Who Is Lagging Behind: Profiling Student Behaviors with Graph-Level Encoding in Curriculum-Based Online Learning Systems

Qian Xiao
School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
xiaoq@tcd.ie

Conn Breathnach
School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
cobreath@tcd.ie

Ioana Ghergulescu
Adaptemy
Dublin, Ireland
ioana.ghergulescu@adaptemy.com

Conor O'Sullivan
Adaptemy
Dublin, Ireland
conor.osullivan@adaptemy.com

Keith Johnston
School of Education
Trinity College Dublin
Dublin, Ireland
keith.johnston@tcd.ie

Vincent Wade
School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
vincent.wade@tcd.ie

Abstract—The surge in the adoption of Intelligent Tutoring Systems (ITSs) in education, while being integral to curriculum-based learning, can inadvertently exacerbate performance gaps. To address this problem, student profiling becomes crucial for tracking progress, identifying struggling students, and alleviating disparities among students. Such profiling requires measuring student behaviors and performance across different aspects, such as content coverage, learning intensity, and proficiency in different concepts within a learning topic.

In this study, we introduce `CTGraph`, a graph-level representation learning approach to profile learner behaviors and performance in a self-supervised manner. Our experiments demonstrate that `CTGraph` can provide a holistic view of student learning journeys, accounting for different aspects of student behaviors and performance, as well as variations in their learning paths as aligned to the curriculum structure. We also show that our approach can identify struggling students and provide comparative analysis of diverse groups to pinpoint when and where students are struggling. As such, our approach opens more opportunities to empower educators with rich insights into student learning journeys and paves the way for more targeted interventions.

Index Terms—Educational technology, Learning systems, Computational behavioral modeling, Data mining, Graph neural networks.

I. INTRODUCTION

Intelligent Tutoring Systems (ITSs) are becoming deeply integrated at all levels of curriculum-based learning, ranging from K-12 education to higher education and professional upskilling. The development of ITSs is often motivated by their potential to provide personalized learning paths that are more suitable for individual students' needs. This motivation aligns with the United Nations' sustainable development goal on quality education (SDG 4): "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" [1]. However, without significantly scaling up the capacity to (a) monitor student progress and (b) provide a comprehensive data-driven framework to guide educational priorities and deepen educators' understanding of students' learning journeys, this surging adoption of digital platforms in education may risk widening achievement gaps among students [2]–[4].

Meanwhile, generative AI is increasingly recognised for its potential positive influence on education, such as enabling broader student access to customised learning materials. However, there is growing concern regarding the assessment of work produced by students using such generative AI tools. For instance, students may use these tools to generate assignment solutions without any genuine comprehension or deep engagement with the subject by themselves. Summative assessments, conducted entirely outside of an invigilated environment, are becoming increasingly unreliable. Current estimates indicate that over 70% of students regularly use ChatGPT [5]. Therefore, in the era of generative AI, much more sophisticated and comprehensive performance management techniques are necessary to adequately reflect student effort and performance. These techniques need to identify where intervention may be required and encourage students to adopt good learning behaviours and practices. This highlights the pressing need for comprehensive monitoring and assessment of various student behaviors and performance measures.

Thus, it is crucial to profile students with a holistic and evidence-based framework to support more accurate and informative assessment [4]. Specifically, given that learner performance is highly integrated and comprises many different interacting factors [4], this framework needs to include multifaceted elements of student behaviours that contribute to academic success. Moreover, to ensure leaving no one behind, we also need student profiling to identify any disparities
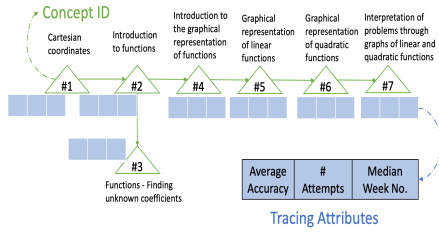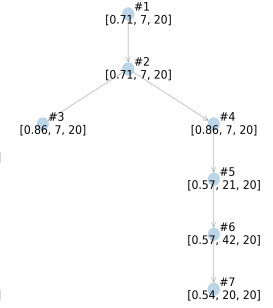
(a) Curriculum Structure for the Topic *Functions I*
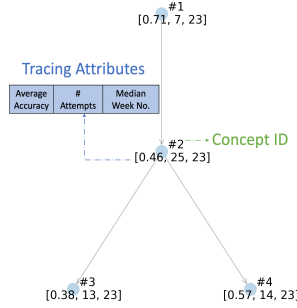
(b) Student 1
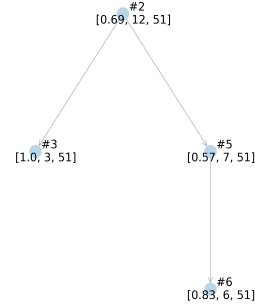
(c) Student 2

(d) Student 3

Fig. 1: Motivating Example: (a) presents a graph illustrating the curriculum structure for the learning topic *Functions I*, where nodes represent concepts (i.e., knowledge items), directed edges represent prerequisite conditions between concepts, and node attributes include statistics of learning traces for each concept; (b) to (d) display the learning paths of three students as aligned with the curriculum structure, along with the corresponding statistics of learning traces as node attributes for each concept.

or significant achievement gaps among individual students as well as between student cohorts. Additionally, it is also essential to create student profiles capable of identifying when and where students are struggling.

Student profiling for ITSs requires a comprehensive assessment of student behaviors and performance across various aspects, particularly in terms of content coverage, learning intensity, and mastery of skills/concepts based on the curriculum structures. Unlike conventional classroom education, teachers do not have full control in ITSs over students' learning pacing and choices of studying materials. Therefore, even though the curriculum remains the same for each student, students' navigation across the concepts in the curriculum and associated learning materials often differ significantly. To attain an objective assessment of student behaviors and performance in ITSs, we need 1) a holistic picture of students' varying learning status across different aspects (e.g. learning paths, intensity, coverage) to understand student progress; 2) the continuous monitoring and identifying when and where students are struggling; 3) the capability to enable comparative analysis among different groups of students for understanding the impact of different choices in learning paths and pacing on student performance.

To better illustrate these challenges, we use one motivating example as depicted in Figure 1. Figure 1a shows an example of concept structures for the mathematical topic *Functions I* in a curriculum-based learning ITS platform. A student's navigation and interactions with the learning materials associated with these concepts in this topic can be represented with attributes on each concept node. In this study, we extract average accuracy, the total number of attempts on the questions associated with a concept and the median week number in the academic calendar from tracing data to be the tracing attributes. In Figures 1b to 1d, we list three graphs that demonstrate three different students' navigation and interactions with the learning materials in Figure 1a. As shown in this example, only student 3 strictly adhered to the curriculum structure, while the other two students skipped learning on

different concepts as indicated in the curriculum structure. Student 1 and 2 covered fewer concepts than Student 3. For each student, the performance (in terms of accuracy) varies on different concepts. For example, student 3 has much better performance in concepts 1-4 than in concepts 5-7. The learning intensity (in terms of the number of attempts on questions) and the timing of attempts (in terms of week number in the academic calendar year) are also very different. Due to such disparity of student behaviors and performance, we need to profile both student performance as well as their behaviors across different aspects to capture such diversity in students' learning journeys. Note that, single statistical indicators, while useful for understanding each aspect independently at a high level, cannot capture such cross-dimensional correlations or provide detailed comparative analysis at fine granularity for each student.

To address these challenges above, we present a novel student profiling approach that uses graph representation learning methods to model student behaviors, performance and learning paths simultaneously. Graphs are effective for representing diverse types of data in real-world applications. Representing student behaviors in graphs provides explicit structural information about the concept orders specified in the curriculum as well as students' choices in concept coverage and learning order. By assigning tracing attributes to the nodes, graphs can also incorporate rich information on student learning status for each specific concept, which is critical for identifying when and where some students are struggling. As such, graph representation learning can account for the differences across various aspects of students' learning journeys.

Concretely, in this study, we explore the use of *InfoGraph*, a graph-level representation learning technique [6], to profile learner behaviors and performance for curriculum-based learning in ITSs. Each student's behaviors and performance are encoded with a vector representation via graph modeling in a self-supervised training manner. This eliminates the need for specifying rules and manual labeling by educators.

We summarize our contributions as follows:

1) We propose `CTGraph`, a Curriculum learning status Tracing approach using Graph-level models in ITSs. Based on *InfoGraph* [6], our approach can effectively capture student behavioral attributes, reflecting their variations in learning paths as aligned to the curriculum structure.

2) We empirically show that `CTGraph` can provide a holistic evidence-based view of student learning status in its latent space, which captures different interacting factors of student behaviors and performance. Our experiments demonstrate that such latent representations are effective in identifying different groups of students from multifaceted aspects and pinpointing students who are struggling.

3) We discover that the latent representations obtained via `CTGraph` can be used to identify student cohort groups, in which student behaviors are mostly similar but exhibit subtle differences in certain behavioral attributes, contributory factors associated with performance, or learning paths. Our experiments show that this method enables comparative analysis among different cohort student groups with fine-grained details to understand when and where students are struggling.

To our knowledge, `CTGraph` is the first self-supervised learning approach for student profiling in curriculum-based ITSs. The goal of `CTGraph` is to provide representations of learning behaviors and performance that can reflect students' learning paths as aligned with the contextual structure of the curriculum. These representations can automatically identify students who are struggling and lagging behind in certain aspects for curriculum-based learning systems. We demonstrate the superiority of `CTGraph` through extensive experiments on various learning topics using a real-world dataset. Moreover, our in-depth analysis also shows that the latent representations learned by `CTGraph` are effective and interpretable for providing an objective measurement of learning status across diverse dimensions.

II. RELATED WORK

Student Profiling. Profiling student behaviors to categorize students into different personas has received growing attention in educational AI research. Its applications include identifying at-risk learners [7]–[9], generating personalized feedback [7], refining the design of learning experiences [10] and informing education priorities [3]. Nevertheless, most recent studies in this line use clustering approaches with metrics based on handcrafted formulation of contributory factors associated with learning. For example, Soussia et al. [7] and Mojarad et al. [11] employed k-means to cluster groups of students with similar performance and behavior characteristics based on specified rules that define consistency, pace and effort. Domenzai et al. [12] used spectral clustering to identifying student profiles based on specified features regarding regularity, control, proactivity, etc.

However, when such handcrafted formulations are applied to large datasets with graphs representing curriculum structures, they can lead to very high dimensional, sparse and non-smooth representations and thus yield poor generalization [13]. In addition, as noted in prior studies ([4], [14]), profiling individual differences and distinct behavior attributes is also important for educators to gain insights into students' learning journeys. However, current clustering approaches categorize students into one specified persona type. Therefore, they may overlook the impact of subtle behavioral differences among individual students.

Graph Neural Network Representation Learning. Graph Neural Networks (GNNs) learn informative representations for nodes or entire graphs through information propagation on graph-structured data. Given the versatility of graph modeling, GNN variants have found success in many diverse applications such as modeling users and personal interests for recommendation [15]–[17], user targeting [18], and graph-based service search [19].

Leveraging the effectiveness of GNNs, our approach jointly characterizes students' historical behaviors and performance while taking into account the contextual structure of the curriculum. To enhance the discriminative capabilities of graph models, we exploit Graph Isomorphism Networks (GINs) [20] in the architecture. GINs can effectively distinguish different structures of graphs at a level similar to the power of the Weisfeiler-Lehman graph isomorphism test.

In addition, as is common in online applications, many real-world datasets exhibit long-tail distributions. In our scenario, the distribution of student-concept interactions also exhibits such skewness, which can compromise the model's generalisation performance. Recent studies show that contrastive learning methods can augment models' generalization capability when learning representations of long-tail distributions [21]. Inspired by this insight, our approach incorporates contrastive learning and several data processing strategies to facilitate contrastive learning during the GIN training phase, which effectively enhance our model's generalization ability.

III. PRELIMINARIES

In this section, we introduce the definitions of preliminary terms used in our framework.

Definition 1 (Curriculum-Structure Graph): A curriculum-structure graph $G_\phi = \{V_\phi, E_\phi\}$ is a directed acyclic graph for the topic ϕ. The node set V_ϕ denotes the set of concepts $\Theta = \{\theta_1, \theta_2, ..., \theta_t\}$ that are included in the topic ϕ. Each edge $e \in E_\phi$ represents the sequential order between two concept nodes, which specifies the prerequisite relationship as defined in the curriculum.

Definition 2 (Multivariate Learning-Tracing Vector): A multivariate learning-tracing vector for the student u is a vector concatenating a sequence of attribute vectors regarding u's behaviors and performance on all attempted questions related to the concept θ. It is represented as $\mathbf{x}^\theta(u) = \mathbf{x}_1^\theta(u) \oplus \mathbf{x}_2^\theta(u) \oplus ... \oplus \mathbf{x}_\mathbf{m}^\theta(u)$, where $\mathbf{x}_\mathbf{i}^\theta(u) \in \mathbb{R}^\ell$ is a vector with a variable length ℓ, which describes one tracing attribute related to student behavior or performance on the concept θ in a numeric vector format.

Definition 3 (Student Curriculum-Based Learning Graph): A student curriculum-base learning graph is a directed graph $G_\phi^u = \{V_\phi^u, E_\phi^u\}$ that represents the learning status and learning path for student u on the topic ϕ. The node set V_ϕ^u denotes the set of concepts covered by student u's attempted questions in the ITS. Each node v in the graph G_ϕ^u is associated with a multivariate learning-tracing vector as the node attributes, which represents student u's learning behaviors and performance regarding the concept θ. The structure of G_ϕ^u is derived from the curriculum-structure graph G_ϕ. The directed edge $e = (v_i, v_j)$ in G_ϕ^u denotes that v_j is the nearest successor of v_i in G_ϕ.

IV. Framework

Our approach focuses on using graph neural networks (GNNs) to model students during curriculum-based learning. Graphs models can explicitly represent the curriculum structure and the details of individual students' learning behaviors and performance on each concept. By employing graph models, we encode all these details into a graph-level representation for each student.

In this section, we present the framework for our approach. We will begin with the overview of our framework and then discuss the framework design in details.

A. Framework Overview

ITS Data Preprocessing. We collected data from the ITS, which includes 1) logs of each student's learning records that trace student behaviors and performance, and 2) the curriculum-structure graphs. In this study, for each concept and each student, we extract tracing attributes to compose the multivariate learning-tracing vector $\mathbf{x}^\theta(u) = \mathbf{x}_1^\theta(u) \oplus \mathbf{x}_2^\theta(u) \oplus \mathbf{x}_3^\theta(u)$. $\mathbf{x}_1^\theta(u)$ represents the average accuracy of all attempts made by student u on questions related to the concept θ, $\mathbf{x}_2^\theta(u)$ denotes the total number of such attempts, and $\mathbf{x}_3^\theta(u)$ corresponds to the median week number in the academic calendar when student u attempted these questions in the ITS. It is worth noting that various other attribute types (e.g., temporal embedding via a transformer) can also be integrated into the tracing vector, capturing more nuanced details of student behaviors from different perspectives.

Graph Construction. For each topic ϕ, we construct a curriculum-based learning graph $G_\phi^u = \{V_\phi^u, E_\phi^u\}$ for each student using *Node absorption* (see details in the next subsection IV-B1). G_ϕ^u captures the details of student u's learning behaviors and performance on the topic ϕ. The graph structure reflects student u's learning path in terms of concepts. The node attributes, represented by the multivariate learning-tracing vectors, capture the details of u's learning behaviors and performance. These attributes reflect the learning proficiency on each concept in the topic ϕ, as well as the learning intensity, and the aggregated temporal information indicating the specific learning time in terms of the academic calendar.

Graph-Level Encoding with InfoGraph. Our goal for graph-level encoding is to represent the rich information about student learning behaviors and performance in student curriculum-based learning graphs with a fixed-length vector. Such vectorized representation can be utilized for student profiling, which enables intuitive visualization and pattern discovery for enhancing educators' understanding of student learning journeys in the ITS.

In this study, we leverage *InfoGraph* for the graph-level encoding procedure. *InfoGraph* learns graph-level representations by maximizing the mutual information between the global graph-level representation and the local patch representations of subgraph structures across different scales centered at each code in the graph. As such, the resultant encoding not only captures the global structure of learning paths but also the properties of tracing attributes associated with the concepts.

The flexibility of *InfoGraph* makes our framework suitable for both relatively small classes with hundreds of students and large classes with thousands of students. It can also accommodate curriculum structures with topics' coverage ranging from a dozen concepts to hundreds of concepts.

Latent-Space Representation Extraction and Visualization. By extracting the vectorized representation from graph-level encoding models, we can explore student representations together in one view (visualization) inside the compressed 3D latent space. The visualization is derived via PCA compression. With the multivariate tracing attributes as probing indicators, the latent representations provide a comparative visualization of each student's behaviours and performance, which highlights the overall performance achieved by each student as well as the aggregated information that summarize their learning behaviors.

More critically, students positioned closely together in latent representation space typically selected similar learning paths on the concepts and have comparable performance. By locating learners in the latent space, we can identify different cohorts of students who vary in the choice of learning paths, as well as in their learning intensity and timing.

B. Diving Into The Framework Design Details

In this subsection, we discuss the details of our design in the framework.

1) Node Absorption: Due to the abundance of question sets available, most student attempted only a subset of problems covering specific concepts within the curriculum. Consequently, for many of these students, there are no logs for certain concepts in the curriculum-structure graph. Graphs with a high number of null values in the node attributes pose challenges for contrastive learning. This is because they lead to a non-smooth distribution over node attributes, resulting in poor generalization performance [22]. To encourage rich and discriminative representations, for each student, we apply a process named *node absorption*. Specifically, for each student, this process removes concept nodes that lack learning logs within the curriculum-structure graph. Nodes are subsequently reconnected if they are the nearest successors to other nodes within this structure. Through this method, we construct the student's curriculum-based learning graph for each student,
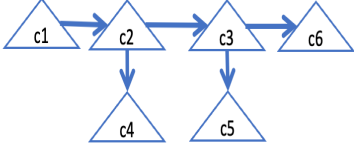
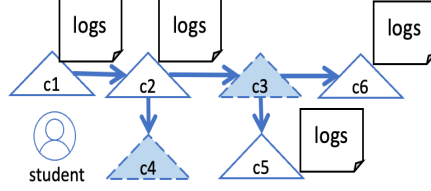Fig. 2: Curriculum-Structure Graph
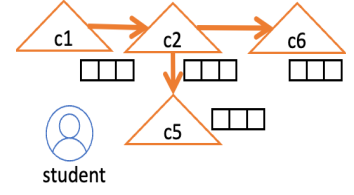

Fig. 3: Node Absorption

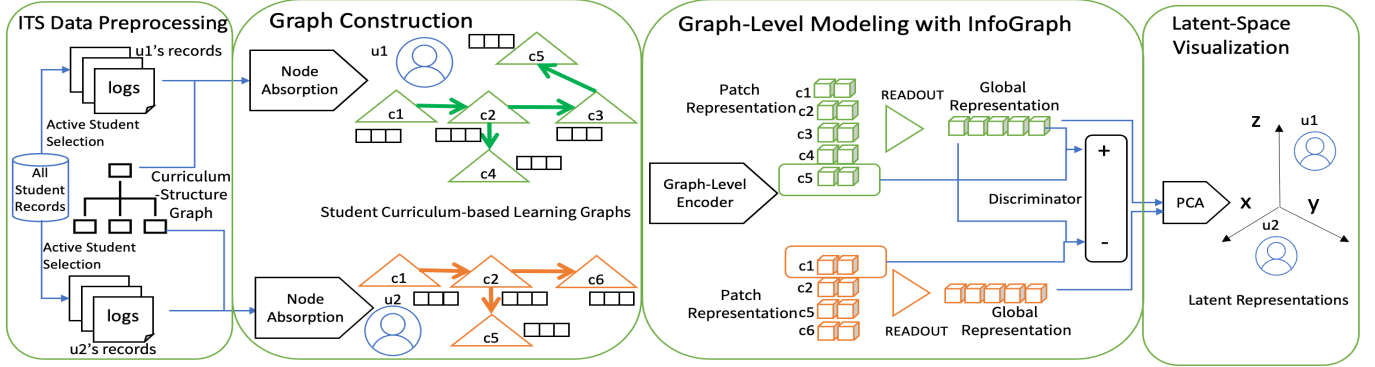Fig. 4: Student Curriculum-Based Learning Graph

Fig. 5: Overview of the `GTGraph` Framework

which uses multivariate learning-tracing vectors as node attributes and eliminates any null values in the attribute vectors.

Figure 3 illustrates an example of node absorption, which is based on the curriculum-structure graph in Figure 2. In this example, the student in Figure 3 did not attempt any questions related to concepts $c3$ and $c4$. Therefore, we first remove the concept nodes $c3$ and $c4$ from the graph. Subsequently, we reconnect the remaining nodes if they are the nearest successors in the corresponding curriculum-structure graph (as depicted in Figure 2). The resulting student curriculum-based learning graph is depicted in Figure 4.

Such graph construction design is critical, which ensures graph-level representations remain discriminative towards other graph instances.

2) Student Selection with Medium-High Concept Coverage Attainment Threshold: For model training, we select students who attempted the questions covering at least a certain percentage of the total number of concepts for a given topic. This is because we encourage our model to learn diverse learning paths undertaken by students. By focusing on these students, we can ensure that each student had sufficient engagement with the ITS and studied questions covering a range of concepts that form diverse learning paths. More importantly, *InfoGraph* is based on contrastive learning during training. Contrastive-learning methods often require a large number of negative samples to be effective [23]. Thus, selecting students with diverse learning paths to generate different negative samples during training is crucial, since learning graph embeddings requires many different graph instances during training.

V. METHODOLOGY

In this section, we formulate the graph representation learning problem formally and present the details of the graph-level encoding and representation learning method.

A. Problem Definition

Graph-Level Representation Learning. Given a set of student curriculum-based learning graphs $\mathbb{G}_\phi = \cup_u^{\mathcal{U}} \{G_\phi^u\}$ on the topic ϕ for all students $\mathcal{U}$ and a positive integer d (the expected embedding size of the representation), our goal is to learn a d-dimensional representation $\mathbf{y}$ for each graph $G_\phi^u \in \mathbb{G}$, where $\mathbf{y}$ is a fixed-length vector $(y_1, y_2, ..., y_d) \in \mathbb{R}^d$.

B. Graph-Level Modeling with InfoGraph

In our approach, we utilize *InfoGraph* to obtain the graph-level representation by maximizing the mutual information between the global representation and the local patch representation. Next, we will illustrate each component of the model separately.

Graph-Level Encoder. We choose Graph Isomorphism Network (GIN) [20] as the graph convolution encoder in the model. The key distinguishing feature of GIN is that it can differentiate graphs that are not isomorphic to each other and hence achieve maximum discriminative power compared to other GNNs. The details of the graph-level encoder are illustrated in Figure 8.

Patch Representation. To obtain the local patch representation, GIN passes message iteratively by updating each concept node's representation with aggregations of its neighbors' representations. After k iterations, the patch representation captures
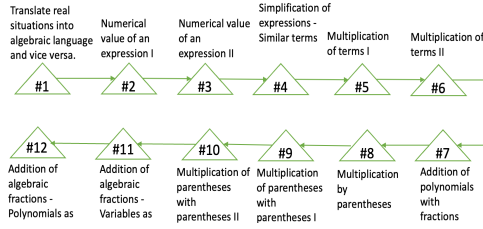
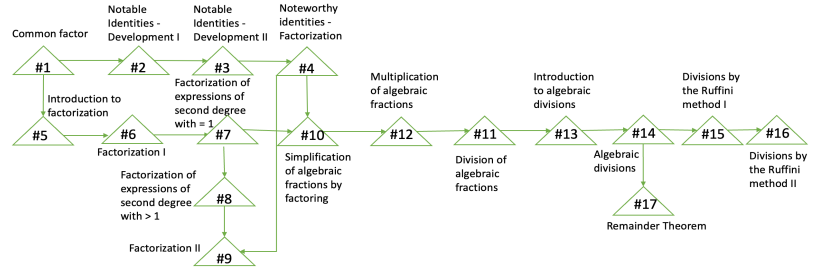**Fig. 6: Curriculum-Structure Graph for the Topic** *Algebra I*

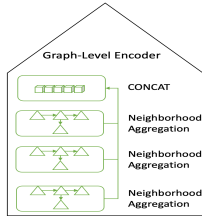**Fig. 7: Curriculum-Structure Graph for the Topic** *Algebra II*

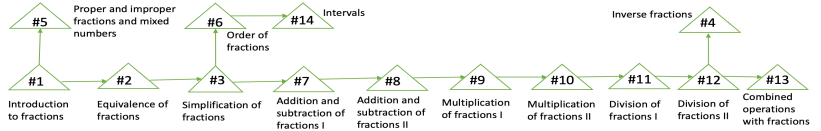**Fig. 8: Graph-Level Encoder Architecture**

Fig. 9: Curriculum-Structure Graph for the Topic *Fractions*

the rich information within its k-hop neighborhood subgraph. Specifically, the update function and the aggregation scheme of the k-th layer in the framework of GIN are as follows:

$$a_{v_i}^{(k)} = \text{AGGREGATE}^{(k)}(h_{v_j}^{(k-1)} : v_j \in \mathcal{N}(v_i))),$$
$$h_{v_i}^{(k)} = \text{COMBINE}((h_{v_i}^{(k-1)}, a_{v_i}^{(k)})$$

where $a_{v_i}^{(k)}$ is the aggregated representation of the concept node v_i's neighbors $\mathcal{N}(v_i)$ and $h_{v_i}^{(k)}$ is the patch representation of node v_i after the k-th layer in the graph encoder. $h_{v_i}^0$ is initialized with the multivariate learning-tracing feature vector. **Global Representation.** We utilize a READOUT function that summarizes all concept nodes' patch representations into a fixed-length vector to obtain the graph-level global representation. That is,

$$h_\psi^{v_i} = \text{CONCAT}(\{h_{v_i}^{(k)}\}_{k=1}^K),$$
$$H_\psi(G_\phi^u) = \text{READOUT}(\{h_\psi^{v_i}\}_{i=1}^N)$$

where ψ denotes the set of parameters of a K-layer graph neural network, $h_\psi^{v_i}$ is the aggregated patch representation summarizing the neighborhood subgraphs centred at the concept node v_i, and $H_\psi(G_\phi^u)$ is the global graph-level representation for the student curriculum-based learning graph G_ϕ^u. In this study, we use sum over all the N concept nodes' patch representations as the READOUT function.
Contrastive Learning. *InfoGraph* maximizes the mutual information (MI) between global and local representations, which is estimated with the discriminator $\mathcal{T}_\beta$ using contrastive learning. $\mathcal{T}_\beta$ is modeled by a feedforward neural network with parameters β. To estimate the maximized MI in practice, we use all possible combinations of global and local patch representations to generate negative samples across all graph instances for contrastive learning, as indicated in *InfoGraph* [6]. The MI estimator, $I_{\psi,\beta}$ maximizes the estimated MI on

all the global/local pairs over the given dataset $\mathbb{G}_\phi$. Formally, for each topic ϕ, the objective is to maximize $I_{\psi,\beta}$ as below:

$$\hat{\psi}, \hat{\beta} = \underset{\psi,\beta}{\text{argmax}} \sum_{G_\phi^u \in \mathbb{G}_\phi} \frac{1}{|G_\phi^u|} \sum_{v_i \in G_\phi^u} I_{\psi,\beta}(h_\psi^{v_i}; H_\psi(G_\phi^u))$$

VI. EXPERIMENTS

In this section, we discuss our analysis and findings of the graph-level representations learned by CTGraph.

A. Datasets

We conduct experiments on a real-world dataset provided by Adaptemy[1]. This dataset was collected from Adaptemy's ITS, which aids students in curriculum-based mathematical learning.

The full curriculum comprises 26 topics in total. We present the results for four primary topics in this curriculum: *Algebra I*, *Algebra II*, *Functions I* and *Fractions*. The curriculum structure for each topic is illustrated in Figures 1a, 6, 7 and 9. As shown, each topic contains concepts (i.e., knowledge items) and the directed edges represent prerequisite conditions between concepts. A concept is a unit of learning. In the ITS, students can access each concept independently. Therefore, when engaging with the ITS, they are not required to follow a predetermined learning path as depicted in the curriculum.

From the original data records of the learning traces, we extract the tracing attributes related to student behaviors and performance. These attributes include the student's average accuracy for each concept within the topic (i.e., the mean value of scores for all attempted questions covering each concept), the number of attempts (i.e., the total count of times a student attempted the questions on each concept), and the median week number in the academic calendar year when

[1]https://www.adaptemy.com/

the student studied each concept. It is worthy noting that, while the attributes used in this study are simple aggregated values based on learning traces, more sophisticated attributes, such as temporal features and concept embedding, can also be incorporated without altering the modeling framework.

Details of the datasets can be found in Table I.

TABLE I: Dataset Statistics

Topic	# Students	# Students with Medium-High Concept Coverage	# Concepts	# Interactions with Questions
Algebra I	9689	1415	12	65675
Algebra II	4106	290	17	32697
Functions I	4648	419	7	24153
Fractions	9243	867	14	63632

B. Experiments Configuration

For each topic, we select students who covered at least 50% of the concepts, indicating medium to high concept coverage attainment in the curriculum structure. The graph-level encoder used in the experiments consists of three hidden layers, each with a hidden dimension of 32. As a result, the global graph-level representation has a dimensionality of 96. We trained all models using the Adam optimizer with a learning rate of 0.01 and a batch size of 128.

C. Analysis

In the following analysis, we seek to answer the following questions:

- Q1: Can the representations in the latent space effectively encode a student's learning path, behaviors, and performance in the curriculum-based ITS? How can these representations be used to monitor student performance and behaviors, as well as identify those lagging behind?
- Q2: How can we use the representations to identify equivalent groups of similar learners, allowing educators to spot commonalities in the behaviors of students who are lagging behind?
- Q3: Can the representations capture the subtle differences in behaviors and learning paths among the students who are lagging behind? Will these differences reveal when and where the students began to lag behind in specific aspects? If so, how can we find students with subtle differences in the latent space?

1) Representations in the Latent Space: For each topic, we extract the latent representation in a 96-dimensional vector for each student. We then derive the 3D visualization by further compressing the latent representations to three dimensions using PCA compression.

Figures 10 to 13 display the 3D visualization of latent spaces for the four topics. As shown in the figures, most students cluster closely together in the latent space, while a small number of outliers are positioned further away from the central clusters. Notably, in Figures 12a to 12c for the topic *Functions I*, the distribution of latent representations clearly show two

clusters. This indicates the behaviors and performance of students in the smaller cluster differ significantly from the students in the larger cluster, as particularly evident in the median week number illustrated in Figure 12c.

Note that the extracted vector representation encompasses tracing attributes reflecting features of learning behaviors, such as proficiency, intensity, and the timing of when the student attempted questions covering certain concepts within the topic. To demonstrate that representations effectively encode behaviors and performance across different dimensions simultaneously, we use color in each figure to highlight the aggregated values of each tracing attribute for all the concepts in a student's curriculum-based learning graph. For example, in Figure 10a, the color represents a student's average accuracy across all the attempts on all the concepts for the topic *Algebra I*. In Figure 10b, the color indicates the total number of attempts on all the concepts, while in Figure 10c, the color corresponds to the median of the median week numbers regarding all the concepts a student studied in this topic.

It is important to note that, those aggregated values of tracing attributes are for illustration purposes only and are not used during model training. As clearly shown in the figures, color gradients in Figures 10 to 13 represent the distribution of these three types of attributes in the latent space, providing evidence that the latent representations effectively encode these tracing attributes simultaneously.

The positioning of each student's representation in the latent space reveals a smooth distribution of different learning behaviors and performance (in terms of tracing attributes and learning paths). When we locate students in the latent space, it is evident that those clustered together achieved similar performance, exhibited comparable tracing attributes and followed similar learning paths. Specifically, in terms of performance, the students identified as outliers typically have an average accuracy in their attempts that is much lower than that of the majority. As such, we can identify students lagging behind by locating these outliers in the latent spaces.

2) Similar Students: Now we illustrate how to utilize the latent spaces from `CTGraph` for identifying similar students who are lagging behind at certain aspects.

We use the topic *Algebra I* as an example. First, we select one student u, whose representation is located far away from the majority in the 3D compressed latent space associated with the topic *Algebra I*. Next, we identify the neighboring students around u by locating those with the closest cosine distance to u in the original 96-dimensional latent space. Figures 14a to 14c show the aggregated statistics for student u (located in the center of the figure with attribute values highlighted in bold and a larger font) and the top 20 neighboring students surrounding u (color-coded with a gradient from purple to yellow, indicating their cosine distance in the original latent space from smallest to largest). Across all three aspects, it's evident that the nearest neighboring students (highlighted in purple) closely resemble student u. The remaining students (colored in orange and yellow) also bear similarities to u in certain aspects. When selecting students further from u in the
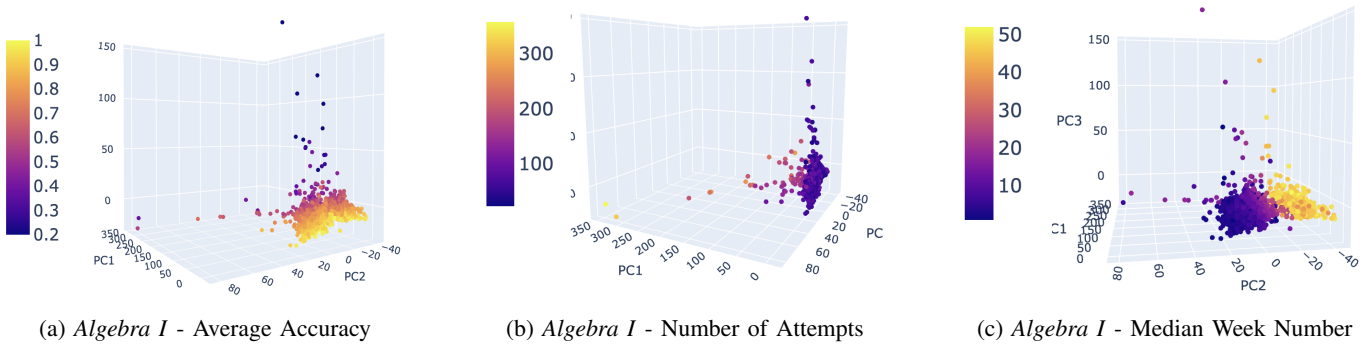
(a) *Algebra I* - Average Accuracy   (b) *Algebra I* - Number of Attempts   (c) *Algebra I* - Median Week Number

Fig. 10: PCA Representations in the Latent Space for the Topic *Algebra I*

(a) *Algebra II* - Average Accuracy   (b) *Algebra II* - Number of Attempts   (c) *Algebra II* - Median Week Number

Fig. 11: PCA Representations in the Latent Space for the Topic *Algebra II*

(a) *Functions I* - Average Accuracy   (b) *Functions I* - Number of Attempts   (c) *Functions I* - Median Week Number
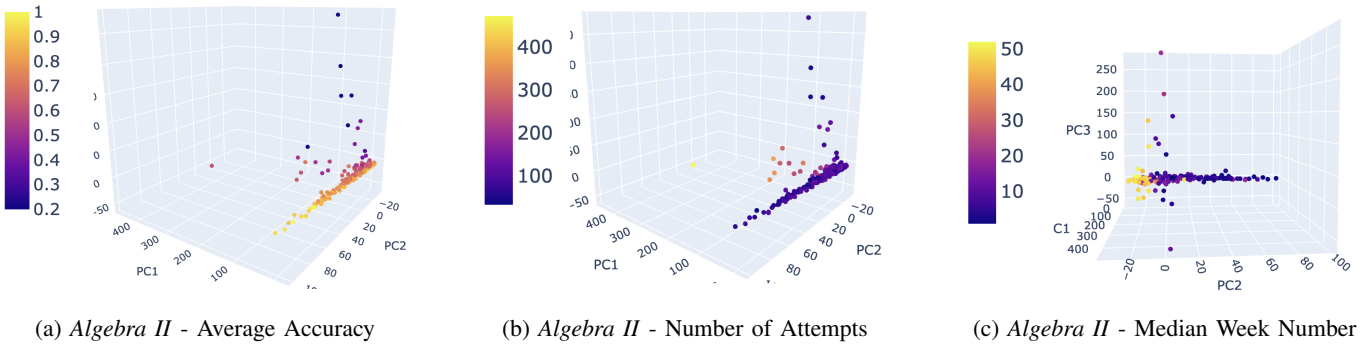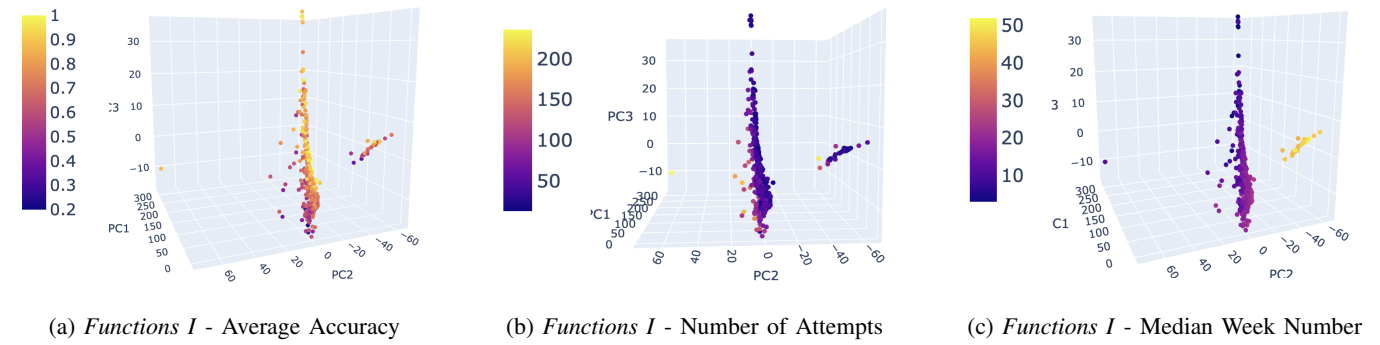
Fig. 12: PCA Representations in the Latent Space for the Topic *Functions I*

(a) *Fractions* - Average Accuracy   (b) *Fractions* - Number of Attempts   (c) *Fractions* - Median Week Number
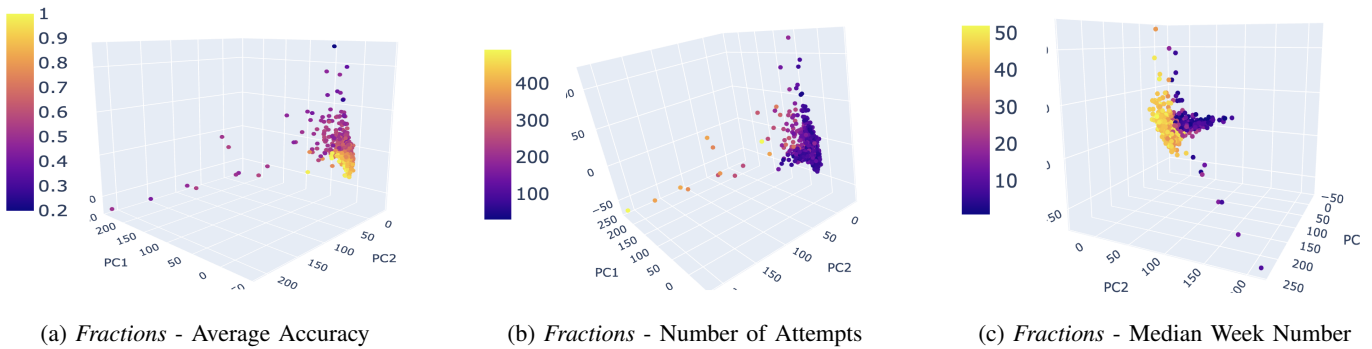
Fig. 13: PCA Representations in the Latent Space for the Topic *Fractions*

(a) Average Accuracy,
Top 20 Similar Students

(b) Number of Attempts,
Top 20 Similar Students

(c) Median Week Number,
Top 20 Similar Students

(d) Average Accuracy,
Top 80 Similar Students

(e) Number of Attempts,
Top 80 Similar Students

(f) Median Week Number,
Top 80 Similar Students

Fig. 14: Similar Students in the Latent Space for the Topic *Algebra I*

(a) Student 1

(b) Student 2

(c) Student 3

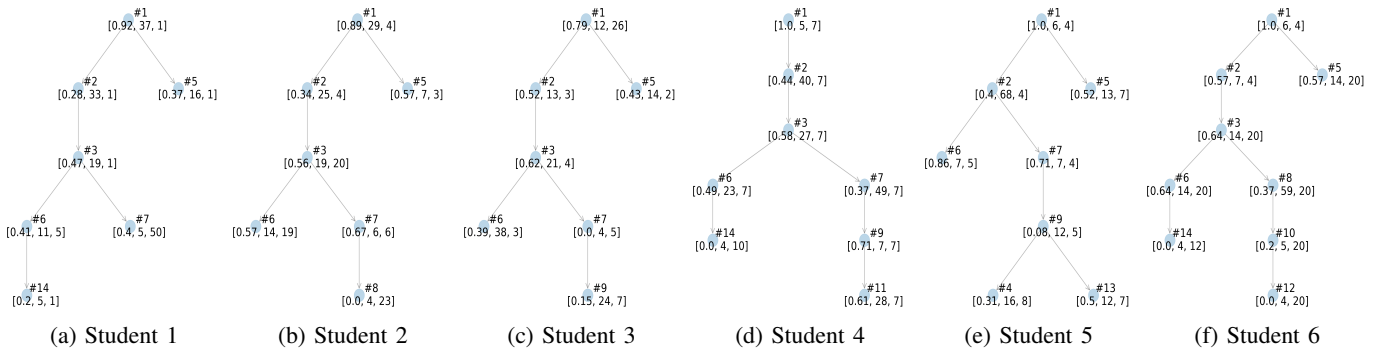(d) Student 4

(e) Student 5

(f) Student 6

Fig. 15: Top 5 Neighboring Students of Student 1 in the Topic *Fractions*

latent space, it is apparent that these individuals significantly differ from u in tracing attributes compared to those who are closer.

Furthermore, in Figures 14d to 14f, we show the top 80 neighboring students of student u. Notably, the students to the right of u, who are also distant from the majority of students, are now included. The aggregated values of these neighboring students' tracing attributes slightly differ from u. As shown in the figures, the representations reveal that these students on the right often have much lower average accuracy in their attempts compared to u, and they generally made fewer attempts than u or tackled these questions at different weeks in the academic calendar.

The aggregated statistics of the tracing attributes for each topic can provide an overview of a student's learning status.

However, for a detailed understanding of student behaviors, performance and learning paths, it is essential to examine the student curriculum-based learning graphs more thoroughly. To further demonstrate that CTGraph captures both the statistical properties of students' learning status and the details of their learning paths relative to the curriculum, we provide an example from the topic *Fractions*. Figure 15 shows the top 5 neighboring students for student 1 in the original latent space (arranged from the nearest to furthest in Figures 15b to 15f). Figures 15a to 15f present the curriculum-based learning graph for each student. It's evident that all five students generally have low accuracy (around 0.5) in their attempts.

Next, we demonstrate that how to examine the learning graphs of similar students to identify the commonality among students lagging behind. The learning graphs explicitly show

the specifics of each student's choices in terms of concept order, intensity, and timing of the attempts. As shown in Figure 15, each student and their nearest left/right neighbors have similar attributes on each concept and their learning paths also demonstrate similarity. Furthermore, it is evident that these students generally achieved high accuracy on Concept 1 but significantly lower accuracy on immediate subsequent concepts (i.e., Concept 2 and Concept 5). After finishing Concept 6, they either skipped all other concepts after on that branch or achieved much lower accuracy (e.g., on Concept 14). In this way, educators can easily find out when and where these students started to fall behind based on the curriculum structure.

3) Comparative Analysis: In this part, we first illustrate the use of latent representation to identify cohort student groups, which are students that exhibit predominantly similar behaviors but also have subtle differences in certain aspects. We then use two different groups as examples and conduct comparative analysis within each group separately.

TABLE II: Comparative Analysis for *Functions I*: Aggregated Values of Tracing Attributes

Student ID	Average Accuracy	# Concepts	# Attempts	Median Week No.
1	0.48	4	103	3
2	0.63	4	144	8
3	0.49	4	83	3
4	0.60	5	118	8
5	0.56	6	203	13

TABLE III: Comparative Analysis for *Algebra II*: Aggregated Values of Tracing Attributes

Student ID	Average Accuracy	# Concepts	# Attempts	Median Week No.
1	0.52	10	179	6
2	0.61	11	113	8
3	0.63	10	122	6
4	0.61	9	96	8
5	0.63	9	93	5
6	0.49	9	119	2
7	0.68	10	110	4
8	0.70	10	90	6
9	0.63	12	88	15
10	0.62	13	106	7

To derive a cohort group, we first select a student as the starting point in the latent space and another student as the ending point. Let s and e denote the representation vectors of these two students. Secondly, we compute vector $v = s - e$. We then search the original latent space for students whose positions are closest to the vector v, as measured by the cosine distance. The top k closest students found in this way, along with the two endpoint students, form a cohort group.

First, we show one example for the topic *Functions I*, in which we demonstrate a cohort group where the students follow similar learning orders on the concepts. Table II shows the aggregated values of tracing attribute for each student in such group, and Figure 16 shows their curriculum-based learning graphs. It is evident that the behaviors of all five

students are largely similar in terms of sequence of learning concepts, average accuracy, the total number of attempts, and timing of learning. However, there are variances in the coverage of concepts as well as subtle difference in accuracy on certain concepts. For instance, Student 4 and Student 5 achieved higher accuracy for Concept 1 and Concept 2 but lower accuracy for the additional concepts they tackled later. This suggests that even though all these students have comparable average performance (in terms of average accuracy) as compared to Student 1-3, Students 4 and 5 actually made significant more effort on advanced concepts. Thus, despite their lower scores on these advanced concepts (i.e., where they started to lag behind), their stronger performance on basic concepts (Concepts 1-2) and efforts to broaden their knowledge base should be highlighted in the assessment of their achievement.

Secondly, we provide another example for the topic *Algebra II*, in which the curriculum structure is more intricate. Table III presents the aggregated values of tracing attributes for each student in a specific cohort group, and Figure 17 depicts their curriculum-based learning graphs. To compare with the majority of students, we check details in the 3D visualization view for the topic *Algebra II* in Figures 11a to 11c. As shown in Figure 11a, compared to the majority of students, students in this cohort group generally achieved much lower average accuracy in their attempts. However, their total numbers of attempts are comparable to the majority.

Furthermore, when compared to the other students in this group, there is also a noticeable expansion in the coverage of concepts for the last two students (i.e., Student 9 and 10). Analyzing their curriculum-based learning graphs in Figure 17i and 17j, we observe that 1) they performed better for the first 12 concepts in the topic's curriculum structure, and 2) their accuracy for the remaining concepts (i.e., Concepts 13-17) they tackled—compared to the rest of the group—is considerably lower. This suggests that these two students exerted greater effort in exploring advanced concepts, even if their performance on these advanced concepts is lower and their average performance remains consistent with the others in the cohort group.

VII. CONCLUSION

In this study, we present `CTGraph`, a graph-level representation learning approach to profile learner behaviors and performance in a self-supervised manner. Our analysis shows the importance of considering the nuanced differences in student learning paths and multivariate metrics related to their behaviors and performance. The intricate details of students' curriculum-based learning graphs reveal the depth and breadth of their learning journeys. Notably, students may achieve similar overall performance but diverge in their engagement with advanced concepts or their mastery of foundational knowledge. This nuanced perspective emphasizes the value of holistic evaluation, recognizing students' efforts and progress in the context of the curriculum structure. As such, our approach can provide educators, pedagogy researchers and
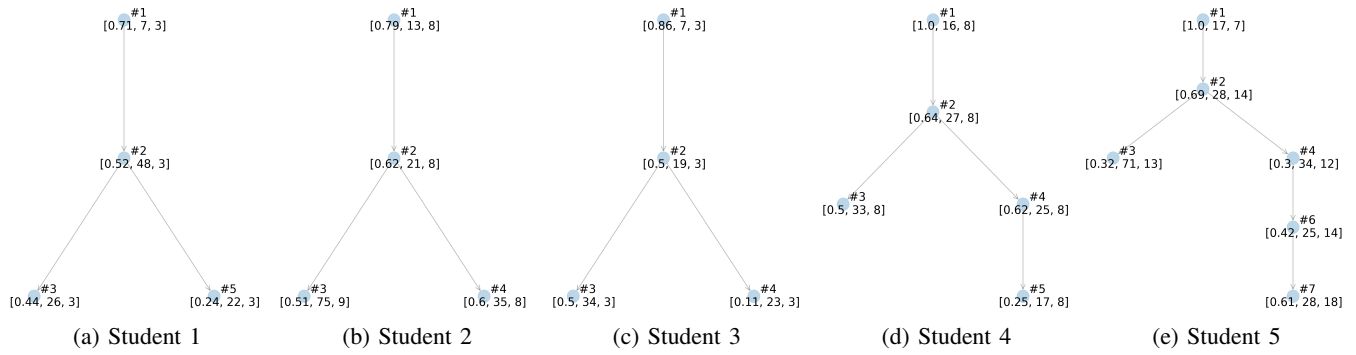
Fig. 16: Comparative Analysis for the Topic *Funtions I* - Student Cohort and Their Curriculum-Based Learning Graphs
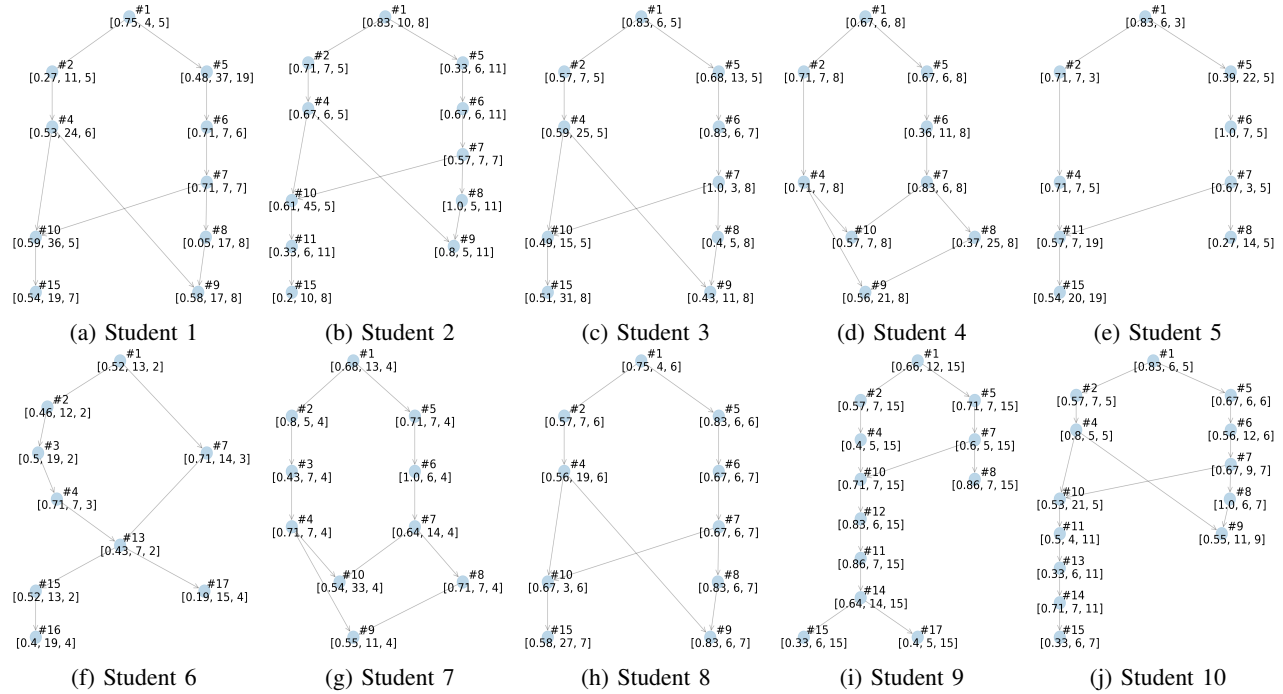

Fig. 17: Comparative Analysis for the Topic *Algebra II* - Student Cohort and Their Curriculum-Based Learning Graphs

curriculum designers with a rich understanding of individual students' learning journeys and the learning effectiveness of various learning strategies and paths. Thereby, it opens new opportunities to develop and design more tailored and effective pedagogical strategies for ITS applications.

REFERENCES

[1] A. Lane, "Open education and the sustainable development goals: Making change happen," *Journal of Learning for Development*, vol. 4, no. 3, 2017.

[2] K. Holstein and S. Doroudi, "Equity and artificial intelligence in education: Will" aied" amplify or alleviate inequities in education?" *arXiv preprint arXiv:2104.12920*, 2021.

[3] L. Crouch, M. Kaffenberger, and L. Savage, "Using learning profiles to inform education priorities: An editors' overview of the special issue," *International Journal of Educational Development*, vol. 86, p. 102477, 2021.

[4] I. Suto, "Creating cambridge learner profiles: A holistic framework for teacher insights from assessments and evaluations." *Research Matters*, vol. 35, pp. 6–26, 2023.

[5] H. Ibrahim, F. Liu, R. Asim, B. Battu, S. Benabderrahmane, B. Alhafni, W. Adnan, T. Alhanai, B. AlShebli, R. Baghdadi *et al.*, "Perception, performance, and detectability of conversational artificial intelligence across 32 university courses," *Scientific Reports*, vol. 13, no. 1, p. 12187, 2023.

[6] F.-Y. Sun, J. Hoffman, V. Verma, and J. Tang, "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization," in *International Conference on Learning Representations*, 2019.

[7] A. Ben Soussia, C. Treuillier, A. Roussanaly, and A. Boyer, "Learning profiles to assess educational prediction systems," in *International Conference on Artificial Intelligence in Education*. Springer, 2022, pp. 41–52.

[8] M. Marras, J. T. T. Vignoud, and T. Kaser, "Can feature predictive power generalize? benchmarking early predictors of student success across flipped and online courses," in *14th International Conference*

on Educational Data Mining, 2021, pp. 150–160.

[9] A. Ben Soussia, A. Roussanaly, and A. Boyer, "An in-depth methodology to predict at-risk learners," in *European Conference on Technology Enhanced Learning*. Springer, 2021, pp. 193–206.

[10] R. M. Quintana, S. R. Haley, A. Levick, C. Holman, B. Hayward, and M. Wojan, "The persona party: using personas to design for learning at scale," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017, pp. 933–941.

[11] S. Mojarad, A. Essa, S. Mojarad, and R. S. Baker, "Data-driven learner profiling based on clustering student behaviors: learning consistency, pace and effort," in *International conference on intelligent tutoring systems*. Springer, 2018, pp. 130–139.

[12] P. Mejia-Domenzain, M. Marras, C. Giang, and T. Käser, "Identifying and comparing multi-dimensional student profiles across flipped classrooms," in *International Conference on Artificial Intelligence in Education*. Springer, 2022, pp. 90–102.

[13] P. Yanardag and S. Vishwanathan, "Deep graph kernels," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1365–1374.

[14] P. A. Alexander and P. K. Murphy, "Profiling the differences in students' knowledge, interest, and strategic processing." *Journal of educational psychology*, vol. 90, no. 3, 1998.

[15] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian, and X. Xie, "Self-supervised graph learning for recommendation," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021, pp. 726–735.

[16] C. Gan, B. Hu, B. Huang, T. Zhao, Y. Lin, W. Zhong, Z. Zhang, J. Zhou, and C. Shi, "Which matters most in making fund investment decisions? a multi-granularity graph disentangled learning framework," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2516–2520.

[17] A. El-Kishky, T. Markovich, S. Park, C. Verma, B. Kim, R. Eskander, Y. Malkov, F. Portman, S. Samaniego, Y. Xiao *et al.*, "Twhin: Embedding the twitter heterogeneous information network for personalized recommendation," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 2842–2850.

[18] D. Yang, B. Hu, X. Yang, Y. Shen, Z. Zhang, J. Gu, and G. Zhang, "Who would be interested in services? an entity graph learning system for user targeting," in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 2023, pp. 3248–3254.

[19] W. Wang, B. Hu, Z. Peng, M. Zhong, Z. Zhang, Z. Liu, G. Zhang, and J. Zhou, "Garcia: Powering representations of long-tail query with multi–granularity contrastive learning," in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 2023, pp. 3182–3195.

[20] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *International Conference on Learning Representations*, 2018.

[21] Y. Wei, X. Wang, Q. Li, L. Nie, Y. Li, X. Li, and T.-S. Chua, "Contrastive learning for cold-start recommendation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5382–5390.

[22] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, "graph2vec: Learning distributed representations of graphs," *arXiv preprint arXiv:1707.05005*, 2017.

[23] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations*, 2018.