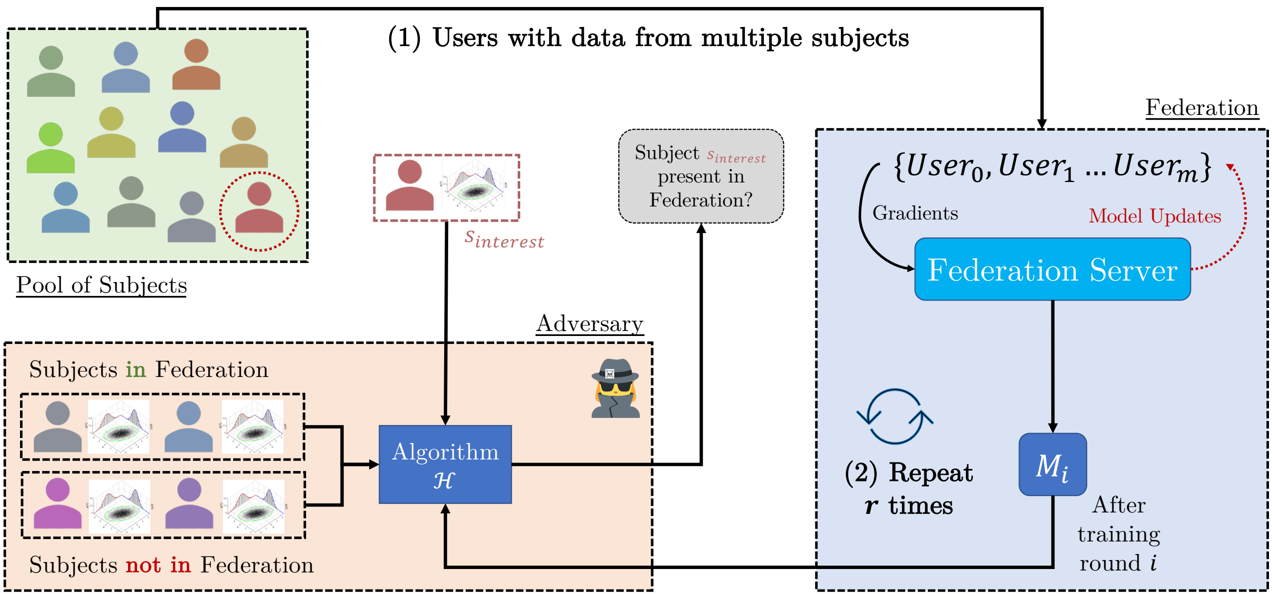


Fig. 2. Information flow for the subject membership inference attack in Federated Learning.

5 Method

Preliminaries. Let m be the total number of users participating in the federation. Let r be the number of rounds for which the global model is trained in the federation, with M_i denote the state of the model after training round i has completed. M_0 thus represents the state of the model before training starts. Let $l_i(x, y)$ be the loss value between the label y and $M_i(x)$, with $M_i(x)$ denoting the model M_i's prediction on point x.

All three of our attacks are based on hypotheses, implied by prior works on the behavior of loss functions on training data [11, 26, 26, 38].

5.1 Loss-Threshold Attack

Hypothesis 1. *If data from a particular subject is present in the federation and is used in training, the global model would be expected to have a lower loss than data from a subject who was not present in any of the users' local datasets.* [11]

Based on this hypothesis, we propose the following attack: record loss values for samples from the subject's distribution and check if any of them have a value less than a particular threshold. If the loss is below the threshold VJM: how do we find this threshold in the algorithm? I see the description of threshold calculation appear later in section 7. Perhaps another way, it would indicate the model has seen that particular data, and thus other data from that subject's distribution, during training.

$$c = \sum_{(d_x,d_y)\sim \mathcal{D}_s} \mathbb{I}[l_r(d_x, d_y) \leq \lambda] \quad (3)$$

The adversary can either check if c is non-zero or derive an additional threshold on this value based on the metric it wishes to maximize, like precision or recall.

5.2 Loss-Across-Rounds Attack

Hypothesis 2. *Loss on training data, and thus data from the training distribution, decreases (and eventually converges, based on the level of overfitting) across iterations by virtue of how learning algorithms (gradient descent, in particular) work. However, data from distributions not seen in the training would probably not exhibit the same trends. It may decrease initially owing to some similarities in the underlying distribution but would likely not decrease consistently or converge to values as low as those for distributions of subjects whose data was present in the federation [39].*

Based on this hypothesis, we propose the following attack: record loss values for samples from the subject's distribution and take note of how the loss values change as training rounds progress. The attack first computes the loss across each training round i:

$$c_i = \sum_{(d_x,d_y)\sim\mathcal{D}_s} l_i(d_x, d_y) \quad (4)$$

Then, the adversary takes note of the number of training rounds where the loss decreases after each round:

$$c = \sum_{i=1}^{r} \mathbb{I}[c_i < c_{i-1}] \quad (5)$$

The adversary can then compute these values for both subjects seen and not seen in the federation VJM: (this appears to imply that the adversary always knows which subjects are in the federation's dataset, can we make this more general to apply the attacks to subjects that are not know to be in the federation's dataset, but can be found to be a part of it by the adversary using the attacks – that will be a purely subject membership inference), and consequently derive a threshold on this value for subject membership VJM: Even if we base the thresholds on subjects known to be a part of the federation's training data, we could still use them to discover membership of subjects that were previously not known to be a part of the training data, right? If so, this point needs to come across from early on. We perhaps need to crisply state what we mean by subject membership inference. It perhaps includes (i) discovery of a subject in the model's training data, and (ii) discovery of additional data on a subject already known to be a part of the training data..

5.3 Neighborhood-Loss Attack

Hypothesis 3. *If the model sees data from a particular subject's distribution, it would be expected to generalize well to data from that distribution. One would expect the model to be robust to small amounts of noise added to data, as opposed to data from distributions of subjects that it has not seen during training.* [26]

Based on this hypothesis, we sample multiple points within L_p-norm balls around each of the datapoints in a sample, and note the fluctuation in loss values:

$$c = \sqrt{\sum_{(d_x,d_y)\sim\mathcal{D}_s} (l_r(d_x, d_y) - l_r(d_x + \epsilon, d_y))^2} \quad (6)$$

Then, similar to the attack described in Section 5.1, we can derive a threshold on c and apply it across all samples, counting how many of them fall under the given threshold. Note that the core idea for sampling points from a neighborhood is similar to the Merlin attack [26]. However, instead of simply counting instances where the loss increases, we track the actual difference in loss values which is additional information that the adversary can utilize.

6 Experiments

As mentioned in Section 4, the subject membership inference threshold (λ) is a tunable parameter. For all of the above attacks, the adversary splits its data into two parts. The first split is used for deriving the thresholds(s) λ, while the second split is used for actual evaluation and reporting results.

Our evaluation largely focuses on a synthetic dataset since it is non-trivial to obtain datasets with a clear notion of "subjects", and even more difficult to control the federation and data attributes that can significantly influence subject membership inference risks. Even though we use the setting of the strongest adversary for all of our configurations, we also test a weaker and more realistic variant for some of the configurations: one where the adversary (federation user) has knowledge of only the subjects in its own data split, and randomly samples subjects not in its set (to be considered as unseen subjects) for computing threshold(s) VJM: The second half of this sentence suggests that unseen subjects' data is also used to compute the thresholds. I think thats not true.. Our results across a variety of different configurations (spanning different levels of inference risk) indicate that weakening an adversary and using user-wise data to compute threshold(s) does not perform much worse than the stronger adversary that is aware of subjects that are used for training across the entire federation. We observe an average drop of .0178 in the attack F-1 scores.

We augment our synthetic data study with evaluation of a variant of the FeMNIST dataset [18] as a representative of the real world setting, where we have little room to tune configuration parameters.

6.1 Synthetic Dataset

Since one of the primary objectives of this research is to study how environmental variables of a federation impact subject membership inference risk VJM: I think I've seen this "primary objective" mentioned a couple times now. Perhaps we need to add an "our contributions" part at the end of introduction that explicitly spells out the contributions we make in the paper, including this one.,

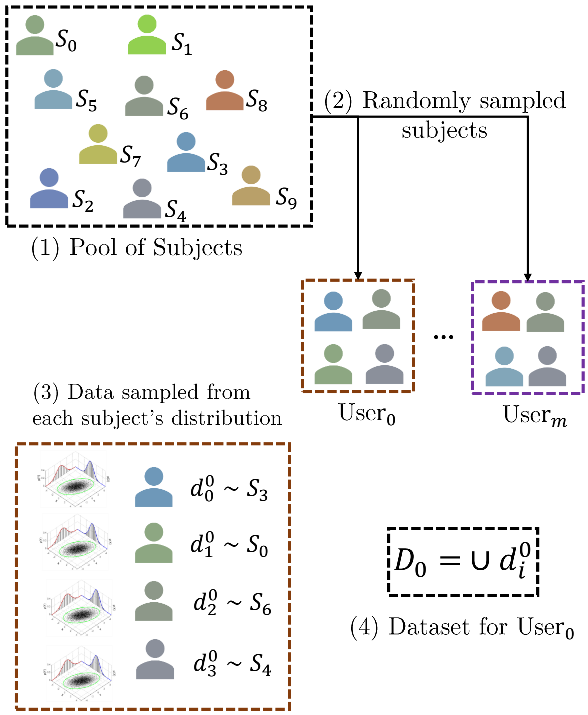


Fig. 3. Dataset creation process for our Synthetic Dataset. Each user is assigned subjects at random, and data from each subject's distribution is sampled to generate a user's dataset.

the ideal setup should allow control over all parameters, even the ones that are usually fixed for a given dataset (e.g. number of subjects per user, items per subject, items per user). For a fully controlled federation environment, we design a synthetic dataset with multiple controllable parameters, quantifying certain aspects of interest in a federation and their impact on subject membership inference risk.

We start with a certain controllable dimensionality for the feature space of data. The ground truth label for each data point is deterministic and computed as the XOR of the features across all dimensions. For a particular data point x with n dimensions:

$$y = \bigoplus_i \mathbb{I}[x_i \geq 0] \tag{7}$$

The data generation process is outlined in Figure 3 and described below:

(1) We model each subject as a parameterized distribution, for which we use a multi-variate Gaussian. We generate random (and valid) mean and covariance matrices for each subject, such that no two subjects have the same parameters to their distributions. Additionally, we enforce a minimum pair-wise separation between all of the subject distributions' means to avoid overlap. This separation is set such that it is not too high to make the subjects too distinct and the inference task trivial, yet low enough to be able to tell any two distributions apart.
(2) Each user in the federation is then assigned a random sample of subjects. These subjects are sampled from the pool of all subjects with replacement, and thus users can have an overlap in the subjects assigned to them.
(3) To construct the user's dataset, data is randomly sampled from distributions of each of the subjects assigned to that particular user. There are two possible extremes when modeling distributions for subjects: each sample being virtually unique and the other with scope for multiple repetitions. The former is more like a patient's blood report readings, while the latter is closer to a customer's shopping cart. We allow for two sampling schemes to capture these two extremes: standard sampling for a multivariate Gaussian and Dirichlet sampling.
(4) The data sampled from each of the user's assigned subjects is then concatenated to form the user's dataset. This process is repeated for all users in the federation.

The number of users, total available subjects, number of subjects per user, and data samples per user, are all controllable parameters of our environment.

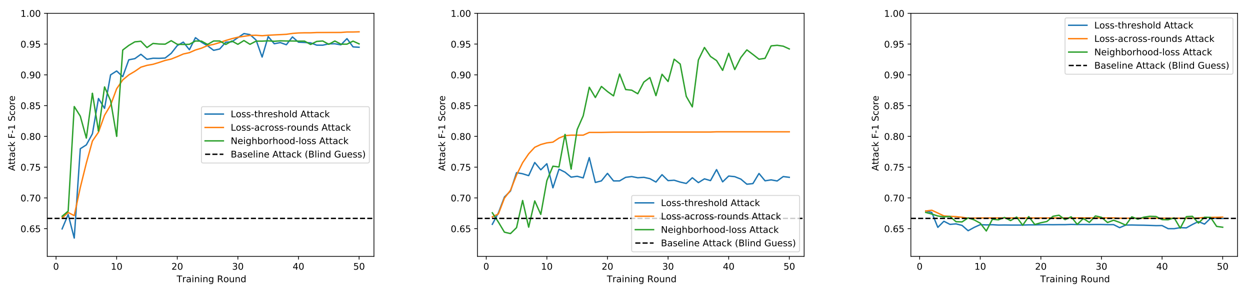
6.2 Results

The success of our inference attacks can depend on several factors:

- **Data Properties**, such as the dimensionality and sampling distribution
- **Model Design and Training**, such as the model architecture and number of training rounds.
- **Federation Properties**, such as the number of users, subjects, and datapoints.

For a comprehensive evaluation of how these factors influence subject membership inference risk, we generate a total of 720 configurations by varying all of the above parameters simultaneously on the synthetic dataset. The exact configuration values are given in Table 1. This extensive grid search is a one-of-its-kind study for Federated Learning systems and meant to expand our understanding of how certain factors, both in and out of the model trainer's control, can influence privacy leakage.

For each experiment, we take a sample of subjects present in the federation along with an equally sized sample of subjects not present in the federation, to use



(a) Config *A*. Final model test accuracy: 99.71%
(b) Config *B*. Final model test accuracy: 50.36%
(c) Config *C*. Final model test accuracy: 65.40%

Fig. 4. Attack F-1 Scores for configurations with varying final test accuracies. Attacks are observed to leak a lot of information, when the models themselves perform exceptionally well on the task (a) or even when they perform marginally better than random (b). However, there are also cases where there is close to little or no leakage when models are neither too accurate nor too inaccurate (c). Parameters for these configurations are given in Table 2.

Configurable	Values Experimented
Sampling Mechanism	{Normal, Dirichlet}
Data Dimensionality	{2, 50, 250, 1000}
Model: Number of Layers	{1, 2, 3}
Model: Number of Epochs	[1, 50]
Users	{10, 100}
Subjects per User	{10, 100, 500}
Items per User	{500, 2000, 10000}

Table 1. Variables for the Synthetic Dataset that we experiment with. Each of these are tried simultaneously, thus yielding all possible configurations with these values.

Parameter	Config *A*	Config *B*	Config *C*
Data Dimensionality	1000	1000	2
Sampling Mechanism	Dirichlet	Normal	Normal
Model Hidden Dimensions	[256, 16, 16, 4]	[8]	[2]

Table 2. Experiment parameters for the configurations given in Figure 4. All of these configurations correspond to 10000 items per user, 10 subjects per user, and 10 users.

as our test set. This process is repeated multiple times, randomly sampling subjects not present in the federation. For computing thresholds across all the attacks, the adversary divides its data into two splits. The first split is used to derive threshold(s), while the second one is used for actually launching the attack.

For computing the F-1 score, we count correctly predicting the presence/absence of a subject's data in the federation as a hit (1) and incorrect as a miss (0). We report mean F-1 scores within 95% confidence intervals for our graphs and tables. Our initial experiments with a few randomly selected configurations show how the *Neighborhood Loss Attack* has the capacity to outperform the other two attacks in some cases, as visible in Figure 4a. However, the *Neighborhood-Loss Attack* is computationally much more expensive than the other attacks, and its performance is close to the others for most of the other cases. Thus, we report F-1 scores with the *Loss-Threshold* attack (Section 5.1) for all our experimental configurations.

AS: Talk about the three configurations for which we have plots and Tables, and why we picked these three (and what kind of cases they correspond to.

6.2.1 Data Properties

Sampling Mechanism. We plot Attack F-1 scores across training rounds, for data distributions with standard and Dirichlet sampling (Figure 5). We observe Dirichlet sampling to exhibit a significantly higher inference risk than the case of regular sampling. This is expected since repeated samples would make inferring a subject's membership easier, almost reducing it to data-point membership inference. These sampling mechanisms represent two extreme cases possible in real-world federation systems: each datum being sampled uniquely (like blood cell counts) versus high density around specific data points (like grocery store purchases). Real-world datasets would be somewhere between these two, and having results for them both gives a good sense of the expected range of inference risk for real-world datasets.

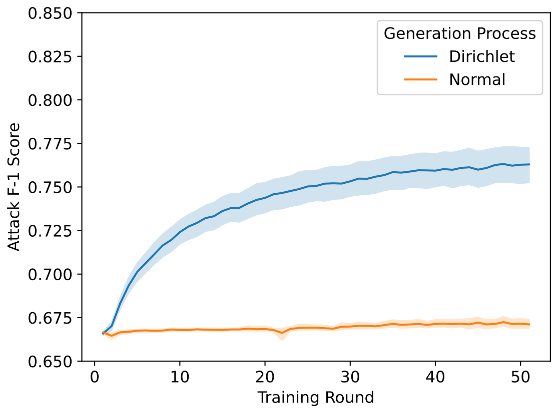


Fig. 5. Attack F-1 Score across training rounds for datasets generation with Standard and Dirichlet Sampling.

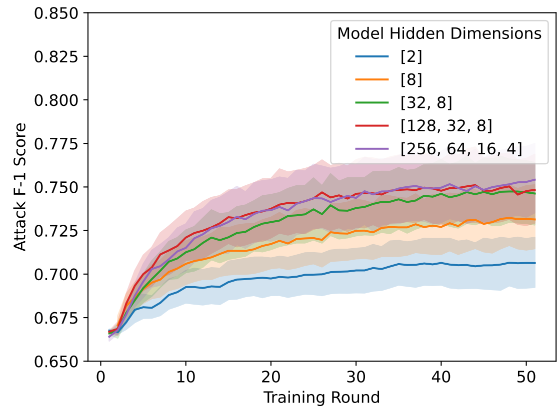


Fig. 7. Attack F-1 Score across training rounds for datasets with different model architectures. Hidden dimensions refers to the number and sizes of intermediate layers for the neural networks used.

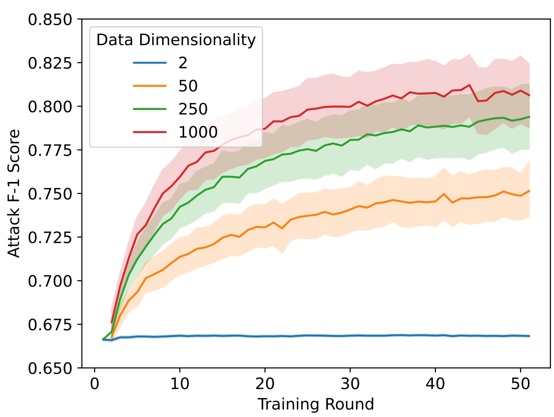


Fig. 6. Attack F-1 Score across training rounds for datasets with different feature dimensionality.

Dimensionality. Inference risk seems to correlate positively with the dimensionality of the feature space (Figure 6), with stagnation in the F-1 scores for inference as the dimensionality increases beyond a certain point. Subject distributions in lower dimensions are likely to be closer to each other. On the other hand, the same number of distributions in a higher-dimensional space would be distributed much more sparsely, owing to the curse of dimensionality. Thus, the latter would be understandably easier to distinguish than the former. Model trainers thus need to be cautious when working with high dimensional data since that makes them highly susceptible to such inference attacks.

6.2.2 Model Design and Training

Model Complexity. We vary model complexity by adjusting both the number of layers and neurons per layer, going from a single hidden layer neural network up to one with four hidden layers (Figure 7). Inference risk seems to increase model complexity but seems to plateau beyond model complexity required for the task. The risk increases as we increase the number of neurons for the same one-hidden-layer architecture and then again on adding an additional hidden layer. However, more complex models seem to exhibit almost similar inference risk beyond that. Interestingly, inference risk for the under-parameterized models is only slightly better than random guessing, suggesting it may be in the model trainers' interest to use models that are not too complex for a given task. This suggestion is in the model trainer's own interest, since using a smaller model also makes the model less likely to overfit.

Model Training. Similar to trends with model complexity, we observe that inference risk increases as the model continue to train and then plateaus towards the latter half of training rounds, which is a few rounds after the model's loss has converged on both train and test data. These observations are clearly visible in all of the previous figures, and especially in Figure 7. Based on these observations, it would make sense not to train the model for too many rounds- only enough to achieve satisfactory performance. Such a decision may hurt the model trainer, since some studies in the literature [40] demonstrate how training beyond convergence can confer benefits like better robustness, generalization, and interpretability.

6.2.3 Federation Properties

For a given number of data points corresponding to a subject, the underlying federation can have several different configurations: different number of users, subjects per user, as well as items per user. Although none of these are in control of an adversary, understanding how they impact subject membership inference risk can be advantageous in both designing and understanding such attacks. We study these trends across varying parameters of the configuration setup and observe very peculiar trends. We split our analyses into two categories: *Few Subjects per User* (10) and *Many Subjects per User* (100/500). We further calculate the total number of items per subject in each configuration, and bin them into three categories: $(4, 100]$ (*low*), $(100, 800]$ (*medium*), and $(800, 2000]$ (*high*).

Few Subjects per User. For the case with only a few subjects per user (Figure 8a), the attack F-1 scores (Y-axis) are higher for the cases with fewer total subjects (blue), compared to settings with more total subjects in the federation (orange). This trend is expected, as having more (subject) distributions in the same feature co-domain would make overlap between distributions more likely, making it harder for an adversary to distinguish between any two distributions. Attack F-1 scores change as the number of items per subject (X-axis) increase, but the trends ore somewhat conflicting for the *low* and *medium* cases of items per subject. For the former, the F-1 scores increase with increase in items. This is expected, since having more items per subject would make it more likely for the model to generalize well to a given subject's distribution (as opposed to overfitting to a few points), making it easier for an adversary to infer membership. Although the attack scores decrease for the *medium* case of items per subject, the decrease is substantial and within error of margin, effectively staying unaffected.

Many Subjects per User. When we have a sufficiently high number of subjects per user, we observe an increase in risk as we increase the number of items per subject (Figure 8b). The gains in attack performance too taper off once there are sufficiently large number of items per subject (*medium* v/s *high*) Since the total number of subjects in the system is high enough, the effects of potential overlap between subject distributions (mentioned earlier) start to converge; the two cases (orange and green) are thus not affected much by an increase in the total number of subjects and are close in their performance.

Our analyses show how configurations with a lot of subjects in the federation increase in susceptibility to subject membership inference as the data available per subject increases. At the same time, configurations with few subjects in the federation are highly likely to leak subject membership.

6.3 High Risk Configurations

To better understand what combinations of the various parameters may make the overall federation more susceptible to these inference attacks, we choose to look at highly successful attacks: ones with both precision and F-1 scores > 0.9. Close analysis of the filtered configurations yields some common configuration attributes:

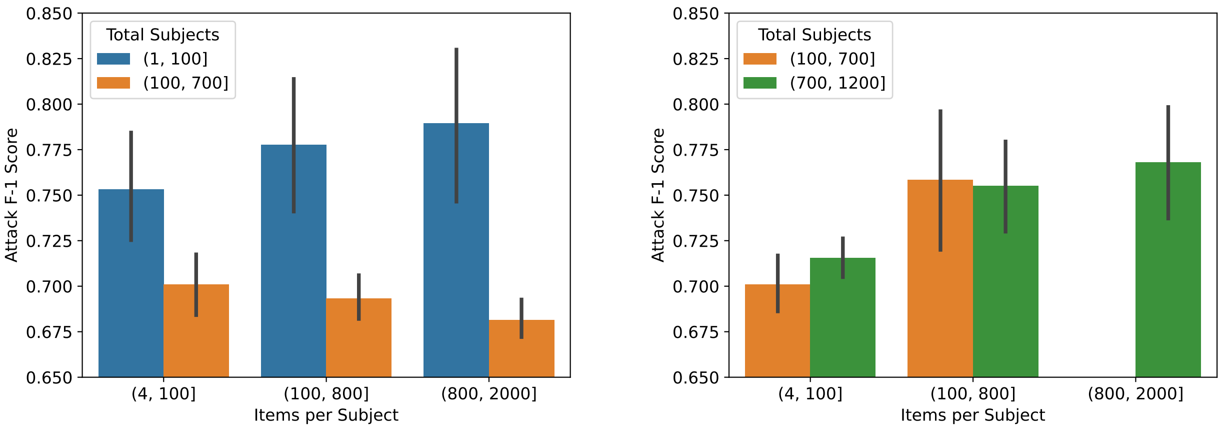
- High data dimensionality: 1000
- Dirichlet sampling while generating data
- Large model architectures: ≥ 3 hidden layers, and
- Models trained for many rounds: ≥ 20.

Looking out for these attributes can help machine learning practitioners identify cases that may be highly susceptible to subject-level membership inference attacks. In addition to these analyses, we also looked at the generalization gap (which we define as the difference between the train and test accuracies) for models and the attacks' F-1 scores. We noted a **negative** correlation score (~ -0.6), which is exactly opposite to what the literature says [11] about inference attacks, and needs further investigation.

6.4 Real-data case study: FeMNIST

Our findings on synthetic data show that certain dataset properties make the inference attacks more difficult - high dimensionality, larger overlap across users, and less exact repetition of data items for the same subject. To test whether these findings are transferable to real-world datasets, we use FeMNIST [18], the federated extended MNIST [41] dataset, which is an image classification task for handwritten digits and letters.

Dataset Description. FeMNIST's digits and letters themselves have been written by $3,500$ distinct individuals, and FeMNIST partitions these images by individual authors. Each author has contributed hundreds of sample images. Ordinarily, FL research experiments [18] map each author to a federation user, resulting in a



(a) 10 Subjects per User

(b) 100/500 Subjects per User

Fig. 8. Attack F-1 Scores while varying number of total subjects and items per subject, for 10 subjects per user (a) and 100 subjects per user (b) in the Federation.

3, 500-user federation. In our experiments, we instead map authors to subjects, resulting in a federation with 3, 500 subjects whose data is randomly scattered among a handful of federation users (16 in our experiments) to emulate a cross-silo FL setting. Multiple federation users may host images from the same subject, though we do not distribute any individual image to more than one federation user. This reconfigured dataset is especially suitable for cross-silo FL and our subject membership attacks study. The data points themselves are 28x28 pixel, black-and-white pictures of a single handwritten character.

Setup. We use the CNN model on FeMNIST appearing in the LEAF data suite [18] as our target model to train. More specifically, the model consists of two Convolution layers interleaved with ReLU activations and Max-pooling layers, followed by two Linear layers. We train the CNN model on the dataset with 3, 500 data subjects. Each subject has ~ 140 data points on average, with its data more-or-less equally spread across 16 federation users. Subject membership inference attacks aim to determine, given some samples from a set of subjects, whether a target subject's data was part of the training set.

Results. We expect this to be a difficult attack for several reasons. The data is high dimensional (784 inputs), does not repeat, and has considerable overlap between users. While differentiating handwriting styles of different individuals is possible to some degree, is not easy (or necessarily reliable) for trained experts to provide forensic matches of handwriting [42]. Some prior

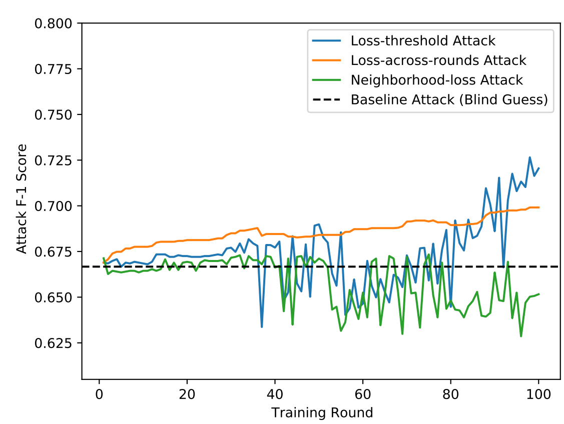


Fig. 9. FEMNIST Model. Final model test accuracy: 86.42%.

work on membership inference for individual data points on MNIST corroborates these prior beliefs [35]. Attack results are plotted in Figure 9, showing how the F-1 Scores are no better than random-guessing for the *Loss-Threshold Attack* for most rounds but do start rising as the model nears completion, which is what is usually released in production.

6.5 Mitigation

One of the most commonly prescribed method for defending against membership inference attack is training ML models with Differential Privacy(DP) [2]. In particular, Federated Learning models can be trained with Local Differential Privacy[43] at various granularities as

Metric	FL	Item	User	Subject
Model Accuracy	.8642	.7818	.3738	.7493
Accuracy	$.61 \pm .05$	$.54 \pm .04$	$.50 \pm .00$	$.50 \pm .00$
Precision	$.57 \pm .03$	$.52 \pm .02$	$.50 \pm .00$	$.50 \pm .00$
Recall	$.96 \pm .00$	$1.0 \pm .00$	$1.0 \pm .00$	$1.0 \pm .00$
F_1 Score	$.71 \pm .03$	$.69 \pm .02$	$.67 \pm .00$	$.67 \pm .00$

Table 3. Model and attack metrics under different DP granularities while using the *Loss-Threshold Attack*, using CNNs on the FEMNIST [18] Dataset.

described before (AS: Add back-reference). From the 720 configurations described earlier, we select the most vulnerable one, and train models on them with DP at $\epsilon = 2.0$ and $\delta = 10^{-5}$ (AS: Same parameters for all three levels of DP? Might want to mention them here again explicitly, along with references to whichever versions we use while training DP models). We train these models for 20 rounds (AS: Is it 40 for FEMNIST?), and use $\sigma = 1.8346$ (AS: May wanna add these parameters for FEMNIST as well). Table 3 shows that DP does help significantly reduce the attack accuracy. Indeed, as has been reported in the literature, this comes at the cost of reduction in model accuracy.

Metric	FL	Item	User	Subject
Synthetic Dataset Config *A*				
Model Accuracy	.9919	.7945	.7290	.6368
Accuracy	$.93 \pm .01$	$.66 \pm .04$	$.59 \pm .02$	$.58 \pm .05$
Precision	$.89 \pm .02$	$.61 \pm .04$	$.55 \pm .02$	$.55 \pm .03$
Recall	$.98 \pm .02$	$.93 \pm .06$	$.98 \pm .02$	$.89 \pm .05$
F_1 Score	$.93 \pm .01$	$.74 \pm .01$	$.71 \pm .01$	$.68 \pm .02$
Synthetic Dataset Config *B*				
Model Accuracy	.5035	.5085	.5018	.5075
Accuracy	$.78 \pm .02$	$.52 \pm .04$	$.50 \pm .01$	$.52 \pm .03$
Precision	$.73 \pm .04$	$.51 \pm .02$	$.50 \pm .00$	$.51 \pm .02$
Recall	$.91 \pm .05$	$.97 \pm .06$	$1.0 \pm .00$	$.98 \pm .03$
F_1 Score	$.81 \pm .02$	$.67 \pm .00$	$.67 \pm .00$	$.67 \pm .01$
Synthetic Dataset Config *C*				
Model Accuracy	.6545	.6291	.8358	.6383
Accuracy	$.53 \pm .04$	$.53 \pm .04$	$.52 \pm .03$	$.50 \pm .01$
Precision	$.51 \pm .01$	$.52 \pm .02$	$.51 \pm .02$	$.50 \pm .01$
Recall	$.98 \pm .03$	$.97 \pm .04$	$.98 \pm .03$	$1.0 \pm .01$
F_1 Score	$.67 \pm .00$	$.68 \pm .01$	$.67 \pm .01$	$.67 \pm .00$

Table 4. Model accuracies and attack metrics (accuracy, precision, recall, F_1 score) under different DP granularities while using the *Loss-Threshold Attack*, using MLPs on the Synthetic Dataset (Section 6.1). Parameters for these configurations are given in Table 2.

7 Conclusion

Privacy in Federated Learning is typically only studied for individual data items or users participating in the federation. However, in complex cross-silo FL settings, we ultimately care about protecting the privacy of individual data subjects. This is the first paper to propose *subject-level* membership inference attacks, which can aid in the empirical measurement of data subject's privacy leakage. We show that the three proposed attack variants are successful in retrieving subject membership information from a wide variety of Federated Learning models. We present a first-of-its-kind, synthetic data generator based study, in which we simulate several hundred FL configurations and measure the accuracy of inference attacks for them. For these simulations, we vary three main aspects of the system - data, model, and FL structure, and find that factors like data distribution, data dimensionality, model complexity, and training protocols, and the size and composition of the federation in terms of the number of users, data subjects and data items, all have a substantial impact on the attack accuracy. By systematically varying some of the variables that can affect attack success, this study provides invaluable practical guidance to model designers and ML practitioners on what makes their models more vulnerable. On the mitigation front, we retrain some of these configurations with Differential Privacy and verify that it indeed helps reduce the attack accuracy significantly, albeit at the cost of model accuracy. Finally, we illustrate the entire process of using subject-level inference attacks to measure privacy leakage and assess the effectiveness of mitigation using a real-world dataset, FEMNIST. We hope that this study will help ML practitioners focus on and protect the most important asset in the FL data ecosystem - the people.

References

[1] J. Konecný, B. McMahan, and D. Ramage, "Federated Optimization: Distributed Optimization Beyond the Datacenter," *CoRR*, vol. abs/1511.03575, 2015. [Online]. Available: http://arxiv.org/abs/1511.03575

[2] 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.

[3] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning Differentially Private Recurrent Language Models,"

in *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[4] Y. Liu, A. T. Suresh, F. X. Yu, S. Kumar, and M. Riley, "Learning discrete distributions: user vs item-level privacy," *CoRR*, vol. abs/2007.13660, 2020. [Online]. Available: https://arxiv.org/abs/2007.13660

[5] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. A. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao, "Advances and Open Problems in Federated Learning," *CoRR*, vol. abs/1912.04977, 2019. [Online]. Available: http://arxiv.org/abs/1912.04977

[6] C. Dwork, A. Roth *et al.*, "The Algorithmic Foundations of Differential Privacy." *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, 2014.

[7] V. J. Maratha and P. Kanani, "Subject Granular Differential Privacy in Federated Learning," *CCS 2021 Workshop on Privacy Preserving Machine Learning (PPML)*, 2021.

[8] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Inference Attacks Against Collaborative Learning," *CoRR*, vol. abs/1805.04049, 2018. [Online]. Available: http://arxiv.org/abs/1805.04049

[9] 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, SP 2019, San Francisco, CA, USA, May 19-23, 2019*. IEEE, 2019, pp. 739–753. [Online]. Available: https://doi.org/10.1109/SP.2019.00065

[10] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[11] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting," in *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 2018, pp. 268–282.

[12] 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*, 2015, pp. 1322–1333.

[13] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 253–261.

[14] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers," *International Journal of Security and Networks*, vol. 10, no. 3, pp. 137–150, 2015.

[15] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, "Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 619–633.

[16] A. Suri and D. Evans, "Formalizing and Estimating Distribution Inference Risks," *arXiv preprint arXiv:2109.06024*, 2021.

[17] 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.

[18] S. Caldas, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar, "LEAF: A Benchmark for Federated Settings," *CoRR*, vol. abs/1812.01097, 2018.

[19] C. Dwork, "Differential privacy," in *Automata, Languages and Programming, 33rd International Colloquium, ICALP*, 2006, pp. 1–12.

[20] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. Guo, "Privacy preservation in federated learning: An insightful survey from the GDPR perspective," *Comput. Secur.*, vol. 110, p. 102402, Nov. 2021.

[21] M. Jegorova, C. Kaul, C. Mayor, A. Q. O'Neil, A. Weir, R. Murray-Smith, and S. A. Tsaftaris, "Survey: Leakage and Privacy at Inference Time," 2021.

[22] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, M. Backes, E. De Cristofaro, M. Fritz, and Y. Zhang, "ML-Doctor: Holistic risk assessment of inference attacks against machine learning models," *arXiv preprint arXiv:2102. 02551*, 2021.

[23] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," 2017.

[24] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models," 2019.

[25] V. Shejwalkar, H. A. Inan, A. Houmansadr, and R. Sim, "Membership inference attacks against NLP classification models."

[26] B. Jayaraman, L. Wang, K. Knipmeyer, Q. Gu, and D. Evans, "Revisiting Membership Inference Under Realistic Assumptions," May 2020.

[27] L. Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[28] F. Mo, A. Borovykh, M. Malekzadeh, H. Haddadi, and S. Demetriou, "Layer-wise Characterization of Latent Information Leakage in Federated Learning," 2020.

[29] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting Gradients – How easy is it to break privacy in federated learning?" Mar. 2020.

[30] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora, "Evaluating Gradient Inversion Attacks and Defenses in Federated Learning," Nov. 2021.

[31] A. Wainakh, F. Ventola, T. Müßig, J. Keim, C. G. Cordero, E. Zimmer, T. Grube, K. Kersting, and M. Mühlhäuser, "User Label Leakage from Gradients in Federated Learning," May 2021.

[32] Y. Liu, X. Zhu, J. Wang, and J. Xiao, "A Quantitative Metric for Privacy Leakage in Federated Learning," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech*

and Signal Processing (ICASSP), Jun. 2021, pp. 3065–3069.

[33] 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*, ser. CCS '17. New York, NY, USA: Association for Computing Machinery, Oct. 2017, pp. 603–618.

[34] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, Apr. 2019, pp. 2512–2520.

[35] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Demystifying Membership Inference Attacks in Machine Learning as a Service," *IEEE Trans. Serv. Comput.*, pp. 1–1, 2019.

[36] 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*, vol. 15, pp. 3454–3469, 2020.

[37] W. Zhang, S. Tople, and O. Ohrimenko, "Leakage of Dataset Properties in {Multi-Party} Machine Learning," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2687–2704.

[38] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[39] J. Wang, C. Lan, C. Liu, Y. Ouyang, and T. Qin, "Generalizing to Unseen Domains: A Survey on Domain Generalization," *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI), Survey Track*, 2021.

[40] V. Papyan, X. Y. Han, and D. L. Donoho, "Prevalence of neural collapse during the terminal phase of deep learning training," *Proceedings of the National Academy of Sciences*, vol. 117, no. 40, 2020.

[41] L. Deng, "The MNIST Database of Handwritten Digit Images for Machine Learning Research," *IEEE signal processing magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[42] U. S. v. Saelee, "162 f. supp. 2d 1097," August 2001, united States District Court, D. Alaska. [Online]. Available: https://law.justia.com/cases/federal/district-courts/FSupp2/162/1097/2319759/

[43] S. Truex, L. Liu, K. H. Chow, M. E. Gursoy, and W. Wei, "LDP-Fed: Federated Learning with Local Differential Privacy," in *Proceedings of the 3rd International Workshop on Edge Systems, Analytics and Networking, EdgeSys@EuroSys 2020, Heraklion, Greece, April 27, 2020*. ACM, 2020, pp. 61–66.