Privacy-Preserving Distributed Link Predictions Among Peers in Online Classrooms Using Federated Learning

Anurata Prabha Hridi
NC State University
aphridi@ncsu.edu

Muntasir Hoq
NC State University
mhoq@ncsu.edu

Zhikai Gao
NC State University
zgao9@ncsu.edu

Collin Lynch
NC State University
cflynch@ncsu.edu

Rajeev Sahay
UC San Diego
r2sahay@ucsd.edu

Seyyedali Hosseinalipour
University at Buffalo–SUNY
alipour@buffalo.edu

Bita Akram
NC State University
bakram@ncsu.edu

ABSTRACT

Social interactions among classroom peers, represented as social learning networks (SLNs), play a crucial role in enhancing learning outcomes. While SLN analysis has recently garnered attention, most existing approaches rely on centralized training, where data is aggregated and processed on a local/cloud server with direct access to raw data. However, in real-world educational settings, such direct access across multiple classrooms is often restricted due to privacy concerns. Furthermore, training models on isolated classroom data prevents the identification of common interaction patterns that exist across multiple classrooms, thereby limiting model performance. To address these challenges, we propose one of the first frameworks that integrates Federated Learning (FL), a distributed and collaborative machine learning (ML) paradigm, with SLNs derived from students' interactions in multiple classrooms' online forums to predict future link formations (i.e., interactions) among students. By leveraging FL, our approach enables collaborative model training across multiple classrooms while preserving data privacy, as it eliminates the need for raw data centralization. Recognizing that each classroom may exhibit unique student interaction dynamics, we further employ model personalization techniques to adapt the FL model to individual classroom characteristics. Our results demonstrate the effectiveness of our approach in capturing both shared and classroom-specific representations of student interactions in SLNs. Additionally, we utilize explainable AI (XAI) techniques to interpret model predictions, identifying key factors that influence link formation across different classrooms. These insights unveil the drivers of social learning interactions within a privacy-preserving, collaborative, and distributed ML framework—an aspect that has not been explored before.

Keywords

Social Learning, Social Network Analysis, Federated Learning, Student Interaction.

Anurata Hridi, Muntasir Hoq, Zhikai Gao, Collin Lynch, Rajeev Sahay, Seyyedali Hosseinalipour, and Bita Akram. Privacy-Preserving Distributed Link Predictions Among Peers in Online Classrooms Using Federated Learning. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 276–288. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870179

1. INTRODUCTION

Students' social interactions, a founding factor of learning according to the *socio-constructivist theory of learning* [38], hold special importance in an online/hybrid learning environment. Unlike traditional settings, where peer-to-peer and peer-to-instructor interactions occur naturally, these environments often rely on limited and deliberately designed interaction channels [59]. This shift in interaction dynamics, shaped by the widespread adoption of modern learning platforms such as massive open online courses (MOOCs) and learning management systems (LMSs), not only alters the way students engage but also provides a structured setting for systematically examining these interactions [51]. Understanding how this shift affects the nature and quality of interactions is crucial, as these interactions play a direct role in shaping students' learning experiences and outcomes. To this end, researchers have drawn parallels between student interactions in online learning environments and user connections in social networks, representing them through *graph-based data structures* known as social learning networks (SLNs) [16, 29]. Analyzing SLNs uncovers valuable insights into interaction patterns and their correlation with student learning, enabling the development of personalized instructional strategies that foster engagement, enhance learning outcomes, and promote academic success [18].

1.1 Background and Motivation

Historically, artificial intelligence/machine learning (AI/ML) techniques have been utilized to analyze SLNs and understand the complexities of social interactions among students [12, 56]. Despite significant progress, the existing works in this domain share a common limitation: *a reliance on centralized learning architectures, where ML models are trained on local/cloud servers with direct access to all training data [19].* Such direct data access typically occurs under two scenarios: (i) data from multiple classrooms are aggregated into a centralized storage system or (ii) model training is performed using data from a single classroom. Nevertheless, while the former approach enhances model performance through access to diverse and high-quality data, it is often restricted by privacy regulations – such as family educational rights and privacy act (FERPA) [1] – that prohibit the transfer of student data across networks due to the risk of data exposure. Also, the latter approach, although privacy-compliant, generally leads to suboptimal model performance due to the limited diversity and volume of available training data.

Recently, federated learning (FL) has emerged as a leading distributed ML technique [47, 37], offering a viable solution to the limitation discussed above. In particular, FL enables collaborative model training across distributed data-collecting entities (referred to as clients) through an iterative process that involves two primary steps. *(Step 1) Local Training:* each client independently trains a local model on its dataset and shares only the model parameters (rather than raw data) with a central server. *(Step 2) Global Aggregation and Broadcast:* the server aggregates the received local model parameters from clients — typically through weighted averaging — to construct a global model, which is subsequently broadcast back to the clients, allowing them to synchronize their local models and commence the next round of local training. By eliminating the need to transfer raw data across networks, FL inherently preserves privacy while facilitating distributed ML across multiple clients. This privacy-preserving property allows FL to enhance both the performance and fairness of model training compared to centralized methods, which rely on isolated data from individual clients [7, 50, 22, 35, 39]. Given these advantages, FL has recently gained significant attention in various domains, including educational applications [21, 20, 55, 14, 24] and social networks [63, 9, 57, 6, 43, 39]. Despite these advances, *the application of FL for harnessing data from multiple classrooms to study SLNs through the lens of predicting student interactions (referred to as link prediction) remains largely unexplored.* Addressing this gap and investigating the performance of FL in this context serves as the primary motivation for this paper.

1.2 Overview and Summary of Contributions

Our key contributions can be summarized as follows:

- To our knowledge, this is the first study that explores the potential of FL, along with its advanced variations, to predict the *occurrence* (or absence) of an *interaction* between students (represented as an *edge* connecting two nodes in an SLN) based on their interaction patterns. Consequently, this work establishes the framework for privacy-preserving distributed link prediction in SLNs.
- To conduct our analysis, we use forum interaction data from five distinct classrooms focused on science and humanities, from which we extract a set of graph-theoretic features. This process reveals a unique challenge in studying FL performance for SLN analysis: *while student interaction behaviors may share commonalities across classrooms, each class — due to its unique subject matter and structure — often exhibits distinct interaction patterns.* Nevertheless, vanilla FL methods, which train a single/universal global model across all classrooms, fail to capture these distinct interaction patterns.
- To address this challenge, we investigate a set of *model personalization strategies* to adapt the global FL model to the data of each classroom. Our findings demonstrate the superior performance of fine-tuned/personalized FL models over both centralized and vanilla FL approaches, achieving notable improvements in both *classification accuracy* and *prediction fairness* metrics. These results highlight the effectiveness of personalization in accounting for classroom-specific differences and pave the way for future work on further personalization of FL models based on student demographics, promoting more inclusive and representative educational environments.
- We then investigate an understudied aspect: *whether the importance of graph-theoretic features varies between science and humanities classrooms due to their distinct interaction patterns.* To this end, we incorporate explainable AI (XAI) techniques — specifically, SHapley Additive exPlanations (SHAP) — into the analysis of fine-tuned/personalized FL models. This examination provides nuanced insights into the varying influence of graph-theoretic features on student interaction predictions across different classroom types, enabling the development of tailored recommendations for educators and institutions to foster meaningful student connections based on course subject and classroom context.

2. RELATED WORK

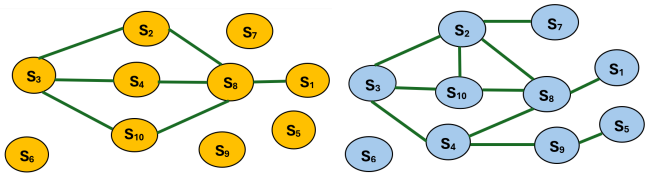
In the following, we first provide an overview of social learning and the use of AI/ML techniques in this domain, and then we delve deeper into the link prediction problem in SLNs.

2.1 Social Learning & the Impact of AI/ML

Social learning in traditional educational settings emphasizes peer interactions, which has been shown to result in enhanced cognitive skills [10]. In this process, students acquire knowledge and develop competencies through social interactions [38]. In recent years, the rise of online learning platforms such as Coursera and edX has significantly expanded the reach of social learning, leading to the formation of SLNs. These platforms support virtual communities and discussion forums, enabling learners to connect, collaborate, and share resources across diverse geographical locations. Also, the growth of online education, particularly accelerated by the COVID-19 pandemic, has led to the widespread adoption of digital learning platforms, with nearly 84% of undergraduate students in the US engaging in at least one online course [32]. These online/digital platforms generate extensive data through student activity logs and peer interactions, which can be modeled as SLNs [61], where *nodes* represent students and *edges* denote interactions. With the advancement of AI/ML techniques, analyzing student activity logs from these platforms has been shown to uncover meaningful interaction behaviors, creating opportunities to enhance the learning environment [5]. By capturing diverse patterns within data, ML models can address learners' individual needs, making SLNs a powerful tool for increasing engagement and offering personalized learning recommendations [29].

2.2 Link Prediction in SLN & Potential of FL

Link prediction has been studied in the context of both SLNs and online social networks (OSNs) — which bear a strong resemblance to SLNs — for decades [31, 41, 16, 48], primarily leveraging neighborhood and path-based features derived from graph topologies [49, 36, 8]. In this domain, various ML methods have been used to improve the accuracy of link predictions [3]. Nevertheless, the existing approaches in this domain rely on *centralized* model training procedures, which suffer from either limited data exposure [27] when trained on isolated data of an individual OSN/SLN or face privacy concerns when the model is trained on the pool of data collected from multiple OSNs/SLNs [43]. This hinders the effectiveness of models derived from OSNs/SLNs in promoting collaboration and personalized learning experiences.



(a) At time $t-1$ (b) At time t

Figure 1: SLNs illustrating students' connections formed at two subsequent time instants, where, S_1, S_2, etc., represent individual students. (a) Out of 10 students, four stay unconnected by the end of $t-1$, and (b) besides existing connections, new links were formed by the end of t that can be attributed to course structure and student interactions.

FL offers a solution by enabling distributed model training across multiple clients while keeping data localized, thereby ensuring privacy and addressing data limitations [47, 42]. Subsequently, by aggregating knowledge from various classrooms, each considered to be a client in the FL setting, FL enhances the generalizability and accuracy of the models trained on multiple classrooms' data. While recent studies have examined the applications of FL in educational settings [20, 55, 14, 58, 33], its potential for link prediction in SLNs remains unexplored — a gap that this work aims to fill.

3. METHODOLOGY

In this section, we first describe the SLN structure and the datasets involved in this study (Sec. 3.1). We then elaborate on the centralized learning (Sec. 3.2) and the FL architecture (Sec. 3.3). Afterward, we describe three personalized FL (pFL) methods that entail tailoring the FL model to individual classrooms (Sec. 3.4). Finally, we delve into the fairness measures and XAI techniques (Secs. 3.5 and 3.6).

3.1 Social Learning Network (SLN) Structure

We begin by describing how SLNs are constructed based on student pairs and their shared contributions to various online activities, including discussions, resource sharing, and collaborative projects. Figure 1 shows the temporal progression of a sample SLN formed from student interaction data, where nodes represent students and edges represent connections or interactions between them in the SLN. Let $t=0$ denote the initial time instant corresponding to the start of a course, and let $\mathcal{S}=\{S_1, S_2, \cdots\}$ represent the set of enrolled students. At any given time instant t, a link (u,v) is established between two students S_u and S_v, where $S_u, S_v \in \mathcal{S}$, if they have interacted prior to time t. Such interactions include initiating a common thread or adding follow-up posts in the same thread [52] and are depicted as bidirectional edges in the SLN to reflect the reciprocal nature of the communication. To model the evolving SLN, we construct a time-varying adjacency matrix $\mathbf{A}^{(t)}=[a^{(t)}_{u,v}]_{S_u,S_v\in\mathcal{S}}$, where each entry $a^{(t)}_{u,v}\in\{0,1\}$ is a binary indicator denoting the presence ($a^{(t)}_{u,v}=1$) or absence ($a^{(t)}_{u,v}=0$) of an edge between students S_u and S_v at time t. Accordingly, the SLN at time instant t can be represented by the time-dependent graph structure $\mathcal{G}^{(t)}=(\mathcal{S},\mathcal{E}^{(t)})$, where $\mathcal{S}$ denotes the set of nodes (students) and $\mathcal{E}^{(t)}$ is the set of edges. A pair (u,v) belongs to $\mathcal{E}^{(t)}$ (i.e., $(u,v)\in\mathcal{E}^{(t)}$) if and only if $a^{(t)}_{u,v}=1$.

3.1.1 Datasets

We consider four MOOC datasets adopted by Sahay et al. [52] and one real-world classroom data of Fall 2021 from a North American university. These datasets encompass a broad spectrum of subjects: *Algorithms: Design and Analysis, Part 1 (algo), English Composition I (comp), Shakespeare in Community (shake), Machine Learning (ml)* and *Software Development Fundamentals (csc)*. When comparing the datasets based on the number of considered student pairs (i.e, student pairs whose interactions were recorded), *'ml'* is the largest, followed by *'algo'*, *'shake'*, *'comp'*, and *'csc'*, which is the smallest (see Table 1). The relatively small number of links compared to the student pairs indicates sparsity of the SLNs. These datasets cover diverse disciplines, comprising three STEM subjects and two humanities subjects, with no overlap (i.e., no shared enrolled students) in the collected data. In the context of constructing SLN graphs, each dataset includes links representing interactions formed among students only [17] throughout the course duration. It is important to note that, in adherence to standard data anonymization practices, the datasets do not contain user-identifying information or student demographic details.

Table 1: Course details and dataset characteristics.

Course Title	Class -room	Nodes	Links	Student Pairs	Course Type
Algorithms Design Analysis I	algo	2080	5416	36390	STEM & MOOC
English Composition I	comp	1707	2731	19049	Non-STEM & MOOC
Shakespeare in Community	shake	1296	3753	26008	Non-STEM & MOOC
Machine Learning	ml	6446	24481	76526	STEM & MOOC
Software Development Fundamentals	csc	366	585	1910	STEM & Non-MOOC

3.1.2 Temporal Representation of SLNs

The MOOC datasets (i.e., *'algo', 'comp', 'shake', and 'ml'*) are static longitudinal data logs capturing interactions between online student pairs, with links formed cumulatively over a certain period. However, since explicit timeline information was not provided, we imposed a temporal structure by assuming the absence of links among certain pairs at an earlier time, thereby reflecting the potential for those links to form later. To emulate this temporal evolution, we created a modified version of the original datasets representing time step $t-1$ by randomly selecting 20% of the data (i.e., student pairs) and removing any existing links between the corresponding student pairs. The assumption is that the removed links would emerge at time step t, as reflected in the original datasets. Using this setup, we employed the data at time step $t-1$ to predict link formations at time step t. Nonetheless, the *'csc'* dataset contains student interaction data recorded continuously throughout the semester. For this dataset, we captured the first instance of link formation between student pairs. Subsequently, we used the SLN topology at week 10 (i.e., time $t-1$) to predict the SLN topology — i.e., the links among students — in weeks 11, 12, 13, and 14 (i.e., time t). To evaluate our prediction performance, we partitioned the data into training and test sets, allocating 80% of the student pairs to the training set (i.e., 80% of student pairs at time $t-1$) and the remaining 20% to the test set.

3.1.3 Feature Description

To extract the features from data points (i.e., student pairs), given the fact that SLNs resemble graph structures, we employ a set of graph-theoretic features used in various link prediction tasks [31]. In a nutshell, these pairwise features measure the similarity between node pairs based either on the similarity of their neighborhoods in the graph or on the connectivity paths between them [44, 41, 64, 46, 15]. In particular, for each of the five SLNs, we computed six features — enumerated from **(Feature I)** to **(Feature VI)** below — for every learner node pair (u, v), where $u \neq v$, corresponding to nodes/students S_u and S_v. Henceforth, we use Γ_u to denote the set of neighboring nodes of S_u, i.e., the nodes that have a link/edge with S_u. Consequently, $|\Gamma_u|$ represents the *degree* of S_u, i.e., the number of edges connected to node S_u.

(Feature I) Jaccard Coefficient [34]: This metric quantifies the similarity between two nodes S_u and S_v through a ratio, capturing the proportion of their shared neighbors as follows:

$$\mathcal{J}_{uv} = \frac{|\Gamma_u \cap \Gamma_v|}{|\Gamma_u \cup \Gamma_v|}, \tag{1}$$

where the numerator represents the number of common neighbors between nodes S_u and S_v, while the denominator indicates the total number of distinct neighbors across both nodes. A higher Jaccard Coefficient value between two students implies a greater proportion of shared connections between them, suggesting they belong to closer social circles.

(Feature II) Adamic-Adar Index [2]: This metric quantifies the likelihood of a connection between two nodes S_u and S_v through the inverse logarithm of the degrees of their shared neighbors as follows:

$$\mathcal{A}_{uv} = \sum_{S_n \in \Gamma_u \cap \Gamma_v} \frac{1}{\log |\Gamma_n|}, \tag{2}$$

where the summation is taken over all common neighbors of S_u and S_v. In essence, the inverse of $\log |\Gamma_n|$ assigns higher weights to neighbors with fewer connections, implying that two students sharing a rarely active peer as a common neighbor are more likely to be connected than when they share a highly active peer.

(Feature III) Resource Allocation Index [67]: This metric, similar to the Adamic–Adar Index (2), quantifies the likelihood of a connection between two nodes based on the inverse of the degrees of their common neighbors as follows:

$$\mathcal{R}_{uv} = \sum_{S_n \in \Gamma_u \cap \Gamma_v} \frac{1}{|\Gamma_n|}. \tag{3}$$

A higher value of the Resource Allocation Index between two nodes/students suggests that their common neighbors have fewer connections — meaning these neighbors are selective in their interactions and are not connected to the majority of students. This selectivity implies a stronger likelihood of a direct connection forming between the two students.

(Feature IV) Preferential Attachment Score [11]: Unlike previous metrics that focus on node neighborhoods, this metric directly assesses the similarity between two nodes S_u and S_v based on their degrees and is defined as follows:

$$\mathcal{P}_{uv} = |\Gamma_u| \cdot |\Gamma_v|. \tag{4}$$

A higher Preferential Attachment Score between two students/nodes indicates a greater likelihood of a link forming between them if one does not already exist. In the context of classrooms, this suggests that students with many existing connections are more likely to develop additional connections.

(Feature V) Cosine Similarity [54]: Also known as the Salton Index, the Cosine Similarity measures the similarity between the two nodes S_u and S_v by computing a value that resembles the cosine of the angle between two vectors as follows:

$$\mathcal{C}_{uv} = \frac{|\Gamma_u \cap \Gamma_v|}{\sqrt{|\Gamma_u| \cdot |\Gamma_v|}}, \tag{5}$$

where the numerator represents the number of common neighbors between nodes/students S_u and S_v, while the denominator represents the geometric mean of their individual degrees. Unlike the Jaccard Coefficient (1), which considers the ratio of shared to total neighbors, the Cosine Similarity focuses on the alignment of the nodes' neighborhoods. A higher Cosine Similarity score between a student pair suggests that their neighborhoods are more aligned, implying stronger similarity and a higher likelihood of connectivity between them.

(Feature VI) Dice Similarity [23]: While the Jaccard Coefficient (1) excludes duplicates when counting the neighbors of two nodes, this metric counts such duplicates, defined as

$$\mathcal{D}_{uv} = \frac{2 \cdot |\Gamma_u \cap \Gamma_v|}{|\Gamma_u| + |\Gamma_v|}, \tag{6}$$

where the numerator represents twice the number of common neighbors between nodes S_u and S_v, and the denominator reflects the total number of neighbors of both nodes. By doubling the weight of shared neighbors, this metric assigns greater importance to common connections compared to the Cosine Similarity (5). A higher Dice Similarity between a student pair indicates a significant overlap in their social circles, suggesting a stronger likelihood of forming a connection.

3.2 Centralized Learning Architecture

Using the above features for every student pair in each of the five SLNs of our interest described in Table 1, we developed a centralized model as a baseline for evaluating the performance of our FL models in the link prediction task. A centralized model refers to a single model trained on the complete set of available data, which is trained and tested using the aggregated data from five distinct classrooms. It is important to note that *this centralized model compromises the privacy of local datasets*, as data from individual classrooms are shared during the training process. Our central model is a convolutional neural network (CNN), following the architecture used in [52], trained using the stochastic gradient descent (SGD) method. To optimize its performance, we conducted hyperparameter tuning with the following configurations: the learning rate choices of {0.1, 0.01, 0.001}, mini-batch size choices of {64, 128, 256, 512, 1024}, and epoch counts of {50, 100, 200}. We monitored for overfitting by comparing the training and validation losses and used a 5-fold cross-validation to tune the hyperparameters.

3.3 Federated Learning (FL) Architecture

Our (vanilla) FL method is built upon the FedAvg architecture [47], utilizing the same CNN architecture employed in the centralized model training. In this setup, each SLN is assumed to reside in a separate storage space with local computing capabilities (e.g., a local server, hereafter referred to as a *client*). A central server — either a cloud server or another local server — coordinates the distributed model training process. In a nutshell, as depicted in Figure 2, the

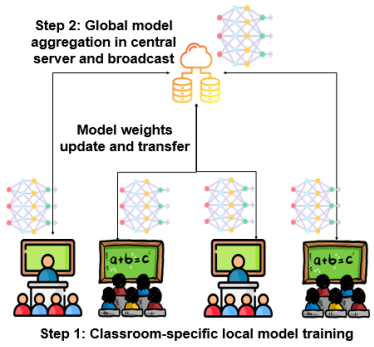


Figure 2: A schematic of FL's distributed training architecture involving the data of multiple classrooms.

training procedure consists of two main phases: *local model training* and *global model aggregation*. During local training, each client first synchronizes its local model with the latest model received from the central server, independently trains its local model on its dataset, and subsequently sends the model parameters to the central server. The central server then performs a global model update by aggregating the received parameters using weighted averaging, thereby forming an updated global model that is redistributed to the clients for the next training round. In the following, we formally describe these processes.

Let $\mathbf{w}^{(k)}$ denote the global model parameters maintained by the central server at the end of the k-th global aggregation round. These parameters are broadcast to clients at the beginning of the subsequent round $k+1$. The model training process initiates at $k=0$, with the central server generating the initial global model $\mathbf{w}^{(0)}$ (e.g., through random initialization) and distributing it to all clients. We next describe the procedure through which clients utilize $\mathbf{w}^{(k)}$ during the k-th global round to derive the updated global model $\mathbf{w}^{(k+1)}$.

Let $\mathcal{I}$ denote the set of clients in the system, and let $\mathcal{D}_i$ with size $|\mathcal{D}_i|$ represent the dataset of client $i \in \mathcal{I}$, corresponding to a server with the data of one of the SLNs. The loss function of client i for an arbitrary model parameter $\mathbf{w}$ is then given by $\mathcal{L}_i(\mathbf{w}) = \frac{1}{|\mathcal{D}_i|}\sum_{d\in\mathcal{D}_i}\ell(\mathbf{w};d)$, where $\ell(\mathbf{w};d)$ is the ML loss function that quantifies the prediction performance of the model parameter $\mathbf{w}$ for data point d (e.g., cross-entropy loss). Upon receiving $\mathbf{w}^{(k)}$, each client i initializes its local model as follows $\mathbf{w}_i^{(k),0} \leftarrow \mathbf{w}^{(k)}$, and subsequently updates its local model through a series of SGD iterations, indexed by e, on its local dataset. Let $e_{\max}$ denote the number of local SGD iterations performed by the clients. The evolution of the local model of client i across the SGD iterations is given by

$$\mathbf{w}_i^{(k),e} = \mathbf{w}_i^{(k),e-1} - \eta\,\widetilde{\nabla}\mathcal{L}_i\big(\mathbf{w}_i^{(k),e-1}\big), \tag{7}$$

where $e \in \{1,\cdots,e_{\max}\}$ indicates the SGD iteration count, η denotes the learning rate, and $\widetilde{\nabla}\mathcal{L}_i(\mathbf{w}_i^{(k),e-1})$ is the stochastic approximation of the gradient of the local loss function, computed as $\widetilde{\nabla}\mathcal{L}_i(\mathbf{w}_i^{(k),e-1}) = \frac{1}{|\mathcal{B}_i^{(k),e}|}\sum_{d\in\mathcal{B}_i^{(k),e}}\nabla\ell(\mathbf{w}_i^{(k),e-1};d)$, where $\mathcal{B}_i^{(k),e} \subseteq \mathcal{D}_i$ denotes a randomly selected mini-batch of data from the local dataset, and $|\mathcal{B}_i^{(k),e}|$ is its size. After completing the local model training, each client sends its latest local model parameters $\mathbf{w}_i^{(k),e_{\max}}$ to the central server. It is worth noting that while the structure of the local models is consistent with that of the global model (i.e., all models share the same CNN architecture as in centralized training), the key difference lies in their trained parameters, which reflect the variations in local datasets and the differing patterns of student interactions they represent.

After the reception of the local models $\{\mathbf{w}_i^{(k),e_{\max}}\}_{i\in\mathcal{I}}$, the central server updates the global model as follows:

$$\mathbf{w}^{(k+1)} = \sum_{i\in\mathcal{I}}\frac{|\mathcal{D}_i|\mathbf{w}_i^{(k),e_{\max}}}{|\mathcal{D}|}, \tag{8}$$

where $|\mathcal{D}| = \sum_{i\in\mathcal{I}}|\mathcal{D}_i|$ denotes the cumulative size of datasets distributed across the clients. The new global model $\mathbf{w}^{(k+1)}$ is then broadcast across the client to synchronize their local models and initiate the next round of local model training and global model aggregations.

In our implementation, the hyperparameters of FL we optimized through a grid search as follows: for each client, the mini-batch size $B = |\mathcal{B}_i^{(k),e}|$, $\forall k, e$, choices of {128, 256} and the number of SGD steps $e_{\max}$ choices of {100, 200}.

3.4 Personalization of FL Models

Computing the maximum differences between the cumulative distribution functions (CDFs) of the features across the five datasets, the Kolmogorov-Smirnov (KS) test confirmed that they follow different distributions, implying that the clients' datasets are non-independent and identically distributed (non-IID). Since each client's dataset exhibits a unique feature distribution, training a single global FL model and applying it uniformly across all clients, as done in vanilla FedAvg, may lead to suboptimal performance. Subsequently, to better capture the nuances of each SLN's data, we implemented three personalized FL (pFL) methods, ranked by their implementation complexity, as detailed below. These methods introduce modifications during the training and post-training phases of FL, allowing the global model to be tailored to individual clients' datasets.

1) FedAvg followed by Local Fine-Tuning (FedAvg+FT): In this pFL method, a post-training fine-tuning process is applied to the global model obtained via FedAvg. Let $\mathbf{w}^{(K)}$ denote the final global model obtained at the end of the K-th global aggregation round of vanilla FedAvg, marking the conclusion of the training phase. In this approach, the server broadcasts $\mathbf{w}^{(K)}$ to all clients. Each client then adapts the global model to its local dataset by performing an additional epoch of local training. Notably, conducting more epochs of local fine-tuning often causes the model to forget the globally shared data patterns across clients, potentially reducing generalization. After this fine-tuning step, the client uses the adapted local model for inference on its dataset.

2) Personalized FedAvg via Meta-Learning using Hessian-Free Approximation Method (PerFedAvg-HF) [26]: This pFL method adopts the same aggregation rule as FedAvg (see Sec. 3.3) but introduces modifications to both its training and post-training phases. The primary goal of PerFedAvg-HF is to obtain a global model during the training phase that is *adaptable* to each client's local dataset at the end of training. To achieve this adaptability, instead of performing standard

SGD iterations over the original loss function $\mathcal{L}_i(\mathbf{w})$ (as in vanilla FedAvg), each client i executes SGD iterations over a modified loss function defined as: $\mathcal{L}_i\big(\mathbf{w} - \eta\, \nabla\mathcal{L}_i(\mathbf{w})\big)$, where η denotes the learning rate. This modification encourages the global model to be more amenable to local adaptation: after one epoch of local fine-tuning, the model can be efficiently adapted to client i's dataset, captured by the update $\mathbf{w} - \eta\, \nabla\mathcal{L}_i(\mathbf{w})$. However, performing SGD over this modified loss function typically requires computing the Hessian matrix, which is computationally intensive. To address this issue, we employ the Hessian-free approximation method proposed in [26]. Similar to FedAvg, after every $e_{\max}$ local SGD iterations over the modified loss function, clients send their updated local models to the server for aggregation according to the same rule as in (8). Also, following the conclusion of the training phase, the final global model is broadcast back to the clients. Afterward, each client performs one additional epoch of fine-tuning on its local dataset, similar to the process described in the FedAvg+FT method.

3) Federated Adaptive Local Aggregation (FedALA) [65]: The local model training and global aggregation processes of FedALA are similar to vanilla FedAvg (see Sec. 3.3); however, instead of directly overwriting/synchronizing the local model with the global model after the reception of the global model at each client, FedALA employs an adaptive local aggregation (ALA) module to element-wise aggregate the received global model from the server $\mathbf{w}^{(k)}$ and the latest local model $\mathbf{w}_i^{(k-1),e_{\max}}$ at each client i. Specifically, instead of initializing the local model as $\mathbf{w}_i^{(k),0} \leftarrow \mathbf{w}^{(k)}$ as in other FL methods, in FedALA the local model is initialized as $\mathbf{w}_i^{(k),0} \leftarrow \mathbf{w}_i^{(k-1),e_{\max}} + (\mathbf{w}^{(k)} - \mathbf{w}_i^{(k-1),e_{\max}}) \odot \mathbf{W}_i$, where $\mathbf{W}_i$ is a learnable, element-wise aggregation weight matrix constrained to $[0, 1]$ via a clipping function to ensure stability and $\odot$ denotes the Hadamard product. This adaptive weighting enables FedALA to capture desired information from the global model for each client while mitigating the impact of undesired information (i.e., those that do not relate to the client's local data), thus enhancing local model personalization. FedALA further considers the fact that lower layers in deep neural networks typically extract fundamental features that are broadly applicable across clients, whereas higher layers capture more task- or client-specific features. Subsequently, in this method, the ALA module is applied only to the top p layers (p is a hyperparameter) during the local model synchronization, allowing clients to retain shared generic representations from the global model in the lower layers without modification, ensuring better generalizability of the local models. Further, FedALA introduces a hyperparameter ζ, determining the fraction of the local dataset utilized to learn the ALA weights $\mathbf{W}_i$ for each client i. The hyperparameters of FedALA were optimized through a grid search: for p, the choices were {1,2}, while for ζ, the choices covered interval $[5, 100]$ with increments of 5.

3.5 Fairness Measures

We consider two widely adopted fairness measures in the literature: *equalized odds* and *equalized opportunity* [30]. Both measures are built upon quantifiable metrics — namely the true positive rate (TPR), which assesses the model's ability to correctly identify students who form links, and the false positive rate (FPR), which evaluates how often students not connected are incorrectly classified as link-forming. Let d denote an arbitrary data point (i.e., a student pair), $\widehat{Y}(d)$ the predicted label, and $Y(d)$ the true label (i.e., 0 or 1 depending on the existence of an edge) of the data point. In this study, our goal is to assess the fairness of the model across various classrooms/SLNs (e.g., ensuring the FL global model treats students from different classrooms equally well). In this context, the *equalized odds* criterion is satisfied when both of the following probabilistic conditions hold for every pair of SLNs/clients $i, i' \in \mathcal{I}$. *(Condition 1) Equality of TPRs:* $\mathbb{P}\big(\widehat{Y}(d) = 1 \big| Y(d) = 1, d \in \mathcal{D}_i\big) = \mathbb{P}\big(\widehat{Y}(d) = 1 \big| Y(d) = 1, d \in \mathcal{D}_{i'}\big)$, *(Condition 2) Equality of FPRs:* $\mathbb{P}\big(\widehat{Y}(d) = 1 \big| Y(d) = 0, d \in \mathcal{D}_i\big) = \mathbb{P}\big(\widehat{Y}(d) = 1 \big| Y(d) = 0, d \in \mathcal{D}_{i'}\big)$. In essence, equalized odds is achieved when both the TPR (i.e., Condition 1) and FPR (i.e., Condition 2) are equal across clients/SLNs. Further, *equal opportunity* is a relaxed version of equalized odds, requiring only the equality of TPRs (i.e., the satisfaction of Condition 1) across clients. In our subsequent simulations, *we evaluate the fairness of the models by examining how closely their TPRs and FPRs adhere to these fairness criteria*, thereby assessing the extent to which the models satisfy equalized odds and equal opportunity.

3.6 Model Explainability

To explain the model's decision-making process and interpret the influence of individual features, we use SHapley Additive exPlanations (SHAP), an XAI method grounded in game theory [45]. SHAP quantifies the contribution of individual features to a model's predictions by calculating *Shapley values* for each feature, providing both global explanations (overall feature importance) and local explanations (feature impact on individual predictions). Formally, SHAP measures

$$\varphi_f = \sum_{\mathcal{F} \subseteq \mathcal{N} \setminus \{f\}} \frac{\big|\mathcal{F}\big|!(n - \big|\mathcal{F}\big| - 1)!}{n!} \Big(v(\mathcal{F} \cup \{f\}) - v(\mathcal{F}) \Big), \quad (9)$$

for every feature f (i.e., the six features discussed in Sec. 3.1.3) where φ_f is the Shapley value for feature f, indicating its contribution to the prediction. The summation in (9) iterates over all possible subsets $\mathcal{F}$ of features excluding f, where $\mathcal{N}$ is the set of all features and n denotes the number of features ($n = 6$ in our case). The term $\big|\mathcal{F}\big|!(n - \big|\mathcal{F}\big| - 1)!/n!$ is a weighting factor that ensures each subset $\mathcal{F}$ is considered fairly by accounting for all permutations in which feature f could be added. The function $v(\mathcal{F})$ represents the *value function* — measuring the model's prediction — of the subset $\mathcal{F}$, and the term $v(\mathcal{F} \cup \{f\}) - v(\mathcal{F})$ measures the marginal contribution of feature f to the prediction performance when added to the subset $\mathcal{F}$. In essence, by examining how the model's predictions change when features are omitted, SHAP identifies the positive or negative influence of each feature. We will utilize SHAP to produce visualizations that illustrate the importance of individual features across different classrooms/SLNs and their impact on the likelihood of link formation between students.

4. RESULTS & DISCUSSIONS

In this section, we report performance evaluation (Sec. 4.1) and fairness comparison (Sec. 4.2) to demonstrate the effectiveness of personalized/fine-tuned FL models over the centrally trained model and discuss insights obtained from applying XAI methods (Sec. 4.3). Finally, we offer implications of this study for the educators and institutions (Sec. 4.4).

Table 2: Comparison of performance (± standard deviation) between implemented methods shows that the personalized federated model FedALA outperformed the centralized model for all datasets. Metric values in **bold** denote the best results.

Course	Metrics	Centralized (η=0.001, E=200, B=256)	FedAvg [47] (η=0.001, K=30, $e_{\max}$=200, B=256)	FedAvg+FT (η=0.0001, B=64, $e_{\max}$=200)	PerFedAvg-HF [26] (η=0.01, B=256, $e_{\max}$=350, K=15)	FedALA [65] (η=0.01, K=30, $e_{\max}$=100, B=128, p=2, ζ=80)
algo	Accuracy(%)	93.62 ± 0.34	93.26 ± 0.26	93.88 ± 0.01	93.7 ± 0.03	$\mathbf{94.01 \pm 0.14}$
	Loss	0.18 ± 0.01	0.2 ± 0.01	0.17 ± 0.01	0.2 ± 0.002	$\mathbf{0.17 \pm 0.002}$
	AUC	0.9 ± 0.002	0.9 ± 0.001	0.89 ± 0.01	0.9 ± 0.01	$\mathbf{0.91 \pm 0.001}$
comp	Accuracy(%)	92.26 ± 0.19	92.22 ± 0.06	92.65 ± 0.15	92.00 ± 0.37	$\mathbf{92.76 \pm 0.04}$
	Loss	0.2 ± 0.01	0.22 ± 0.2	0.18 ± 0.003	0.21 ± 0.001	$\mathbf{0.18 \pm 0.002}$
	AUC	0.86 ± 0.001	0.86 ± 0.01	0.85 ± 0.002	0.84 ± 0.001	$\mathbf{0.86 \pm 0.01}$
shake	Accuracy(%)	90.52 ± 0.18	90.27 ± 0.08	91.2 ± 0.05	90.98 ± 0.13	$\mathbf{91.21 \pm 0.05}$
	Loss	0.22 ± 0.003	0.23 ± 0.01	0.2 ± 0.01	0.23 ± 0.001	$\mathbf{0.2 \pm 0.001}$
	AUC	0.84 ± 0.01	0.84 ± 0.01	0.8 ± 0.01	0.83 ± 0.001	$\mathbf{0.84 \pm 0.01}$
ml	Accuracy(%)	86.31 ± 0.15	84.93 ± 0.17	86.7 ± 0.34	86.93 ± 0.31	$\mathbf{87.2 \pm 0.32}$
	Loss	0.33 ± 0.002	0.37 ± 0.01	0.32 ± 0.01	0.36 ± 0.002	$\mathbf{0.32 \pm 0.01}$
	AUC	0.85 ± 0.002	0.82 ± 0.01	0.86 ± 0.001	0.86 ± 0.001	$\mathbf{0.86 \pm 0.01}$
csc	Accuracy(%)	76.57 ± 0.59	75.52 ± 1.77	79.71 ± 0.004	81.48 ± 0.15	$\mathbf{82.11 \pm 2.51}$
	Loss	0.62 ± 0.05	0.56 ± 0.01	0.44 ± 0.02	0.48 ± 0.03	$\mathbf{0.41 \pm 0.01}$
	AUC	0.58 ± 0.01	0.61 ± 0.01	0.7 ± 0.02	0.76 ± 0.01	$\mathbf{0.77 \pm 0.01}$

4.1 Link Formation Prediction

Here, we present the performance evaluation of our proposed link prediction methodology. We compare our approach to a centralized baseline model for link prediction, along with FedAvg and variations of pFL. As discussed in Sec. 3.2, the centralized training model serves as the *ideal baseline* as it has access to the pool of data of all the SLNs and does not adhere to privacy restrictions.

To quantify the performance, we calculated the accuracy, loss, and area under curve (AUC) of the models on each SLN's test data and reported them in Table 2. AUC measures the probability that a randomly chosen positive instance (i.e., a connected student pair) is ranked higher than a randomly chosen negative instance (i.e., a non-connected pair) by the model. An AUC of 0.5 indicates no discriminative power (equivalent to random guessing), while an AUC of 1.0 signifies perfect class separation. As seen in Sec. 3.1, SLNs are often sparse graphs [61] showing class imbalance (i.e., most of the student pairs are not connected); hence, AUC unveils how well classes are separated [53].

We present our results in Table 2, where the top row lists the various methods along with their corresponding hyperparameters, where E denotes the number of training epochs in the centralized training scenario, while K represents the number of global aggregation rounds in the FL settings and the remaining notations are defined in Sec. 3.3. All the models reached convergence under the specified settings.

Inspecting the results in Table 2, *FedAvg* exhibits a slightly worse performance than the *centralized* baseline, unveiling the fact that SLNs data are non-IID and training a single global model and applying it across all the SLNs, as done in vanilla *FedAvg*, may not be an optimal choice. Nevertheless, the results demonstrate that as we move toward the pFL methods, we obtain performance gains over the *centralized* baseline. These gains stem from the fact that pFL methods tailor individual models to the SLNs that can well-capture the local data characteristics and student interaction patterns in each SLN. In particular, *FedALA* achieves the best performance across all the methods for all the classrooms/SLNs with the highest accuracy, lowest loss, and highest AUC.

Table 3: TPR and FPR values for centralized, federated, and personalized FedALA models across different classrooms.

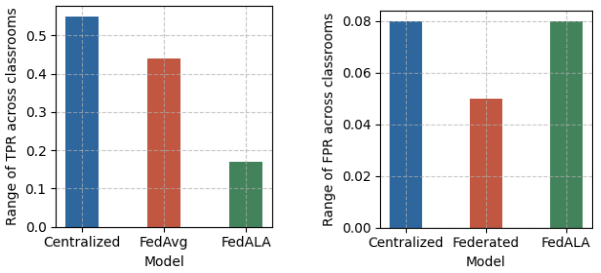
Course	Centralized (TPR, FPR)	FedAvg (TPR, FPR)	FedALA (TPR, FPR)
algo	(0.83, 0.03)	(0.81, 0.05)	(0.84, 0.03)
comp	(0.72, 0.04)	(0.76, 0.05)	(0.74, 0.04)
shake	(0.72, 0.05)	(0.77, 0.07)	(0.68, 0.04)
ml	(0.81, 0.1)	(0.71, 0.08)	(0.85, 0.11)
csc	(0.28, 0.02)	(0.37, 0.03)	(0.68, 0.05)

Assessing the *centralized* model performance across various SLNs, *csc* possesses the worst accuracy, indicating that its dataset highly differs — both in terms of its size and its data patterns — from other datasets. This suggests that *csc* may benefit the most from the pFL methods. The results confirm this hypothesis with *csc* obtaining the highest performance improvements when comparing its accuracy between the *centralized* and *FedALA* methods — approximate 5.54% increase in accuracy and 19% increase in AUC. Also, examining the overall results across all datasets reveals a consistent trend of improved performance when using pFL methods over the centralized baseline. In particular, the best-performing pFL method (i.e., FedALA) achieves an average improvement of 1.6% in accuracy and a 4.2% increase in mean AUC compared to the centralized model, highlighting the effectiveness of FL model personalization in handling diverse SLNs.

Conducting a paired t-test to assess the statistical significance of the performance differences with and without model personalization, we obtained a p-value less than 0.05 at a 95% confidence interval. This result indicates that the improvements achieved by *FedALA* over the *centralized* model are statistically significant. Beyond this notable performance gain, *FedALA* and all the other pFL methods offer the added benefit of preserving data privacy, further underscoring their practical advantages in real-world scenarios.

4.2 Fairness Evaluation

Next, we examine the fairness of the models obtained from different methods, based on the methodology outlined in Sec. 3.5. To this end, we compare how close the TPR and FPR were across the classrooms/SLNs through the values



(a) TPR range (b) FPR range

Figure 3: The range (difference between the highest and lowest) of TPR and FPR values across all classrooms/SLNs.

presented in Table 3. In this comparison, we focus solely on the best-performing pFL model identified in Table 2, namely *FedALA*, to highlight how personalization impacts fairness relative to the vanilla *FedAvg* and *centralized* methods.

We first focus on the TPR values in Table 3 and observe a higher consistency across the TPR values for *FedALA* compared to both *centralized* and *FedAvg* methods. This implies that *FedALA* adheres to the equal opportunity measure more closely. To quantify this finding, we compute the equal opportunity difference (EOD) metric [25], which measures the difference between TPR values across the classrooms, referred to as the *range* of TPR. A smaller *range* of TPR is desirable as it reflects more consistent positive predictions across different student pairs belonging to different SLNs, contributing to model fairness — the most fair model would have the *range* of 0. We depict the *range* of TPR of the methods in Figure 3(a), which verifies that *FedALA* achieves the lower TPR *range*, implying its higher adherence to equal opportunity measure. On the other hand, inspecting the FPR values in Table 3 there seem to be no conclusive results: the range (the difference between the maximum and minimum value computed over the SLNs) of FPR for *FedALA* is on par with *centralized* but is worse than *FedAvg* as also observed in Figure 3(b). So, a concrete conclusion on the equal odds measure cannot be drawn from the results. In summary, by exhibiting a better TPR consistency performance in terms of EOD, *FedALA* outperforms the other methods in terms of model fairness. Also, the results open a door to future studies on improving the adherence of FL models to the equal odds measure in SLNs.

4.3 Contributions of Features in Predictions

We next discuss the explainability of the models, based on the methodology outlined in Sec. 3.6. We showcase Shapley values in different SHAP plots that illustrate the importance/impact of each feature on the models' predictions. In the following, we study the SHAP plots in terms of both individual level interpretations for a student pair as well as global level interpretations for all student pairs.

4.3.1 Individual Level Explanation

SHAP allows us to perform data instance-specific explanations based on the computed Shapley values, where each feature gets a score representing its contribution to moving the prediction away from the base value (mean prediction value for the dataset). Focusing on the predictions made by *FedALA* model, Figure 4 shows the most important features and their corresponding values for a linked student pair in *algo* class. The base value (i.e., 0.1464) represents the average model output across the dataset before incorporating individual feature effects. The final predicted value is 0.81, which is much higher than the base value, meaning that the contributing features collectively increased the prediction. Each feature's impact is visualized by arrows: the longer the arrow for a feature, the greater the impact of that feature on the prediction. Note that the red-colored features increased the link formation probability higher, while the blue-colored features did the opposite. As can be seen, high *Jaccard Coefficient, Cosine Similarity*, and *Resource Allocation Index* values helped the model in its predictions, whereas low *Preferential Attachment Score* and *Adamic-Adar Index* values have often lowered the prediction.

This observation highlights two important points: (i) features *do not* contribute equally to the model's predictions, and (ii) there may be *significant differences* in the contribution of individual features, underscoring the importance of understanding feature relevance for informed model interpretation. This motivates our subsequent discussions, where we discuss the importance of the features between the pairs of students in all classrooms and investigate their impacts.

4.3.2 Global Level Explanation

We next provide dataset-wide explanations on trends of feature importance, focusing on the FedALA model. In Figures 5(a)-(d), and 6, we illustrate the relative feature importance using SHAP plots to assess feature contributions at a global level for all student pairs for both MOOC and non-MOOC courses/SLNs. The X-axis represents each feature's relative impact on the model's predictions. We summarize the main takeaways of these results below:

STEM and MOOC SLNs (*algo* & *ml*): *Resource Allocation Index* and *Cosine Similarity* (Figures 5(a) & (b)) have the most impact on the model prediction performance. Inspecting the logic behind these two metrics, discussed in Sec. 3.1.3, this outcome matches the intuition that STEM classrooms expect students to collaborate for learning and achieving success, and thus any two students working on similar assignments will likely share several mutual peers, increasing the probability of their direct collaboration. Furthermore, two students with similar coding performance on assignments will likely discuss solutions in MOOC forums or work together on projects. On the other hand, in a MOOC format course, due to the lack of face-to-face interactions, students tend to have more diverse conversations with random peers. Hence, two students who communicate with similar neighbors more exclusively have a higher potential to connect as part of the same online sub-community.

Non-STEM and MOOC SLNs (*comp* & *shake*): *Resource Allocation Index* and *Preferential Attachment Score* (Figures 5(c) & (d)) have the most impact on the model performance. Inspecting the logic behind these two metrics, discussed in Sec. 3.1.3, this outcome also matches the intuition that non-STEM classrooms encourage open-ended discussions, and thus, students might frequently respond to different peers in a MOOC discussion forum. As a result, a highly active pair of students has a higher probability of corresponding with one another. Also, similar to the STEM and MOOC SLNs discussed above, since correspondences in the MOOC format are more diverse, a high *Resource Allocation Index* can

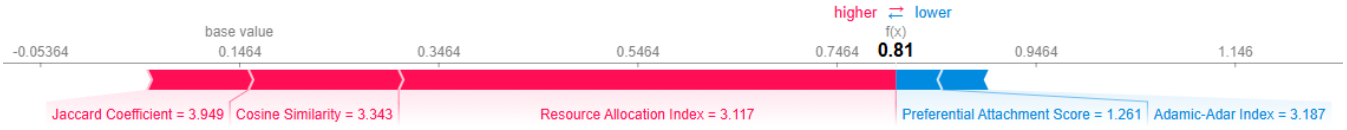
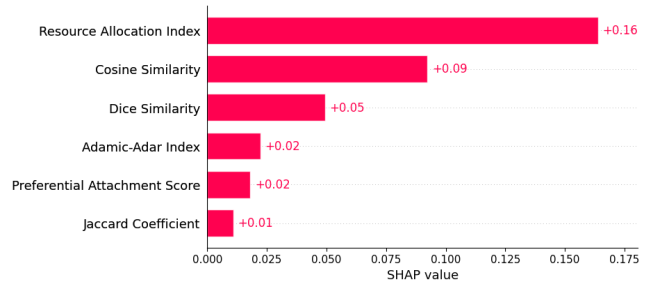
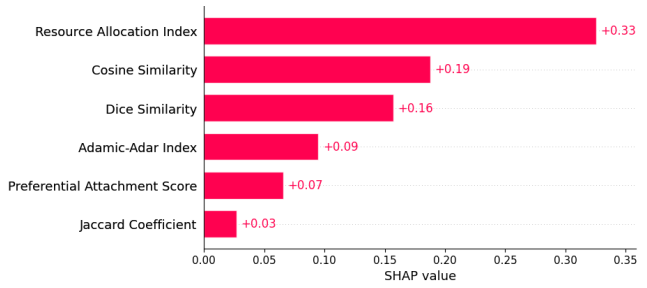


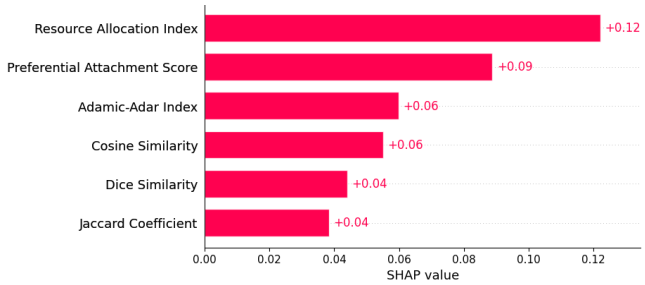
Figure 4: Force plot for a learner pair in *algo* classroom, presenting the relationship between feature values and prediction.



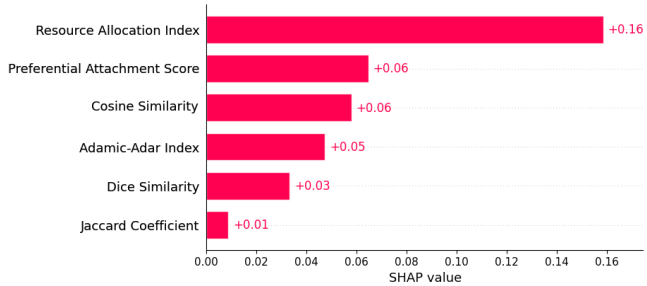
(a) algo



(b) ml



(c) comp



(d) shake

Figure 5: Relative importance of features across all MOOC SLNs show the same set of top-2 features for (a) and (b), which are STEM courses, and the same set of top-2 features for (c) and (d), which are non-STEM courses.

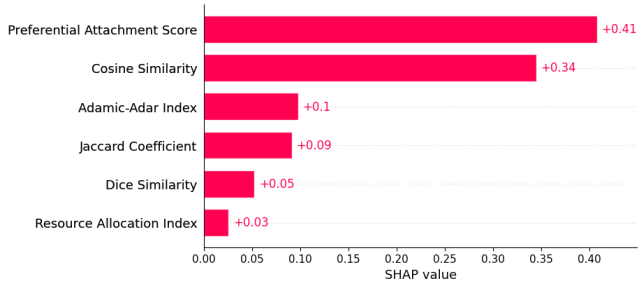


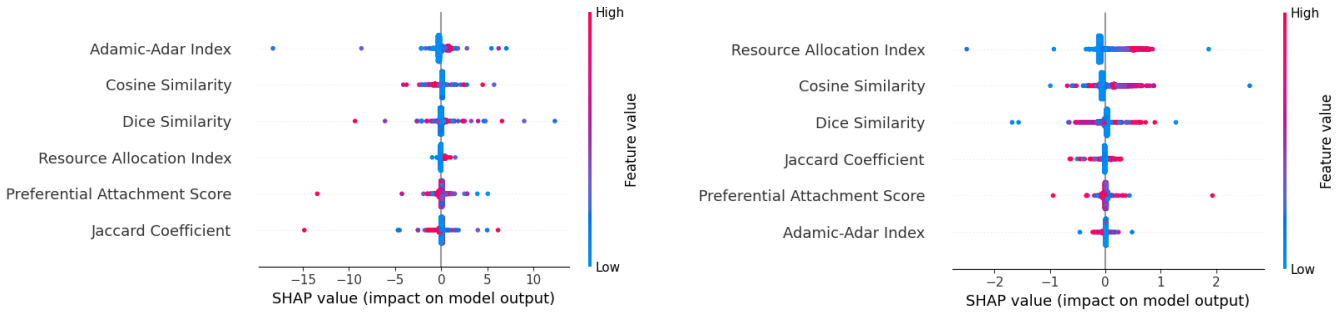
Figure 6: Feature importance in STEM & non-MOOC *csc*.

demonstrate students who have formed an online community.

STEM and non-MOOC SLN (*csc*): *Preferential Attachment Score* and *Cosine Similarity* (Figure 6) have the most impact on the model performance. Inspecting the logic behind these two metrics, discussed in Sec. 3.1.3, this outcome matches the intuition that in a non-MOOC STEM classroom, students tend to work in structured groups formed during in-person interactions, and thus, the proportion of shared neighbors plays an important role in increasing the possibility of future link formation. Furthermore, since online correspondence in this type of class is less compared to non-STEM and MOOC, students who have a tendency to be active online have a higher tendency to communicate.

In summary, *Resource Allocation Index* turns out to be the most prominent feature for MOOC courses — implying that common neighbors and neighborhood influence forming connections in these settings. Interestingly, this feature is the least important for the non-MOOC course — implying less reliance on shared connections to predict interactions. Such differences in feature importance can highlight classrooms' unique needs and aid instructors with intervention guidance.

To further understand the relationship between feature values and predicted outcomes and how it may differ upon conducting model personalization, we present summary plots for all student pairs in *algo* classroom in Figure 7. Each point in the plot represents the Shapley value for a student pair, and features are sorted according to their importance. Data points with overlapping Shapley values oscillate along the Y-axis. The color gradient reflects feature values, with red indicating high values and blue representing low values. Between the *centralized* and *FedALA* models, we observe a notable change in the order of the features. This reveals a nuanced phenomenon: *FedALA obtains the performance gains reported in Table 2 by putting a different emphasis on the features, which are tailored to each SLN.* Interestingly, from being the most important feature for the *centralized* model, *Adamic-Adar Index* becomes the least important feature after personalization in *FedALA*, implying the shift of focus to *Resource Allocation Index* that boosts low-degree nodes more aggressively than *Adamic-Adar Index* for this MOOC



(a) Centralized (b) FedALA

Figure 7: Feature importance explainability plots for the *algo* classroom under (a) centralized and (b) FedALA model training. The results unveil a change in the significance/contribution of individual features from the *centralized* to *FedALA* model.

STEM course. These findings can be intuitively explained further in the context of an algorithm online course, where learners do not randomly connect with others but build their network based on shared connections. This might happen as learners trust and engage with those connected to their existing network. As a result, students with common neighbors are more likely to connect in discussions, projects, and coding forums, leading to better engagement and learning success. These observations can offer distinguishing patterns that have the potential to help instructors take more control and provide better learning environments.

The results from other datasets, which are omitted here due to their qualitative similarity to Figure 7, similarly demonstrate that the FedALA model effectively captures the unique interaction dynamics and identifies features that are more relevant within each classroom.

4.4 Implications for Design

The experimental results collectively showed that FL can learn distinct behavioral patterns for students in each classroom, especially when fine-tuned/personalized locally for better adaptation to the individual data patterns in each SLN. This alleviates the need for using a single centralized model that needs direct access to data from all classrooms, which can cause privacy concerns. The results further provide opportunities to improve students' learning experiences through timely and effective interventions. For instance, link prediction can help identify students who may engage less socially, allowing teachers to target interventions like pairing them with peers to foster collaboration. Such interventions can help struggling students, promote stronger peer connections, and increase overall class performance [13, 40]. Another example is when students who may benefit the most from collaborative study or project groups are identified. For example, the odds of students forming links can be correlated with their personal characteristics, such as academic strengths, learning styles, engagement patterns, etc. Coupled with their class performance measures, the results can further inspire methods to identify characteristics of successful collaboration among students, leading to intentional pairings that can lead to more balanced and effective group dynamics.

5. LIMITATIONS & FUTURE WORK

Focusing on the limitation of our work, the datasets used for this study did not have student demographic information due to data anonymization practices, preventing us from fine-tuning and testing our pFL models in various student demographic groups. We aim to work on this in the future to demonstrate FL's potential to address the bias against underrepresented/minority students.

Focusing on future work, investigation of the relationship between predicted links and actual student learning gains in a real-world scenario is a promising research direction. In addition, it is worth studying few-shot learning techniques [28, 4] to adapt models trained in large classrooms in the pFL setting to smaller or underrepresented ones with limited data. Furthermore, using large language models (LLMs) [66, 60, 62, 68] to (a) generate human-readable explanations of link predictions, aiding educators in understanding student networks, (b) extract rich semantic features from students' forum posts, chat logs, and written submissions for improved link prediction, and (c) suggest personalized interventions (e.g., peer group formations) based on student interaction histories, all are exciting research opportunities.

6. CONCLUSION

This study presented a novel application of federated learning (FL) for link prediction in social learning networks (SLNs). By modeling SLNs from multiple classrooms as graph structures representing student interactions, we computed topological features for student pairs and utilized them for model training and testing. Unlike centralized training methods that require aggregating raw data, our FL-based approach allowed each classroom/SLN to train its model locally while only transmitting model parameters to a server for aggregation, preserving the privacy of student interaction data. To further enhance performance, we incorporated model personalization techniques, enabling the FL model to adapt to the distinctive characteristics of individual classrooms. Our experimental results demonstrated that personalized FL models outperformed both vanilla FL and centralized models, highlighting the impact of unique student interaction patterns inherent to individual SLNs. In addition, we explored the explainability of the trained models using explainable AI (XAI) techniques. This analysis revealed differences in the importance of various topological features for link prediction across different classroom types (e.g., STEM vs. non-STEM, MOOC vs. non-MOOC), offering valuable insights into how interaction patterns vary by context.

7. REFERENCES

[1] Family educational rights and privacy act (FERPA). https://epic.org/family-educational-rights-and-privacy-act-ferpa/, Online; accessed 19 Feb. 2025.

[2] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[3] S. Aghababaei and M. Makrehchi. Interpolative self-training approach for link prediction. *Intelligent Data Analysis*, 23(6):1379–1395, 2019.

[4] C. Aguirre, K. Sasse, I. Cachola, and M. Dredze. Selecting shots for demographic fairness in few-shot learning with large language models. *arXiv preprint arXiv:2311.08472*, 2023.

[5] S. Amershi, C. Conati, et al. Combining unsupervised and supervised classification to build user models for exploratory learning environments. *Journal of educational data mining*, 1(1):18–71, 2009.

[6] 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.

[7] M. Asad, A. Moustafa, and T. Ito. Federated learning versus classical machine learning: A convergence comparison. *arXiv preprint arXiv:2107.10976*, 2021.

[8] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 635–644, 2011.

[9] J. Baek, W. Jeong, J. Jin, J. Yoon, and S. J. Hwang. Personalized subgraph federated learning. In *International conference on machine learning*, pages 1396–1415. PMLR, 2023.

[10] A. Bandura. Social learning theory. *Englewood Cliffs*, 1977.

[11] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[12] M. Bashiri and K. Kowsari. Transformative influence of llm and ai tools in student social media engagement: Analyzing personalization, communication efficiency, and collaborative learning. *arXiv preprint arXiv:2407.15012*, 2024.

[13] S. Berlinski, M. Busso, and M. Giannola. Helping struggling students and benefiting all: Peer effects in primary education. *Journal of Public Economics*, 224:104925–104948, 2023.

[14] S. Bhattacharya, P. S. Vyas, S. Yarradoddi, B. Dasari, S. Sumathy, R. Kaluri, S. Koppu, D. J. Brown, M. Mahmud, and T. R. Gadekallu. Towards smart education in the industry 5.0 era: A mini review on the application of federated learning. In *Proceedings of the 2023 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 602–608. IEEE, 2023.

[15] M. Bojanowski and B. Chroł. Proximity-based methods for link prediction in graphs with r package'linkprediction'. *ASK. Research and Methods*, 29(1):5–28, 2020.

[16] C. G. Brinton, S. Buccapatnam, L. Zheng, D. Cao, A. S. Lan, F. M. Wong, S. Ha, M. Chiang, and H. V. Poor. On the efficiency of online social learning networks. *IEEE/ACM Transactions on Networking*, 26(5):2076–2089, 2018.

[17] R. Brown, C. F. Lynch, M. Eagle, J. L. Albert, T. Barnes, R. S. Baker, Y. Bergner, and D. S. McNamara. Good communities and bad communities: Does membership affect performance? In O. C. Santos, J. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, M. C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, and M. C. Desmarais, editors, *Proceedings of the 8th International Conference on Educational Data Mining, EDM 2015, Madrid, Spain, June 26-29, 2015*, pages 612–613. International Educational Data Mining Society (IEDMS), 2015.

[18] B. V. Carolan. *Social network analysis and education: Theory, methods & applications*. Sage Publications, 2013.

[19] S. Choudhary, K. Sharma, and M. Bajaj. Social networks analysis and machine learning: an overview of approaches and applications. In *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, pages 123–128, 2023.

[20] Y.-W. Chu, S. Hosseinalipour, E. Tenorio, L. Cruz, K. Douglas, A. Lan, and C. Brinton. Mitigating biases in student performance prediction via attention-based personalized federated learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3033–3042, 2022.

[21] Y.-W. Chu, S. Hosseinalipour, E. Tenorio, L. Cruz, K. Douglas, A. S. Lan, and C. G. Brinton. Multi-layer personalized federated learning for mitigating biases in student predictive analytics. *IEEE Transactions on Emerging Topics in Computing*, X:1–15, 2024.

[22] T. D. V. and P. D. V. Evaluation of privacy-preserving techniques: Bouncy castle encryption and machine learning algorithms for secure classification of sensitive data. *International Journal of Intelligent Systems and Applications in Engineering*, 11:429 –, Jul. 2023.

[23] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[24] M. Ebrahimi, R. Sahay, S. Hosseinalipour, and B. Akram. The transition from centralized machine learning to federated learning for mental health in education: A survey of current methods and future directions. *arXiv preprint arXiv:2501.11714*, 2025.

[25] Y. H. Ezzeldin, S. Yan, C. He, E. Ferrara, and A. S. Avestimehr. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 7494–7502, 2023.

[26] A. Fallah, A. Mokhtari, and A. Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in neural information processing systems*, 33:3557–3568, 2020.

[27] A. Gharahighehi, K. Pliakos, and C. Vens. Addressing the cold-start problem in collaborative filtering through

positive-unlabeled learning and multi-target prediction. *IEEE Access*, 10:117189–117198, 2022.

[28] Y. Guo and L. Zhang. One-shot face recognition by promoting underrepresented classes. *arXiv preprint arXiv:1707.05574*, 2017.

[29] N. Gurjar. Leveraging social networks for authentic learning in distance learning teacher education. *TechTrends*, 64(4):666–677, 2020.

[30] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3323–3331, 2016.

[31] M. A. Hasan and M. J. Zaki. *A Survey of Link Prediction in Social Networks*, pages 243–275. Springer US, Boston, MA, 2011.

[32] B. Hollister, P. Nair, S. Hill-Lindsay, and L. Chukoskie. Engagement in online learning: student attitudes and behavior during covid-19. In C. Blake, editor, *Frontiers in education*, volume 7, page 851019, Lausanne, Switzerland, 2022. Frontiers Media SA.

[33] A. P. Hridi, R. Sahay, S. Hosseinalipour, and B. Akram. Revolutionizing ai-assisted education with federated learning: A pathway to distributed, privacy-preserving, and debiased learning ecosystems. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 297–303, 2024.

[34] P. Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912.

[35] Z. L. Jiang, J. Gu, H. Wang, Y. Wu, J. Fang, S.-M. Yiu, W. Luo, and X. Wang. Privacy-preserving distributed machine learning made faster, 2022.

[36] Z. Jie, Y. Li, and R. Liu. Social network group identification based on local attribute community detection. In *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pages 443–447. IEEE, 2019.

[37] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

[38] G. Kanselaar. Constructivism and socio-constructivism. *Journal of Educational Psychology*, 94(3):1–10, 2002.

[39] M. Khelghatdoust and M. Mahdavi. A socially-aware, privacy-preserving, and scalable federated learning protocol for distributed online social networks. In *International Conference on Advanced Information Networking and Applications*, pages 192–203. Springer, 2022.

[40] H. Li, S. Zhang, S. Lee, J.-E. Lee, Z. Zhong, E. Weitnauer, and A. F. Botelho. Math in motion: Analyzing real-time student collaboration in computer-supported learning environments. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 533–541, 2024.

[41] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth International Conference on Information and Knowledge Management*, pages 556–559, 2003.

[42] J. Liu, J. Huang, Y. Zhou, X. Li, S. Ji, H. Xiong, and D. Dou. From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems*, 64(4):885–917, 2022.

[43] Z. Liu, L. Yang, Z. Fan, H. Peng, and P. S. Yu. Federated social recommendation with graph neural network. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–24, 2022.

[44] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.

[45] S. Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[46] L. Ma, H. Han, J. Li, H. Shomer, H. Liu, X. Gao, and J. Tang. Mixture of link predictors on graphs. *arXiv preprint arXiv:2402.08583*, 2024.

[47] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282, Fort Lauderdale, FL, USA, Apr. 2017. PMLR.

[48] M. Mezghani, M. Washha, and F. Sèdes. Online social network phenomena: buzz, rumor and spam. In Z.-H. Zhou and C. Zhang, editors, *How information systems can help in alarm/alert detection*, pages 219–239. Elsevier, Oxford, UK, 2018.

[49] C. P. Muniz, R. Goldschmidt, and R. Choren. Combining contextual, temporal and topological information for unsupervised link prediction in social networks. *Knowledge-Based Systems*, 156:129–137, 2018.

[50] S. Peng, Y. Yang, M. Mao, and D.-S. Park. Centralized machine learning versus federated averaging: A comparison using mnist dataset. *KSII Transactions on Internet and Information Systems (TIIS)*, 16(2):742–756, 2022.

[51] B. Rubin, R. Fernandes, M. D. Avgerinou, and J. Moore. The effect of learning management systems on student and faculty outcomes. *The internet and higher education*, 13(1-2):82–83, 2010.

[52] R. Sahay, S. Nicoll, M. Zhang, T.-Y. Yang, C. Joe-Wong, K. A. Douglas, and C. G. Brinton. Predicting learning interactions in social learning networks: a deep learning enabled approach. *IEEE/ACM Transactions on Networking*, 31(5):2086–2100, 2023.

[53] T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):1–21, 2015.

[54] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[55] D. Sengupta, S. S. Khan, S. Das, and D. De. Fedel: Federated education learning for generating correlations between course outcomes and program outcomes for internet of education things. *Internet of Things*, 25:101056–101082, 2024.

[56] D. Williams-Dobosz, R. F. L. Azevedo, A. Jeng, V. Thakkar, S. Bhat, N. Bosch, and M. Perry. A social network analysis of online engagement for college students traditionally underrepresented in stem. In *LAK21: 11th International Learning Analytics and*

Knowledge Conference, pages 207–215, 2021.

[57] C. Wu, F. Wu, Y. Cao, Y. Huang, and X. Xie. Fedgnn: Federated graph neural network for privacy-preserving recommendation. *arXiv preprint arXiv:2102.04925*, 2021.

[58] J. Wu, Z. Huang, Q. Liu, D. Lian, H. Wang, E. Chen, H. Ma, and S. Wang. Federated deep knowledge tracing. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 662–670, 2021.

[59] T.-m. Wut and J. Xu. Person-to-person interactions in online classroom settings under the impact of covid-19: a social presence theory perspective. *Asia Pacific Education Review*, 22(3):371–383, 2021.

[60] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, Y. Wang, and E. Chen. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357–186381, 2024.

[61] Y. Xu, C. F. Lynch, and T. Barnes. How many friends can you make in a week?: Evolving social relationships in moocs over time. *International Educational Data Mining Society*, 2018.

[62] F. Yang, Z. Chen, Z. Jiang, E. Cho, X. Huang, and Y. Lu. Palr: Personalization aware llms for recommendation. *arXiv preprint arXiv:2305.07622*, 2023.

[63] T.-Y. Yang, C. G. Brinton, and C. Joe-Wong. Predicting learner interactions in social learning networks. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 1322–1330. IEEE, 2018.

[64] S. Yun, S. Kim, J. Lee, J. Kang, and H. J. Kim. Neo-gnns: Neighborhood overlap-aware graph neural networks for link prediction. *Advances in Neural Information Processing Systems*, 34:13683–13694, 2021.

[65] J. Zhang, Y. Hua, H. Wang, T. Song, Z. Xue, R. Ma, and H. Guan. Fedala: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11237–1fedl2p1244, 2023.

[66] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

[67] T. Zhou, L. Lü, and Y.-C. Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71:623–630, 2009.

[68] A. Zytek, S. Pido, S. Alnegheimish, L. Berti-Equille, and K. Veeramachaneni. Explingo: Explaining ai predictions using large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1197–1208. IEEE, 2024.