

FedPDD: A Privacy-preserving Double Distillation Framework for Cross-silo Federated Recommendation

Sheng Wan*
Dept. of CSE
SUSTech & HKUST
Hong Kong, China
swanae@cse.ust.hk

Dashan Gao*
Dept. of CSE
SUSTech & HKUST
Hong Kong, China
dgaoaa@cse.ust.hk

Hanlin Gu
Webank
Shenzhen, China
ghltsl123@gmail.com

Daning Hu†
Dept. of Finance
SUSTech
Shenzhen, China
hdaning@gmail.com

Abstract—Cross-platform recommendation aims to improve recommendation accuracy by gathering heterogeneous features from different platforms. However, such cross-silo collaborations between platforms are restricted by increasingly stringent privacy protection regulations, thus data cannot be aggregated for training. Federated learning (FL) is a practical solution to deal with the data silo problem in recommendation scenarios. Existing cross-silo FL methods transmit model information to collaboratively build a global model by leveraging the data of overlapped users. However, in reality, the number of overlapped users is often very small, thus largely limiting the performance of such approaches. Moreover, transmitting model information during training requires high communication costs and may cause serious privacy leakage. In this paper, we propose a novel privacy-preserving double distillation framework named FedPDD for cross-silo federated recommendation, which efficiently transfers knowledge when overlapped users are limited. Specifically, our double distillation strategy enables local models to learn not only explicit knowledge from the other party but also implicit knowledge from its past predictions. Moreover, to ensure privacy and high efficiency, we employ an offline training scheme to reduce communication needs and privacy leakage risk. In addition, we adopt differential privacy to further protect the transmitted information. The experiments on two real-world recommendation datasets, HetRec-MovieLens and Criteo, demonstrate the effectiveness of FedPDD compared to the state-of-the-art approaches.

I. INTRODUCTION

Benefiting from the explosion of data, deep learning-based recommendation systems have gained significant attention by overcoming obstacles of conventional models and achieving high recommendation quality [1]. Unfortunately, in reality, this wealth of data is often separated into different platforms and owned by different entities. For example, people can chat with friends on WhatsApp, watch favorite videos on TikTok or Youtube, and buy wanted stuff on Amazon. Collecting these features from different platforms can help to build a more accurate user profile and provide a better recommendation. However, cross-silo collaborations among different platforms are restricted by data protection regulations such as General

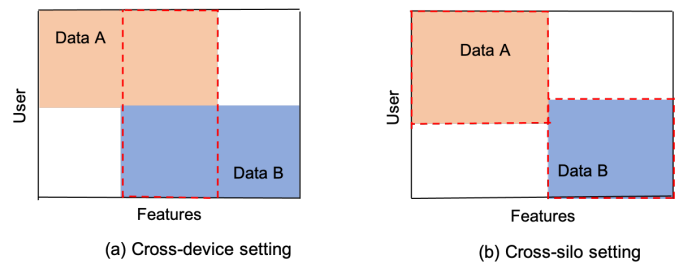


Fig. 1. In cross-device FL setting, participants are a large number of individual customers (2C) that have the same feature space. In cross-silo FL setting, participants are a small number of business partners (2B) that have partially overlapped user spaces and different feature spaces. Here we assume that there is no feature overlapping in our cross-silo FL setting. Data in the red box are used for training.

Data Protection Regulation (GDPR) and data cannot be centralized for training.

To tackle the privacy issue for cross-silo recommendation, a practical solution is Federated Learning (FL) [2], [3]. FL enables multiple parties to collaboratively train a global model while private data resides locally on the data owners and therefore can largely reduce systemic privacy risks. Existing cross-silo FL methods [4]–[6] try to fix this problem by using the overlapped samples across participants and viewing it as a multi-view learning problem. The performance of such approaches highly relies on the number of overlapped users between parties. However, in reality, such overlapped data is often limited and thereby may cause the performance to be even worse than the locally trained models. Moreover, these approaches transmit feature or model information during training, which requires high communication costs and has serious privacy weaknesses. Recent studies [7]–[9] show that sharing model or feature information could still lead to private data leakage.

To address these challenges, we propose a novel privacy-preserving double distillation framework named FedPDD for cross-silo federated recommendation. We design a double distillation strategy, which enables local models to learn both implicit knowledge from themselves and explicit knowledge from the other party. Specifically, we distill implicit knowledge

*These authors contributed equally to this work.

†Corresponding author.

from the past predictions of local models and distill explicit knowledge from the ensemble predictions of current local models. The key idea is that we provide multiple informative sources for local model training. Therefore, by learning from these sources, FedPDD is able to enhance model performance and generalization ability. Moreover, we employ an offline distillation strategy and only transmit model output during training. Parties only communicate with the server during the federated ensemble stage and the size of the model output is much smaller than the model itself. Accordingly, our training strategy largely reduces communication needs and limits the exposure of private information to the server. In addition, we adopt differential privacy [10] to further protect the communication process. We experiment on two real-world recommendation datasets showing that FedPDD significantly boosts local model performance by up to 3.94% and outperforms the state-of-the-arts by up to 3.98%.

Overall, our main contributions are as follows:

- We propose a novel privacy-preserving double distillation method named FedPDD for cross-silo federated recommendation. Our method enables local models to learn from not only private labels and the other party but also itself, which enhances model performance and generalization ability when overlapped samples are limited.
- We employ an offline training strategy to reduce communication needs and privacy leakage risk. Moreover, we adopt differential privacy to further protect the communication process and provide a theoretical privacy analysis of FedPDD.
- We conduct experiments on two public real-world datasets and the results demonstrate the effectiveness of FedPDD by up to 3.98% further improvements compared to the state-of-the-art approaches.

II. BACKGROUND AND RELATED WORK

A. Federated Learning

Federated Learning [2], [3] allows multiple participants to collaboratively train a global model while keeping training data in local. The data resides locally on the data owners and therefore largely reduces systemic privacy risks. Based on scenarios, FL can be divided into two kinds of settings: cross-device FL and cross-silo FL [3]. As shown in Figure 1, in the cross-device setting, participants are a large number of individual customers (2C) that have the same feature space, while participants are a small number of business partners (2B) that have partially overlapped user spaces and different feature spaces in the cross-silo setting. In this work, we focus on the cross-silo setting that features are not overlapped.

Liu et al. [4] first proposed a transfer learning method named federated transfer learning (FTL) to transfer knowledge through the overlapped user data between parties. They assumed that only one party owns the labels and aims to improve the model performance by leveraging the knowledge (i.e., features) from other parties. Under such an assumption, their approach merely leveraged the overlapped users across parties.

TABLE I
DEFINITION OF NOTATIONS

Notation	Description
D^k	local dataset on party k
D^c	overlapped data shared between two parties
α	ratio of overlapped data to the training set
f^k	local model on party k
$f_{b(n)}^k$	best local model on party k in round n
z^k	logit output of local model f^k
z^t	ensemble logit
z'^t	perturbed ensemble logit
p^k	soft target output of local model f^k
p^t	ensemble teacher knowledge
p_b^t	self teacher knowledge
σ_T	general softmax function
T	distillation temperature

Existing cross-silo FL methods [5], [6] mostly follow this direction which views the distributed features of overlapped users in different parties as different views of these data and regards it as a multi-view learning problem. Specifically, Feng et al. [5] established a Multi-participant Multi-class Vertical Federated Learning (MMVFL) framework. They utilized multi-view learning methods to securely share label information between participants. Kang et al. [6] proposed a self-supervised multi-view learning method called FedMVT under the cross-silo FL setting. They built a model based on the overlapped data to predict the missing features. However, these studies build models on the overlapped user data across parties and prediction accuracy of the global model highly relies on the amount of overlapped data. When such data is limited, the performance of these approaches may be even worse than the local fine-tuned models. In contrast, FedPDD leverages both the overlapped data and the non-overlapped data to enhance model performance through knowledge distillation and can achieve superior performance compared to the state-of-the-arts when overlapped data is limited.

B. Federated Recommendation System

Inspired by the success of federated learning [11], federated recommendation system is proposed to address the privacy and data silo problems in the recommendation system [2]. Existing federated recommendation works such as federated matrix factorization methods [12]–[16] and federated collaborative filtering methods [17], [18] mainly focus on the cross-device FL setting. They adopt the idea of the typical federated learning algorithm FedAvg [11] and average the updates of gradient information from participants. However, this line of methods has serious privacy issues. Sharing gradient information could lead to private data leakage as proofed by [14].

We notice that there is a contemporary work [19] similar to our approach. The authors propose a cross-silo federated recommendation framework using split knowledge distillation. However, they assume that there are massive unlabeled overlapped data between parties and ignore privacy issues, which is a key concern in cross-silo FedRec. Exposing more overlapped data during training will inevitably lead to a higher risk of privacy leakage and largely increase the privacy budget.

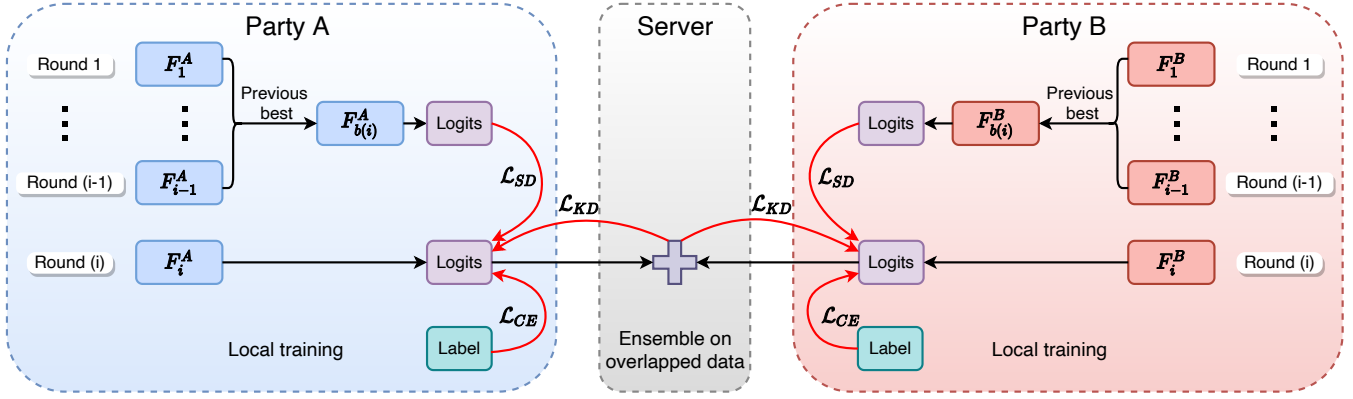


Fig. 2. The overview of our proposed FedPDD. During training, each party trains its local model via three kinds of knowledge from the ground truth labels, ensemble of local models and past predictions of local models.

C. Federated Knowledge Distillation

Knowledge distillation [20] distills knowledge from a teacher model to a student model through the soft target of the teacher model. More precisely, the student model learns from the teacher model by imitating the soft target distribution of the teacher model as model output through a Kullback-Leibler (KL) divergence loss defined as:

$$L_{KD}(\mathbf{p}, \mathbf{q}) = T^2 KL(\mathbf{p}||\mathbf{q}), \quad (1)$$

where \mathbf{p} and \mathbf{q} are the soften output of the student model and teacher model. T is the temperature parameter. Denote the student logit as z_s and teacher logit as z_t . Then $\mathbf{p} = \text{softmax}(z_s/T)$ and $\mathbf{q} = \text{softmax}(z_t/T)$.

Ensemble distillation [21] ensembles knowledge from multiple teacher networks to yield better teacher knowledge. Current federated knowledge distillation methods [22]–[26] adopts this idea to fuse heterogeneous local models into a global model and share model outputs instead of gradients. FedMD [22] first trained student models through averaged logits of each sample on a public labeled dataset to reduce communication costs during training. Along this direction, Lin et al. [23] designed a more robust model distillation framework named FedDF, which allows for heterogeneous client models and data. Chang et al. [24] used group knowledge transfer to reduce the communication and computation overload for edge devices. Gong et al. [25] proposed to use cross-domain unlabelled public data to protect private data privacy. These studies are designed for the cross-device setting and transfer knowledge through labeled public datasets [22], [27], or unlabeled public datasets [23], [25], [26]. The improvement of their model performance unavoidably benefits from the extra public data. Li et al. [26] proposed a practical one-shot federated learning algorithm for cross-silo FL setting. However, they still leverage an unlabeled public dataset to transfer knowledge.

Along with these studies, FedPDD is designed to transfer knowledge through limited overlapped user data for cross-silo FedRec. Unlike the above methods that only explore explicit knowledge from the other party, we propose a double distillation strategy that enables local models to fully exploit

both explicit knowledge from the other party and implicit knowledge from itself.

III. METHODOLOGY

A. Problem Statement

Consider two parties A, B and a central server. Each party has a private labeled dataset $D^k := \{\mathbf{x}_i^k, y_i\}_{i=1}^{|D^k|}$ and can independently design its model f^k , $k = \{A, B\}$. There exists limited overlapped users shared between two parties $D^c := \{\mathbf{x}_i^A, \mathbf{x}_i^B, y_i\}_{i=1}^{|D^c|}$. Let $\alpha = \frac{|D^c|}{|D^A| + |D^B|}$ denotes the overlapped data ratio. Our target is to improve the performance of the local models in both parties as well as the global model in the central server without exchanging raw data. Moreover, due to privacy concerns, the private information exposed to the server or the other party should be as less as possible. We summarize the notations used in this paper in table I.

B. Overview of FedPDD

The overall pipeline of FedPDD consists of three steps: pretraining, federated ensemble and local training. In the first stage, local models are pretrained on the local private datasets from scratch. In the federated ensemble stage, each party makes predictions on the overlapped user data using its pretrained local model. Then the central server aggregates the local predictions from two parties to obtain the ensemble teacher knowledge and distributes it to parties. The aggregation is protected by differential privacy as illustrated in section III-E. In the local training stage, we propose a double knowledge distillation strategy to fully explore both implicit knowledge from itself and explicit knowledge from the other party. That is to say, local models have two teachers and learn from multiple informative sources simultaneously during training. Therefore, by providing more teachers, our double distillation strategy enhances model performance and generalization ability. The details are illustrated in section III-C and III-D.

We denote one local training and one federated ensemble as a round and repeat these two steps until models achieve convergence. Note that all training is performed offline and

communications are only required in the federated ensemble stage. This offline learning strategy largely decreases communication needs and privacy leakage risk. Algorithm 1 summarizes the whole training process of FedPDD.

C. Distilling Implicit Knowledge

In order to fully explore the implicit knowledge of local models, we propose a self-distillation strategy to enable the local model to distill knowledge from itself. That is to say, the teacher is the student model itself. Specifically, the student in the previous rounds becomes the teacher of itself in the current round. The key idea is that we regard the previous student model outputs as different views of the features, which provides more information for the training [28].

We design a simple but effective method to obtain the self-teacher model. We compare all intermediate local models in previous rounds as the candidates and select the one with the best performance as the teacher model for the current round. Considering deep learning models usually have a large model size, we only maintain one historical best model for each party and replace it whenever there is an improved one.

Denote the best local model on party k in previous $n - 1$ round as $f_{b(n-1)}^k$. To learn from the implicit knowledge, we let the output of local model f^k approximate the output of the teacher $f_{b(n-1)}^k$ in round n through the self-distillation (SD) loss L_{SD}^k . It is given by the KL divergence between local model output \mathbf{p}^k and the teacher model output \mathbf{p}_b^k :

$$L_{SD}^k(\mathbf{p}^k, \mathbf{p}_b^k) = T_{SD}^2 KL(\mathbf{p}^k || \mathbf{p}_b^k), \quad (2)$$

where T_{SD} is the self-distillation temperature.

D. Distilling Explicit Knowledge

We adopt ensemble distillation [21] to leverage explicit knowledge from the other party through the overlapped user data. The key idea is that the ensemble of student models often yields improvements in system performance compared to the performance of individual models. To distill the explicit knowledge, we regard the ensemble results of local model predictions on the overlapped data as the ensemble teacher knowledge. By imitating this teacher knowledge, local models are able to learn from the other party.

Denote the ensemble teacher knowledge as \mathbf{p}_c^t . The ensemble distillation loss is given by the KL divergence between local model output \mathbf{p}^k and the ensemble teacher knowledge:

$$L_{KD}^k(\mathbf{p}^k, \mathbf{p}_c^t) = T_{ED}^2 KL(\mathbf{p}^k || \mathbf{p}_c^t), \quad (3)$$

where T_{ED} is the ensemble distillation temperature.

E. Federated Ensemble

During federated ensemble stage, communication is protected by two levels of privacy. First, we only send model outputs instead of sending model parameters or gradients. Second, if the central server is curious, directly updating local logits may have the risk of privacy leakage. Inspired by PATE [29], we perturb the local output logits with a Gaussian noise to ensure a higher privacy guarantee.

Algorithm 1 Proposed FedPDD algorithm

Input: Local dataset D_A, D_B , overlapped dataset D_c , round number n , temperature T , trade-off weights β, γ, w .

Output: Best local models $f_{b(n)}^A, f_{b(n)}^B$

```

1: Let  $i = 1$ .
2: while  $i \leq n$  do
3:   // Perform local training
4:   for  $k \in \{A, B\}$  do
5:     while not converge do
6:        $f_{b(i)}^k = f_{b(i-1)}^k$ 
7:       Compute Loss based on equation 10
8:       Compute gradients and update  $f^k$ 
9:       if  $f^k$  is better than  $f_{b(i)}^k$  then
10:        Update  $f_{b(i)}^k = f^k$ 
11:       end if
12:     end while
13:   end for
14:   // Perform federated ensemble
15:   for  $k \in \{A, B\}$  do
16:     for each overlapped data  $x_c$  in  $D_c$  do
17:       Compute  $z_c^k$  and perturb with Gaussian noise
18:       Send the perturbed logit  $z_c'^k$  to the server
19:       Server computes the ensemble soft target distribution  $\mathbf{p}_c^t$  and send it back to parties
20:     end for
21:   end for
22: end while
23: return best local models  $f_{b(n)}^A, f_{b(n)}^B$ 

```

Consider an overlapped sample $x_c \in D^c$. In round n , party k first presents local prediction through local best model $f_{b(n)}^k$ obtained from local training. Denote the output logit of $f_{b(n)}^k$ as z_c^k :

$$z_c^k = f_{b(n)}^k(x_c). \quad (4)$$

The perturbed ensemble logit $z_c'^t$ is a linear combination of the perturbed local output logits of $f_{b(n-1)}^k$. It can be expressed as:

$$z_c'^k = f_{b(n)}^k(x_c) + \mathcal{N}(0, \sigma^2) \quad (5)$$

$$z_c'^t = w z_c'^A + (1 - w) z_c'^B, \quad (6)$$

where w is the ensemble weight and σ is the variance of Gaussian noise. Then the soft target distribution \mathbf{p}_c^t of x_c can be defined as

$$\mathbf{p}_c^t = \sigma_T(z_c'^t) \quad (7)$$

$$\sigma_T(z_c'^t) = \frac{\exp(z_c'^t / T_{ED})}{\sum_{i=1}^n \exp(z_{ci}^t / T_{ED})}, \quad (8)$$

where σ_T is the general softmax function tuned by the ensemble temperature T_{ED} . The standard softmax function can be viewed as a special case of the general softmax function with $T = 1$.

F. Local Training

We use an offline training scheme due to efficiency and privacy concerns. Offline training can largely reduce communication needs and therefore limit the exposure of private information to the server. During the local training, each party trains its local model f^k parameterized by $f^k(\theta)$. θ^* is optimized by minimizing the training objective function \mathcal{L}_{train} :

$$\theta^* = \operatorname{argmin}_{\theta} \mathcal{L}_{train}. \quad (9)$$

In this stage, we leverage three kinds of knowledge to enhance the model performance: direct knowledge from private labeled data, implicit knowledge from the best local model in the previous round and explicit knowledge from the other party. The direct knowledge is learned through the cross-entropy loss \mathcal{L}_{CE} computed by the local model outputs \mathbf{p}^k and the ground truth labels y . Our overall training objective function \mathcal{L}_{train} is a weighted combination of three loss terms. Combining equation 2 and 3, \mathcal{L}_{train} can be written as:

$$\mathcal{L}_{train} = \mathcal{L}_{SD} + \beta \mathcal{L}_{KD} + \gamma \mathcal{L}_{CE}, \quad (10)$$

where γ and β are the corresponding trade-off weights. We simply remain these weights to be unchanged during training. More experimental details are introduced in section IV-C.

Note that local models have converged on their private data during the pretraining stage, which means that the initial best local models already contain valuable information. Therefore, both the self teacher knowledge and ensemble teacher knowledge are informative from the first round.

G. Inference

In the inference phase, given a sample x , a party first checks whether the test sample x is aligned with the other party. If the sample is aligned between both parties, two parties first infer locally through their obtained best local models and then ensemble the local predictions to give a joint prediction as the final result. Otherwise, the party directly returns the prediction of its local model as the final result.

H. Communication Analysis of FedPDD

We analyze the communication cost of FedPDD in this section. Assume that the communication cost of updating or downloading a record (i.e. logit of the local prediction for m class classification) from the server once is m . The total number of communication rounds until local models converge is n . Then the overall communication cost is $2mn|D^c|$. We can see that overall communication cost relates to three factors: the number of communication rounds, the amount of overlapped data involved in training and the size of updates. Our offline training strategy only has $O(1)$ communication rounds as shown in section IV-E1 and the size of model output m is also $O(1)$ for each record. Therefore, the communication cost of FedPDD is $O(|D^c|)$. In contrast, the online training strategy requires $O(100)$ communication rounds and transmitting model information which usually involves thousands of parameters will cost more communication overload.

I. Privacy Analysis of FedPDD

In this section, we follow the prior works [30], [31] and give the privacy analysis of FedPDD.

Definition 1: (Differential Privacy [31]). A randomized mechanism $M : \mathcal{X} \rightarrow \mathcal{Y}$ is (ϵ, δ) -DP if for every pair of datasets $X, X' \in \mathcal{X}$ that only differ in one sample, and every possible output $T \in \mathcal{Y}$. The following inequality holds:

$$\mathbb{P}[M(x) \in E] \leq e^\epsilon \mathbb{P}[M(x') \in E] + \delta, \quad (11)$$

where $\epsilon, \delta \geq 0$ are privacy loss parameters.

Definition 2: (l_2 -sensitivity). The l_2 -sensitivity of a function $f : \mathcal{X} \rightarrow \mathbb{R}^d$ is

$$\Delta_2(f) = \max_{X, X' \in \mathcal{X}} \|f(X) - f(X')\|_2. \quad (12)$$

Definition 3: (Analytic Gaussian Mechanism [30]). Let $f : \mathcal{X} \rightarrow \mathbb{R}^d$ be a function with global L_2 sensitivity Δ . For any $\epsilon \geq 0$ and $\delta \in [0, 1]$, the Gaussian output perturbation mechanism $M(x) = f(x) + Z$ with $Z \sim \mathcal{N}(0, \sigma^2 I)$ is (ϵ, δ) -DP if and only if

$$\Phi\left(\frac{\Delta}{2\sigma} - \frac{\epsilon\sigma}{\Delta}\right) - e^\epsilon \Phi\left(-\frac{\Delta}{2\sigma} - \frac{\epsilon\sigma}{\Delta}\right) \leq \delta, \quad (13)$$

where Φ is the CDF function of $\mathcal{N}(0, 1)$.

Definition 4: (Composition of DP Algorithms [32], [33]). Suppose $M = (M_1, M_2, \dots, M_k)$ is a sequence of algorithms, where M_i is (ϵ_i, δ_i) -DP, and the M_i 's are potentially chosen sequentially and adaptively. Then M is $(\sum_{i=1}^k \epsilon, \sum_{i=1}^k \delta)$ -DP.

For a meaningful privacy guarantee, we have $\delta = o(\frac{1}{n})$, where n is the size of the dataset. By fixing privacy budget ϵ , we calibrate the noise with proper σ given by definition 3. Therefore, our method preserves (ϵ, δ) -DP.

IV. EXPERIMENTS

In this section, we evaluate our proposed FedPDD on two public real-world datasets. We aim to answer the following two questions through our experiments:

- Q1: How well does FedPDD perform compared to the SOTA baselines on the benchmark datasets?
- Q2: How does differential privacy influence the performance of FedPDD?

A. Datasets

The statistics of two benchmark public datasets are summarized in Table III.

HetRec-MovieLens Dataset This dataset ¹ is a heterogeneous dataset for movie recommendation with 86W ratings. It is an extension of the MovieLens10M dataset, which contains personal ratings and tags about movies. The movies are linked to Internet Movie Database (IMDb) and RottenTomatoes movie review systems, greatly extending the feature space. The dataset is converted as a classification task by taking instances with a rating lower than 3 as negative instances, otherwise positive. In our setting, party A holds 21 features mainly coming from MovieLens, including user ID, movie ID, tags

¹<https://grouplens.org/datasets/hetrec-2011/>

TABLE II
MAIN RESULT (OVERLAPPED DATA RATIO $\alpha = 0.1$)

Model settings	HetRec-MovieLens			Criteo		
	Local A	Local B	Joint Prediction	Local A	Local B	Joint Prediction
DeepFM [34]	79.65%	80.36%	-	72.74%	72.93%	-
Ensemble [35]	-	-	80.53%	-	-	73.08%
FTL [4]	-	-	78.94%	-	-	73.90%
PFML [36]	80.26%	80.62%	81.32%	74.28%	74.26%	75.53%
FedKD [25]	81.70%	81.60%	81.71%	74.17%	74.28%	75.41%
FedPDD (ours)	82.91%	82.88%	82.92%	75.36%	75.07%	76.68%

TABLE III
STATISTICS OF BENCHMARK DATASETS

Name	#users	#movies	#feat. fields	#ratings
HetRec-MovieLens	2,113	10,197	33	855,598
Criteo	-	-	39	1000,000

and etc., while party B holds 13 features from RottenTomatoes, including movie information, critics score, Rotten data and etc.

Criteo The Criteo dataset ² is generated from the original criteo dataset by randomly sampling 1 million instances. The task is to predict the ad click-through rate (CTR). It consists of 13 numerical features as well as 26 categorical features. We randomly split both numerical and categorical features into two parts with 19, 20 features respectively for two parties.

B. Baselines

We compare FedPDD with the following baselines:

Locally trained DeepFM: DeepFM [34] can handle high-order feature interaction of user embeddings and item embeddings in a centralized manner. Each party locally trains a DeepFM model on its own private dataset, and can not make use of the private features from the other party.

Ensemble of DeepFM: Each party first locally trains a DeepFM model, then the local model outputs are aggregated to obtain the final joint prediction. We average the predictions from local models for model aggregation.

FTL baseline: The FTL [4] approach maps the samples from heterogeneous feature spaces from two parties into a common latent space. Then, the two feature representations are concatenated and input to a classifier for prediction. We train a unique feature extractor for feature extraction in each party and then collaboratively train a classifier for label prediction.

PFML baseline: The PFML [36] approach integrates deep mutual learning [37] into the local update process in each party to improve the performance of both the global model and the personalized local models.

FedKD baseline: The FedKD [25] algorithm uses ensemble distillation for robust model fusion on heterogeneous local models. To transfer knowledge, an unlabeled dataset is used to sample data for all participants to compute logits and distill knowledge. We adopt this approach to our setting. In each communication round, all parties perform ensemble distillation, in which the local model parameters are evaluated on the aligned samples to generate logit outputs that are used to train each student model.

C. Experiment Settings

For experiments, we randomly sample 80% data as a training dataset and the rest for testing. Then we randomly split the dataset into two local datasets according to the overlapped data ratio. For the test dataset, we assume that all the data is aligned and shared by two parties.

We adopt DeepFM [34] as the backbone, which is proposed to handle sophisticated feature interactions behind user behaviors for recommendation tasks. For training, we follow our two-stage training process as described in section III. The model is optimized by Adam [38]. We set the learning rate for both local models to 0.001, the weight decay to 0.0001, the number of communication rounds to 5, and the batch size to 1024. The training process stops when the training achieves the maximum communication round. For experiments on the HetRec-MovieLens dataset, we set the temperature T to 30, the trade-off loss weights β, γ to 10, and the ensemble weight w to 0.5. For experiments on the Criteo dataset, we set the temperature T to 30, the trade-off loss weights β, γ to 3, and the ensemble weight w to 0.5. For all the ablation settings, we conduct experiments three times and report the average.

We use accuracy as the metric to evaluate our experiment results. The closer the value of accuracy to 1, the better the performance of prediction is.

D. Main Results

Comparison with local training. To demonstrate the effectiveness of our proposed method, we first compare it with the local fine-tuned baselines (i.e. Local A and Local B in Table II). From Table II, it is observed that 1) both local models benefit significantly from our approach, with an increment of 3.26%/2.44% on the HetRec-MovieLens and 2.62%/2.14% on Criteo. 2) The joint prediction further brings extra performance gain to the local models by around 0.04% and 1.61%, respectively. These results show that our approach successfully transfers knowledge between two parties and therefore improves the performance of their local models.

Comparison with SOTAs. FTL-based methods [4] leverage the overlapped data to make a joint prediction. Therefore, we only compare it with the joint prediction of FedPDD. From Table II, we can see that the joint prediction of FedPDD is better than FTL by 3.98% on HetRec-MovieLens and 2.78% on Criteo. These results show that our proposed framework outperforms the FTL-based methods in the situation where overlapped data is limited.

²<https://labs.criteo.com/category/dataset/>

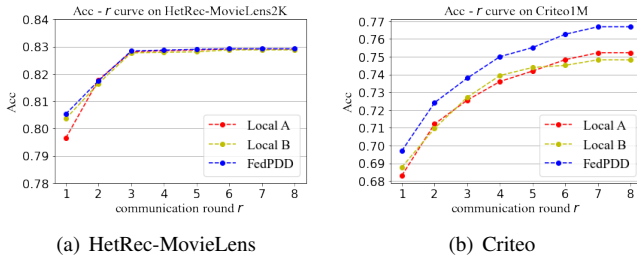


Fig. 3. The relationship between communication round r and performance of FedPDD during training

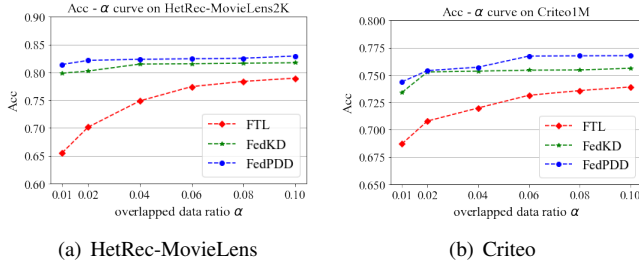


Fig. 4. The comparison between FedPDD and FTL baseline when the overlapped data ratio α decreases

We also compare FedPDD with two knowledge distillation-based federated learning strategies PFML [36] and FedKD [25]. From Table II, we can find that the local model performance of FedPDD outperforms the PFML baseline by 2.46% and 1.92% on two datasets on average. Besides, our FedPDD outperforms FedKD by an additional 1.25% and 0.99% on two datasets on average. This indicates that our double distillation method can generate better teacher logits from not only the ensemble of cross-party local models but also the previous local models of the same party, thereby effectively enhancing the local model performance.

For joint prediction on aligned test samples, the predictions of both local models are averaged as the final result. Therefore, the performance of federated joint prediction mainly depends on the performance of local models. From Table II, we can observe that FedPDD outperforms FedKD by 1.21% and 1.27% on two datasets, respectively, and outperforms PFML by 1.60% and 1.15% on two datasets, respectively. This is reasonable as the local models trained by FedPDD achieve higher accuracy than FedKD and PFML on both datasets. Meanwhile, more accurate joint predictions can in turn transfer more knowledge to both local models.

E. Hyper-parameter Tuning

1) *Effect of communication round r* : From Figure 3, we can find that FedPDD converges after 5 rounds on the HetRec-MovieLens dataset and 7 rounds on the Criteo dataset, which demonstrates that FedPDD only requires a few times of communication need between two parties during training.

2) *Effect of overlapped data ratio α* : As we mentioned previously, the major challenge of the multi-view federated learning problem is that the overlapped data is often limited. We adjust the overlapped data ratio α from 0.1 to 0.01 to test the effectiveness of FedPDD. The experimental results are shown in Figure 4. We can observe that when α decreases to

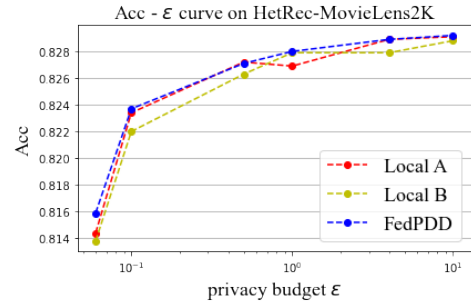


Fig. 5. The impact of DP parameter ϵ on model performance on HetRec-MovieLens dataset.

TABLE IV
SENSITIVITY TO T ON HETEREC-MOVIELENS2K

T	1	10	20	30	40	50
FedPDD	0.8258	0.8260	0.8270	0.8291	0.8270	0.8267
Local A	0.8254	0.8252	0.8260	0.8288	0.8259	0.8256
Local B	0.8268	0.8261	0.8270	0.8292	0.8270	0.8267

0.01, the performance of the FTL approach drops significantly by 13.42% while FedPDD still remains above 0.8 on the HetRec-MovieLens dataset. On the Criteo dataset, FedPDD almost remains the same while the FTL baseline drops around 5.20% when α decreases. These results demonstrate the effectiveness of FedPDD in our setting.

3) *Effect of temperature T* : In knowledge distillation, the temperature is used to soften the probability output, leading the students to pay more attention to the small number of logits [20]. In this section, we conduct experiments to find out the influence of temperature on our model. We let $T = T_{SD} = T_{ED}$. In Table II, we set the temperature to 30 to highlight the best performance of FedPDD on two benchmark datasets. Here we fine-tune the temperature T over a large range from 1 to 50 in the online scheme. The results in Table IV and V show that the careful selection of temperature can bring a little performance enhancement on local models.

4) *Effect of differential privacy budget ϵ* : In Figure 5, we demonstrate the effect of differential privacy on the HetRec-MovieLens dataset. We change the ϵ from 0.05 to 10 to explore the change in local model performance and federated model performance. It can be observed that the accuracy drops only by 1.53% for local models in FedPDD and 1.30% for joint prediction of FedPDD, respectively. The performance of local models outperforms locally trained baselines when $\epsilon > 0.05$.

V. CONCLUSION

In this paper, we propose a novel cross-silo federated recommendation framework FedPDD. We design a double distillation strategy that leverages knowledge not only from the ensemble of local models but also from previous local models to efficiently improve the model performance. Besides, FedPDD largely reduces communication needs and privacy leakage risk by utilizing an offline training strategy and only transmitting model output during training. Additionally, differential privacy is introduced to protect the communication process with a higher level of privacy protection. Experimental

TABLE V
SENSITIVITY TO T ON CRITEO

T	1	10	20	30	40	50
FedPDD	0.7520	0.7508	0.7521	0.7522	0.7514	0.7507
Local A	0.7473	0.7487	0.7489	0.7482	0.7481	0.7480
Local B	0.7657	0.7664	0.7664	0.7668	0.7660	0.7663

results demonstrate that our approach can effectively exploit both implicit knowledge and explicit knowledge and thereby enhance the performances of both local and overall joint prediction tasks. Moreover, our framework can also be adopted to learn and predict financial risks associated with various internet finance platforms with heterogeneous information features and strong privacy-preserving needs.

VI. ACKNOWLEDGMENT

The authors gratefully acknowledge funding from Guangdong Province Focus Research Project (Grant Number: 2019KZDZX2014), Guangdong Province Research Fund (Grant Number: 2019QN01X277), National Natural Science Foundation of China (Grant Numbers: 71971106, 72001099), and Shenzhen Humanities & Social Sciences Key Research Bases. We would like to show our gratitude to Guangneng Hu, Ce Ju, Ben Tan and Prof. Qiang Yang for their advice on the earlier manuscript and we thank all the reviewers for valuable comments.

REFERENCES

- [1] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Computing Surveys (CSUR)*, vol. 52, no. 1, pp. 1–38, 2019.
- [2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.
- [3] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.
- [4] Y. Liu, Y. Kang, C. Xing, T. Chen, and Q. Yang, "A secure federated transfer learning framework," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 70–82, 2020.
- [5] F. Siwei and Y. Han, "Multi-participant multi-class vertical federated learning," *arXiv preprint arXiv:2001.11154*, 2020.
- [6] Y. Kang, Y. Liu, and T. Chen, "Fedmvt: Semi-supervised vertical federated learning with multiview training," *arXiv preprint arXiv:2008.10838*, 2020.
- [7] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *NeurIPS*, vol. 32, 2019.
- [8] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients-how easy is it to break privacy in federated learning?" in *NeurIPS*, vol. 33, 2020, pp. 16937–16947.
- [9] J. Sun, A. Li, B. Wang, H. Yang, H. Li, and Y. Chen, "Soteria: Provable defense against privacy leakage in federated learning from representation perspective," in *CVPR*, 2021, pp. 9311–9319.
- [10] 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.
- [11] 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.
- [12] S. Ying, "Shared mf: A privacy-preserving recommendation system," *arXiv preprint arXiv:2008.07759*, 2020.
- [13] L. Yang, B. Tan, B. Liu, V. W. Zheng, K. Chen, and Q. Yang, "Practical and secure federated recommendation with personalized masks," *arXiv preprint arXiv:2109.02464*, 2021.

- [14] D. Chai, L. Wang, K. Chen, and Q. Yang, "Secure federated matrix factorization," *IEEE Intelligent Systems*, vol. 36, no. 5, pp. 11–20, 2020.
- [15] E. Yang, Y. Huang, F. Liang, W. Pan, and Z. Ming, "Fcmf: Federated collective matrix factorization for heterogeneous collaborative filtering," *Knowledge-Based Systems*, vol. 220, p. 106946, 2021.
- [16] M. E. Eren, L. E. Richards, M. Bhattarai, R. Yus, C. Nicholas, and B. S. Alexandrov, "Fedsplit: One-shot federated recommendation system based on non-negative joint matrix factorization and knowledge distillation," *arXiv preprint arXiv:2205.02359*, 2022.
- [17] L. Minto, M. Haller, B. Livshits, and H. Haddadi, "Stronger privacy for federated collaborative filtering with implicit feedback," in *Fifteenth ACM Conference on Recommender Systems*, 2021, pp. 342–350.
- [18] M. Ammad-Ud-Din, E. Ivannikova, S. A. Khan, W. Oyomno, Q. Fu, K. E. Tan, and A. Flanagan, "Federated collaborative filtering for privacy-preserving personalized recommendation system," *arXiv preprint arXiv:1901.09888*, 2019.
- [19] W. Li, Q. Xia, J. Deng, H. Cheng, J. Liu, K. Xue, Y. Cheng, and S.-T. Xia, "Semi-supervised cross-silo advertising with partial knowledge transfer," *arXiv preprint arXiv:2205.15987*, 2022.
- [20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [21] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," in *Proceedings of the 23rd ACM SIGKDD*, 2017, pp. 1285–1294.
- [22] D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," *arXiv preprint arXiv:1910.03581*, 2019.
- [23] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *NeurIPS*, vol. 33, pp. 2351–2363, 2020.
- [24] C. He, M. Annavaram, and S. Avestimehr, "Group knowledge transfer: Federated learning of large cnns at the edge," *NeurIPS*, vol. 33, pp. 14068–14080, 2020.
- [25] X. Gong, A. Sharma, S. Karanam, Z. Wu, T. Chen, D. Doermann, and A. Innanje, "Preserving privacy in federated learning with ensemble cross-domain knowledge distillation," in *AAAI*, vol. 36, no. 11, 2022, pp. 11891–11899.
- [26] Q. Li, B. He, and D. Song, "Practical one-shot federated learning for cross-silo setting," *arXiv preprint arXiv:2010.01017*, vol. 1, no. 3, 2020.
- [27] J.-H. Ahn, O. Simeone, and J. Kang, "Cooperative learning via federated distillation over fading channels," in *IEEE ICASSP*. IEEE, 2020, pp. 8856–8860.
- [28] Z. Allen-Zhu and Y. Li, "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning," *arXiv preprint arXiv:2012.09816*, 2020.
- [29] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *arXiv preprint arXiv:1610.05755*, 2016.
- [30] B. Balle and Y. Wang, "Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *ICML*. PMLR, 2018, pp. 394–403.
- [31] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.
- [32] 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.
- [33] C. Dwork and J. Lei, "Differential privacy and robust statistics," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 2009, pp. 371–380.
- [34] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "Deepfm: a factorization-machine based neural network for ctr prediction," *arXiv preprint arXiv:1703.04247*, 2017.
- [35] T. G. Dietterich *et al.*, "Ensemble learning," *The handbook of brain theory and neural networks*, vol. 2, no. 1, pp. 110–125, 2002.
- [36] R. Yang, J. Tian, and Y. Zhang, "Regularized mutual learning for personalized federated learning," in *Proceedings of The 13th Asian Conference on Machine Learning*, vol. 157. PMLR, 17–19 Nov 2021, pp. 1521–1536.
- [37] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *CVPR*, 2018.
- [38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.