Educational Digital Twin: Tackling Complexity in Educational Big Data

Luwen Huang
Department of Computer Science
University of Texas at Austin
luwen@cs.utexas.edu

Karen E. Willcox
Oden Institute for Computational Engineering & Sciences
University of Texas at Austin
kwillcox@oden.utexas.edu

***Abstract*—“Everything is bigger in Texas” and this includes the big data challenges of the state’s educational system. But just as big is the opportunity for digital twin technologies to improve decision-making in education, with potential to improve student outcomes. This paper formulates an Educational Digital Twin, a novel approach to understanding, modeling, and analyzing educational data to address the complexities inherent in student pathways. By “student pathways”, we refer to the joint facets of student behavior—evolving dynamically over time and encompassing not only course enrollments but also behavioral and academic factors. Unlike traditional approaches that rely on static analyses of historical data, the Educational Digital Twin applies the digital twin paradigm to model these pathways as a living construct that dynamically evolves alongside the physical world. At the heart of the Educational Digital Twin is a knowledge graph that organizes data into a semantic structure. By employing graph-theoretic methods, we manipulate this structure to derive multi-granular insights that inform decision-making. Our evaluation demonstrates how the Educational Digital Twin not only facilitates the dynamic integration of new data but also scales efficiently to handle complex datasets covering millions of students across hundreds of demographic and academic dimensions.**

***Index Terms*—Educational big data, Complex networks, Network visualization, Graph modeling, Digital twin**

I. Introduction

Academic trajectories are highly nonlinear and recognizing their time-dependent dynamic nature is essential to extracting useful data-driven insights—both at the system-wide and individual student levels. At community colleges, which comprise 42% of first-time freshmen in the United States [1], students take extremely diverse paths [2]–[4]. Students may drop out, change majors, or enroll in a diverse range of classes, leading to a variety of educational trajectories. We refer to these sequences of actions as *student pathways*, which are shaped by numerous factors, including demographics [5], coursework [6], and institutional characteristics [7]. These pathways are complex to navigate and model, posing challenges for students and researchers alike [8], [9].

Modeling student pathways is important because these pathways encode information that relates to a student’s eventual outcomes. For example, students completing required coursework within four semesters at a community college are more likely to transfer to a four-year institution than those spending six years [10]. Moreover, mapping pathways lets institutions predict behaviors and identify students at risk of dropping out early. However, modeling student pathways is challenging because of the dynamic nature of the data: Students carry shifting personal profiles, enroll in different institutions, register for different courses, obtain grades, and earn awards, while institutions frequently change course offerings and update requirements. Modeling single facets of these behaviors or treating the data as static fails to provide a comprehensive understanding of the system.

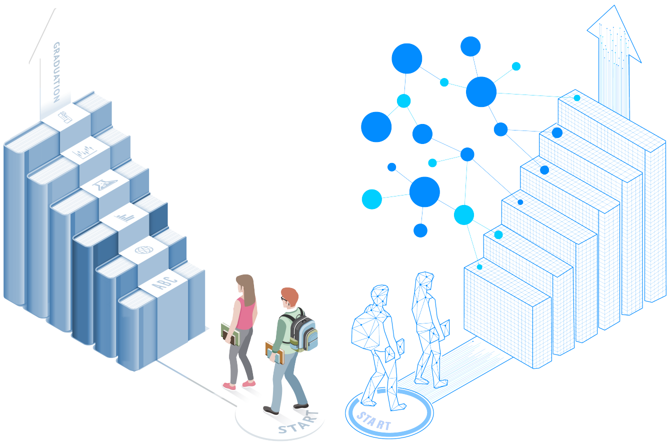


Fig. 1: The Educational Digital Twin, a paradigm for understanding, modeling, and analyzing student pathways and their dynamic evolution over time.

This paper models student pathways in public higher education on the scale of the entire state of Texas. This involves complexities characterized by the “5 V’s” of Big Data: (1) *Variety*, where a comprehensive view of student behaviors requires diverse data sources; (2) *Volume*, where hundreds of Texas public postsecondary institutions serve millions of students every year; (3) *Velocity*, where data have varied frequencies, from semester-based grade reports to to multi-year policy changes; (4) *Veracity*, where data reliability varies widely, from highly accurate course records to less reliable self-reported student intentions; and (5) *Value*, where, properly structured and modeled, the data give insights that support student interventions and policy making across Texas.

To address these Big Data challenges in educational pathways, we propose adopting the Digital Twin paradigm. A Digital Twin is "a set of virtual information constructs that mimics the structure, context, and behavior of a natural, engineered, or social system (or system-of-systems), is dynamically updated with data from its physical twin, has a predictive capability, and informs decisions that realize value." [11] Despite its growth in fields such as engineering and healthcare, the application of digital twins in education remains largely unexplored. Unlike traditional educational data analytics, a Digital Twin goes beyond just simulation and modeling, and as emphasized in [11], "the bidirectional interaction between the virtual and the physical is central to the digital twin."

We propose an approach toward an *Educational Digital Twin* (depicted in Figure 1), a set of constructs designed to reflect and evolve with an educational system dynamically. At the core of this paradigm is a knowledge graph that organizes data into a semantic, mathematical graph structure. By applying graph-theoretic operations, this evolving model facilitates information querying, self-updates, and representation of multiple granularity levels—from individual student actions to broader educational policy context. This is a foundational step toward an eventual educational digital twin. Our contributions in this paper are twofold:

- A foundational approach toward an **Educational Digital Twin**, a paradigm that dynamically models big data related to educational pathways, adapting to changes over time.
- The **Educational Digital Twin Knowledge Graph** (EDT-KG), a semantic, mathematical structure developed to support scalable computations and bidirectional data flow as a basis for digital twins in education.

We develop these contributions in the context of a real-world large-scale implementation for the datasets used across the Texas higher education system.

The remainder of the paper is organized as follows. §II discusses prior work relating to digital twins and modeling educational systems. §III presents the Educational Digital Twin and §IV describes its evaluation for use cases in support of decision-making for Texas higher education.

II. Related Work

The utility of digital twins for tracking, intervention, and improved decision-making has been widely recognized [12]–[14] across diverse fields including healthcare [15], [16], energy [17], and aerospace engineering [18]. Yet, digital twins remain underexplored in education, where data are often treated as static. Graph-based models have addressed educational data in various contexts, such as representing learning objectives [19]–[22] and supporting adaptive learning [23]–[27]. Additionally, network analysis has been applied to educational data for insights into curricular structures and learning patterns [28]–[34]. While studies such as [35]–[37] investigate course enrollment patterns, focusing on data within individual institutions, they do not discuss other behavioral patterns, intermediary achievements, and preparedness across institutions. Our approach incorporates these essential elements to model the diverse and high-dimensional nature of student pathways across two- and four-year institutions and time horizons at the Texas-wide scale.

(a) Selected columns from Report CMB001 ("Student")

student	fice	gender	ethnic	ecodis	...
15	1	1	2	1	
16	2	0	7	1	
17	2	0	7	1	

(b) Selected columns from Report CMB00S ("Student Schedule")

student	course	grade	credit	fcl	...
15	Math 1	A	TRUE	1	
15	Eng 1	B	TRUE	0	
16	Math 1	C	TRUE	3	
17	Eng 1	A	FALSE	4	
17	Econ 2	A	FALSE	4	

(c) Selected columns from Report CMB009 ("Graduation"):

student	fice	degree	level	major	type	...
9	1	AA	1	012	Academic	
13	1	ATC	2	156	Technical	
25	2	CCC	5	213	Tech-Prep	

TABLE I: Examples of data with Texas Higher Education Coordinating Board structure. The structure of these files is available in public data manuals. The data values themselves in these examples are notional.

III. The Educational Digital Twin

We begin by describing the structure of the Texas domain data that inform the design of the digital twin. We then describe the theoretical construction of the digital twin's core component, the knowledge graph. Subsequently, we explore the graph-theoretic operations that enable bidirectional flow between the digital and physical realms.

A. Data-centric Digital Twin Formulation

The semantic structure of EDT-KG is defined to reflect the structure of the datasets curated by the Texas Higher Education Coordinating Board (THECB). We present this structure, noting that the structure of THECB datasets is publicly available in published data manuals.[1] To comply with data access agreements and preserve privacy, the values of all data entries (e.g., enrollment numbers, course outcomes) contained within our published examples here are notional, other than cases where the data are available publicly from various educational websites (e.g., institutional names, course names, course sequences, credit requirements, etc.).

The THECB dataset offers a multi-faceted view of student and institutional data, organized into various official reports. For example, the "Student Schedule Report" (Report CBM00S) details student course enrollments. Each report is stored as a binary SAS table-based file, representing data for a specific

[1] See, for example, https://www.texaseducationinfo.org/Home/Us/About%20Our%20Data, accessed Sept. 2024.

```
Associate Degree - Biology
Recommended Course Sequence
Semester I
ENGL 1301 - Composition I
BIOL 1406 Cellular and Molecular Biology
Select one of the following:
    * MATH 1314 - College Algebra
    * MATH 2412 - Pre-calculus MATH
...
Semester II
ENGL-1302 English Composition II
BIOL-1407 Structure and Function of Organisms
...
```

(a) HTML from Austin Community College's website showing recommended course selections for the Biology major.

```
Biology Field of Study (FOS)
1. BIOL 1406
Choose one of the following:
* BIOL 1306
* BIOL 1106
2. BIOL 1407
Choose one of the following:
* BIOL 1307
* BIOL 1107
...
```

(b) Credit requirements for the Biology Field of Study, published as a PDF file by THECB.

Fig. 2: Unstructured educational data from public web sources.

timestep and type of institution. The frequency of these timesteps varies—some reports are generated only during the summer, while others are produced in the spring. Table 2 illustrates the structure of a subset of data from three selected reports: the Student Report (Report `CBM001`, Table Ia), the Student Schedule Report (Report `CBM00S`, Table Ib), and the Graduation Report (Report `CBM009`, Table Ic). For brevity, we display only a limited selection of columns to illustrate the tabular form of the official data; for instance, the Student Report in 2017 contains 61 columns. The structure and data schema of the reports can change from one timestep to another. In developing our digital twin, we primarily utilize four reports: Student, Student Schedule, Texas Success Initiative, and Graduation. The raw data from these reports are compiled into 242 binary files encompassing the records of over six million students over a decade.

Alongside the THECB reports, our Educational Digital Twin incorporates publicly available data from educational websites. These data include institutional recommendations for course sequences by major. For instance, Fig. 2a displays unstructured text from Austin Community College advising on course selections for the Biology major. Additionally, we use official published THECB guidelines that specify credit requirements for various majors. An example of this is shown in Fig. 2b, where the unstructured text outlines the credit requirements for the Biology field of study.

B. Educational Digital Twin Knowledge Graph

Preliminaries: A graph, or equivalently, network, is represented by a tuple $G = (V, E)$ where V is the set of vertices, or nodes, and E is the set of edges $E \subseteq V \times V$. To denote the edge e between vertices v and v', we write $e \coloneqq (v, v')$. In a data model, vertices and edges can be given types, representing the entities and their relationships. We use the notation $v{:}\sigma$ and $e{:}\varsigma$ to express that vertex v is of type σ and edge e is of type ς, respectively. Vertices and edges can also be further augmented with other arbitrary data attributes. To refer to a given attribute `foo` on a vertex (resp. edge), we write `v.foo` (resp. `e.foo`).

We construct EDT-KG as a typed graph model $G = (V, E)$. We develop the structure of the graph by identifying key entities that are integral to the educational system. Initial key entities include Student, Course, Assessment, Award, Major, and Institution, each represented as a vertex with a specific type. For each entity, a set of permissible attributes is defined to carry relevant information. For example, the Student entity includes attributes such as `ethnicity` and `gender`. Edges in EDT-KG represent relationships between entities and, like vertices, are also typed. Each type of edge is named using a verb that reflects the directionality and nature of the relationship. For example, the edge type *achieves* is used to denote that a Student has completed an Assessment, indicating the edge points from Student to Assessment. Similarly, a Student *enrolls in* a Course, defining a directional relationship from the student to the course.

We define queries using logical predicates that describe the properties and relationships between vertices and edges in the graph. A query is formally expressed as a predicate P over the sets of vertices V and E such that

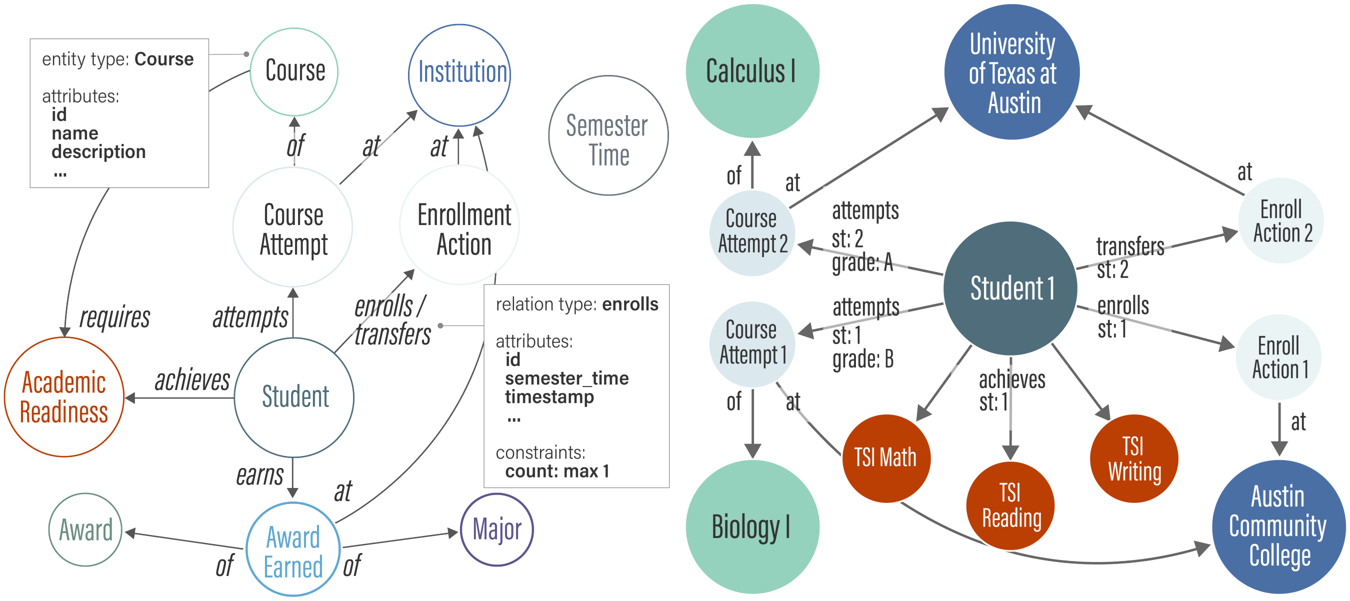
$$P(V, E) \coloneqq \{(v, e) \in V \times E \mid \Phi(v, e)\}, \tag{1}$$

where $\Phi(v, e)$ is a logical formula that specifies the conditions vertices and edges must satisfy. This formula can incorporate various attributes of vertices and edges, such as types, properties, and the existence of paths or subgraphs.

The design of EDT-KG is driven by considerations of expressivity, ease-of-use, and performance, particularly in deciding whether data should be represented as a vertex, an attribute on a vertex, or an attribute on an edge. For example, the edge attribute `ecodis`, which indicates a student's economically disadvantaged status, is associated with the edge types *enrolls* and *transfers*. Alternatively, this attribute could be modeled as a separate vertex linked by a new edge type *has status* connecting Student to a hypothetical ecodis node. The adopted design, as illustrated in Fig. 3a, is the culmination of multiple iterations involving refactoring, testing, and user consultations to ensure it meets the needs of diverse stakeholders. Fig. 3b displays a notional instantiation of the resulting design.

C. Graph Transformations of the Educational Digital Twin Knowledge Graph

We construct graph transformations to aggregate and manipulate data at different levels of granularity for performant, scalable querying. Formally, graph transformations are defined



(a) The EDT-KG ontology

(b) Example instantiation of EDT-KG

Fig. 3: The knowledge graph layer of our Educational Digital Twin.

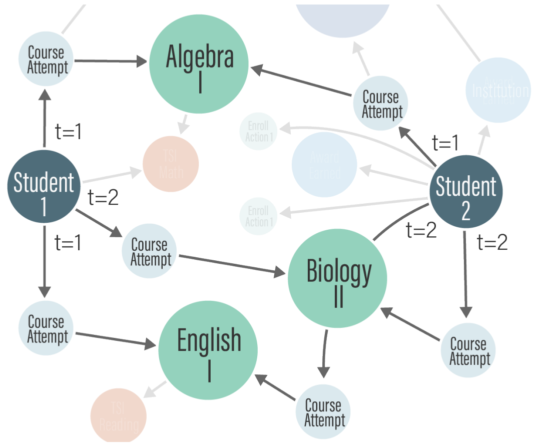


Fig. 4: Original graph G, highlighting the selection Φ.

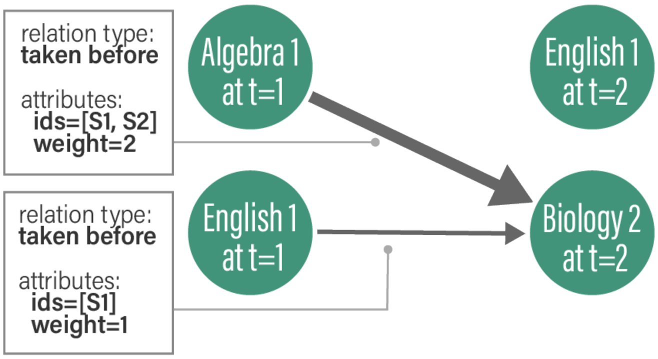


Fig. 5: Transformed graph G' shows four κ vertices. Students S_1 and S_2 have taken Algebra 1 in the first semester followed by Biology 2 in the second semester.

through a series of operations that operate on input graph G, resulting in a new graph G'. The transformation process can be described as a three-step process that involves: 1) **Selection**: a selection $\Phi(v, e)$ of vertices $(v \subseteq V, e \subseteq E)$, 2) **Projection**: a projection of the selection onto a new set of vertices v' and edges e', and 3) **Aggregation**: a traversal of the graph to aggregate quantities to form new attributes of v' and e' in G'.

Example: To capture aggregate patterns in course progression, we construct a graph transformation, $G \rightarrow G'$, which tracks course sequences across semesters. The transformation comprises three steps (illustrated in Fig. 4):

1) **Selection**: Define the selection $\Phi(v, e)$, where v holds for vertices of type Student with outgoing edges of types *attempts* and *of*. This selection isolates a subgraph containing vertices of types Student, CourseAttempt, and Course, along with their connecting edges.
2) **Projection**: For each CourseAttempt and Course, create a vertex κ:CourseAttemptedInTimestep for each unique course attempt per semester. For consecutive course pairs $(\kappa_t, \kappa' t + 1)$ taken across timesteps t and $t+1$, add a *taken before* edge from κ_t to $\kappa t + 1$.
3) **Aggregation** Assign each *taken before* edge an `ids` attribute with student IDs for $(\kappa_t, \kappa'_{t+1})$ and a `weight` attribute for future, frequency-based queries. The transformed graph G' is shown in Fig. 5.

To highlight efficient graph traversal of transformed graph

Algorithm 1 Compute support of course pathways on transformed graph G'

```
set Stack $S := \{\}$, vertices $V \leftarrow \{\kappa \mid (\kappa.\texttt{id} = T\}$, Hash $H := P \to \mathbb{N}$
for $v_i \in V$ do
    set Stack $p_i = \{v_i\}$; $H[p_i] = |r|$; push $S$, $V$
    while $S \neq \emptyset$ do
        $v \leftarrow$ S.pop()
        if $v \in V$ then
            set previous edge ids $r := U$
        end if
        for edge $e \in \{v \leftarrow w$:taken before$\}$ do
            compute intersection $c \leftarrow r \cap$ e.ids
            if $c \neq \emptyset$ then
                $r \leftarrow$ e.ids
                push $S$, $w$; push $p$, $w$
                set $H[p] = \min\{H[p], |c|\}$
            else
                initialize new path $p' \leftarrow$ p.copy()
                set $H[p'] = H[p]$; $p \leftarrow p'$
            end if
        end for
    end while
end for
```

G', we describe a traversal algorithm, detailed in Alg. 1, to identify the most common course pathways of length K for students transferring at timestep K. Starting from the final timestep T vertices $\kappa_{t=K}$, the algorithm computes the intersection I_t of each edge's `ids` attribute with the previous edge's `ids`, initialized to the universal set. When I_t is empty, pruning occurs by backtracking, leveraging the downward closure property: if a sequence of courses (e.g., x_1, x_2) is taken by two students, any extension (e.g., x_1, x_2, x_3) cannot exceed this count. This process continues until paths of length K are recorded, maximizing pruning to compute pathway support.

Remark. *The time complexity of Algorithm 1 is $O(N + R \times S)$, where N is the number of* CourseAttemptedInTimestep *vertices, R is the number of taken before edges, and S is the average size of the* `ids` *array in G'.*

In the worst case with no pruning, the traversal may visit all vertices and edges. For every edge, it executes an intersection operation which can be done in linear time $O(S)$ by comparing two given `id` arrays in sorted order. Thus, the time complexity is $O(N + R \times S)$; however, in practice, S is quite small when starting from the last timestep and pruning consistently eliminates a majority of pathways through the graph.

IV. Evaluation

This section presents two demonstrations of the EDT-KG, highlighting its impact on educational decision-making and its scalability compared to conventional methods.

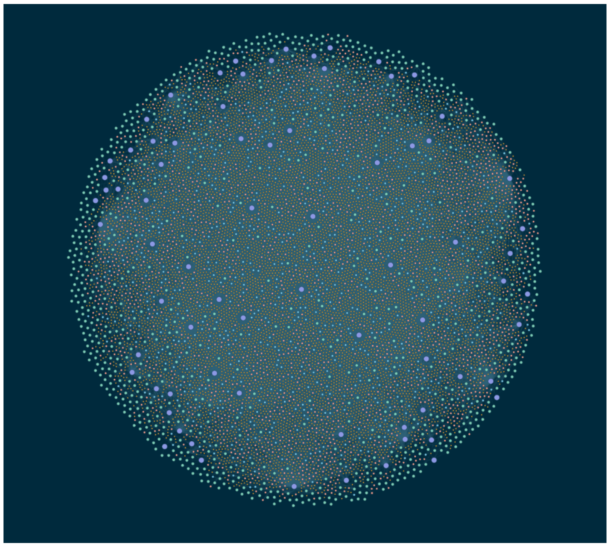


Fig. 6: EDT-KG, visualized with every node and edge.

A. Multi-scale modeling to support decision-making

Although the entirety of EDT-KG, visualized in Fig. 6, provides detailed modeling at a fine level of granularity, extracting insights at a coarser level may be important for informing some decisions but challenging due to the complexity and density of the data. To address these challenges, we develop two specific graph transformations that facilitate analysis at varying scales of granularity.

First Transformation: Academic Readiness and Course Enrollment: This transformation, $G \to G'$, explores the relationships between academic readiness and courses taken across different majors. Academic readiness is defined through three dimensions: "Texas Success Initiative Mathematics" (TSI Math), "Texas Success Initiative Reading" (TSI Reading) and "Texas Success Initiative Writing" (TSI Writing). These dimensions reflect students' semester-wise readiness in mathematics, reading, and writing and are represented in EDT-KG as Readiness vertices. These vertices receive incoming *requires* edges from Course vertices and *achieves* edges from Student vertices, as depicted in Fig. 3a. While some courses have prerequisite requirements of other courses, many entry-level courses are linked to these TSI dimensions. Given the broad array of courses available and the large number of students who do not meet TSI requirements, it is useful to explore how students select courses with identical requirements and whether there are discernible patterns in course selection relative to TSI requirements fulfilled, which may vary by major.

To construct transformation G', we create new nodes typed CourseTakenInMajor for each course within each major. Each node is assigned attributes `num_students` and `average_grade`, aggregated from all attempts of that course within the specified major. We then identify all courses sharing the same TSI requirements. For each set of courses with

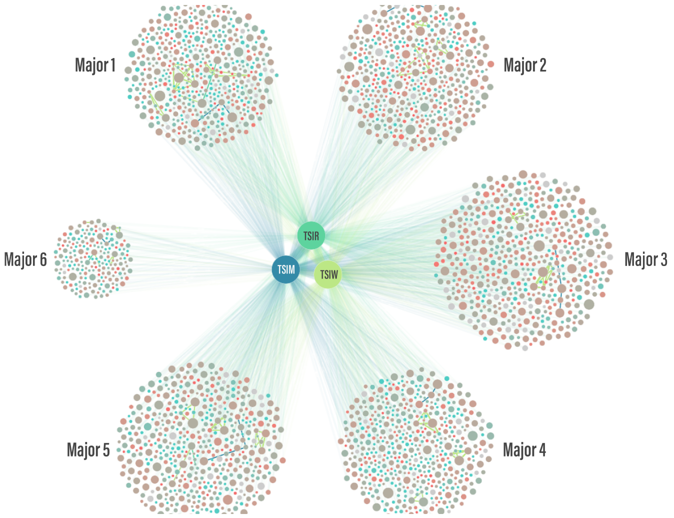


Fig. 7: Graph transformation $G \rightarrow G'$: Exploring relationships between academic readiness, course enrollment, and student outcomes across majors. The three central nodes are the TSI READINESS nodes.

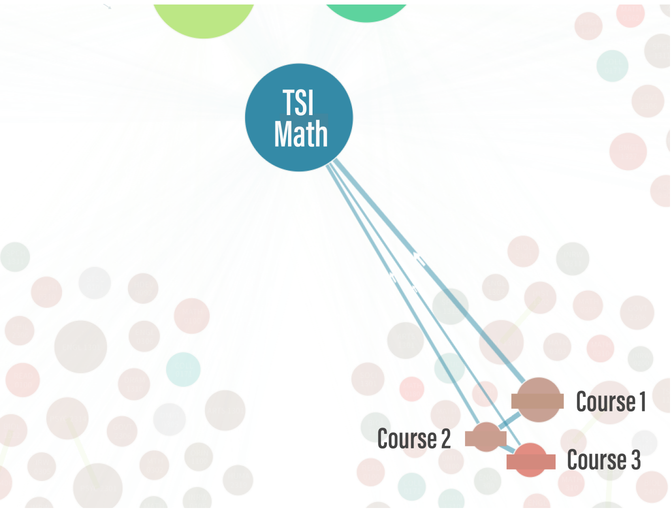


Fig. 8: Zoomed-in portion of $G \rightarrow G'$: `Course 1`, `Course 2` and `Course 3` have the same TSI Math requirement but different ratios of students who do not meet the requirement.

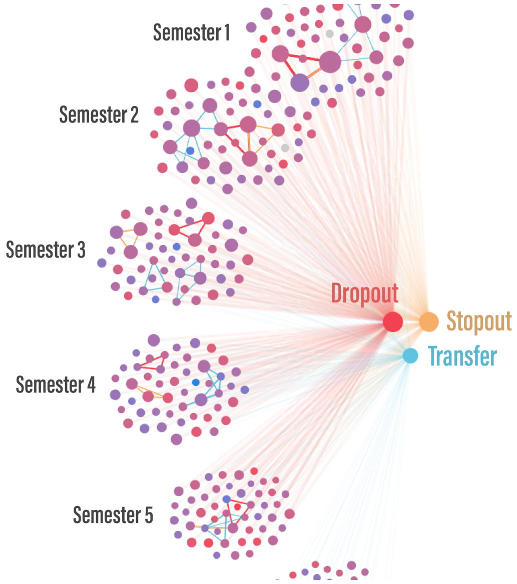


Fig. 9: Graph transformation G' visualizing the relationships between courses and courses, and outcomes across semesters.

identical requirements, we create *has similar requirements* edges between each course pair in the group. For each course and corresponding TSI requirement, we identify students who do not meet the requirement and attempted the course, and create an edge typed *has students who do not meet requirement*. We assign the edge attribute `ratio_students`, quantifying the ratio of students not meeting the requirement for that course.

Fig. 7 depicts the resulting transformation G'. In the visualization, the three central nodes are the TSI READINESS nodes. These READINESS nodes are surrounded by the COURSETAKENINMAJOR nodes, grouped into clusters that correspond to different majors. For clarity, this visualization shows only six of the 417 majors analyzed. In the zoom-in shown in Fig. 8, we observe that `Course 1`, `Course 2`, and `Course 3` all share identical TSI requirements. However, the proportions of students meeting these TSI requirements vary across these courses, as indicated by the differing thicknesses of the edges connecting them. This variation in compliance with TSI requirements across courses reveals important factors that influence student course selection and adherence to academic standards. This analysis helped guide discussions around curriculum design and educational policies regarding readiness requirements.

Second Transformation: Student Pathways and Graduation Outcomes: This second transformation explores how graduation outcomes relate to course selection over time. We create COURSEATTEMPTEDINTIMESTEP vertices for each course attempted in each semester and *taken before* edges between courses attempted to represent sequential course attempts.

We define three additional vertex types representing specific instances of graduation outcomes: `Dropout`, `Stopout`, and `Transfer`. A `Dropout` refers to students who stop enrollment for at least two consecutive semesters, a `Stopout` to those who pause enrollment for one semester but return the following semester, and a `Transfer` to those who enroll in a four-year institution subsequently. We link COURSEATTEMPTEDINTIMESTEP vertices to OUTCOME vertices with *preceding* edges, which represent taking a course and resulting in an outcome at the semester's end. We also introduce *taken together* edges to indicate groups of courses most frequently taken together before a specific graduation outcome.

Fig. 9 visualizes this second transformation. Three OUTCOME vertices lie at the center right, with typed edges among COURSEATTEMPTEDINTIMESTEP vertices. Fig. 10 shows

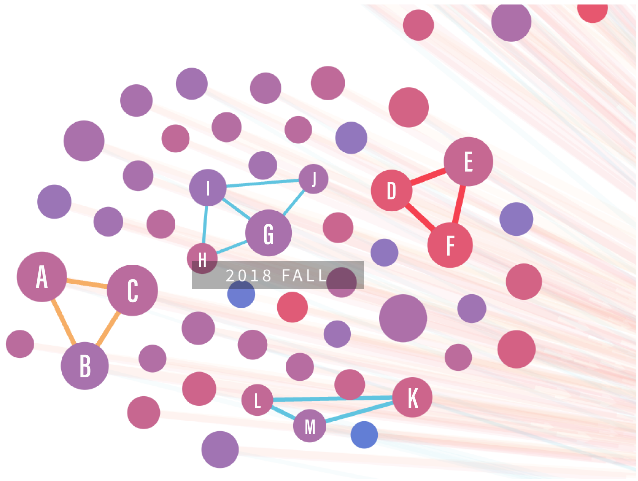
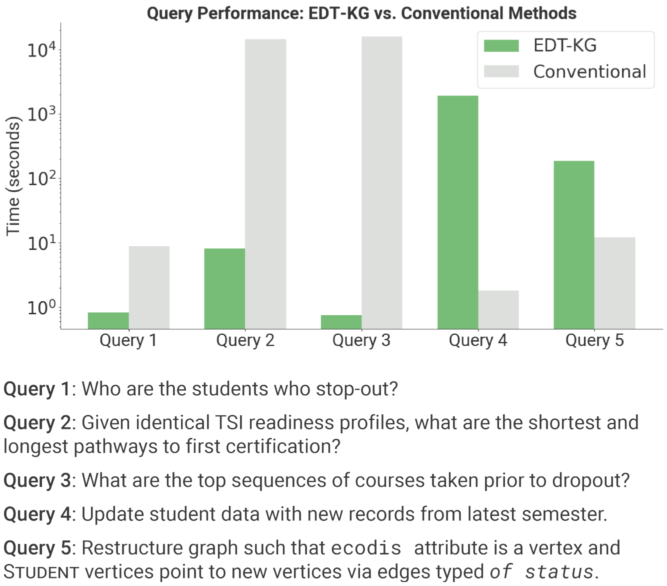


Fig. 10: Zoom-in of G'' with common sets of courses taken before transfer (blue), dropout (red) and stopout (yellow).



Query 1: Who are the students who stop-out?

Query 2: Given identical TSI readiness profiles, what are the shortest and longest pathways to first certification?

Query 3: What are the top sequences of courses taken prior to dropout?

Query 4: Update student data with new records from latest semester.

Query 5: Restructure graph such that `ecodis` attribute is a vertex and STUDENT vertices point to new vertices via edges typed *`of status`*.

Fig. 11: Comparison of query times in EDT-KG versus conventional methods.

a zoom-in of several common pathways. One consists of courses `A-B-C`, leading to `Stopout` and another consists of courses `D-E-F`, leading to `Dropout`. Three other pathways, highlighted in blue, show the three groups of courses most commonly taken together before transfer. In our discussions with decision-makers, these insights revealed previously unknown interactions between course attempts and outcomes.

B. Performance comparison

To assess scalability, we conduct a series of tests involving knowledge retrieval and update. For comparative analysis, we use conventional methods in educational research that involves manipulating raw data with tabular structures, using tools such as R, Stata and Python. Our objective is not a strict laboratory-style comparison with Python and Pandas but to highlight the practical differences one might expect when employing EDT-KG in real-world scenarios as opposed to traditional educational data analytics. Therefore, we follow standard educational analytics protocols without special optimizations.

Fig. 11 shows the query times for a subset of queries executed with EDT-KG, compared to conventional methods. These queries are part of a broader set of over 20 queries, selected here for illustrative purposes. These five queries illustrate the differences between read (Queries 1–3) and write (Queries 4 and 5) operations. Query 1 retrieves demographic profiles of students who stop-out, where graph traversal in EDT-KG is much more efficient. Query 2 retrieves pathways to a student's first certification, often spanning multiple institutions. With conventional methods, Queries 2 and 3 were so memory-intensive that pathway searches were limited and still took nearly three hours. These queries were also challenging to write and debug with conventional methods, underscoring EDT-KG's advantage in semantic clarity. Query 3, finding common course sequences before dropout, highlights EDT-KG's capability to traverse multi-timestep paths efficiently. In Query 4, the longer time required to update EDT-KG compared to reading tabular data reflects the bulk data updates involved; however, single data point updates are faster with EDT-KG. Query 5, which modifies the ontology to accommodate data changes, contrasts with the conventional need for additional dataframes. Although time-consuming, these updates are essential in the digital twin's physical-to-virtual flow, preserving alignment with real-world data.

V. Conclusion

The digital twin paradigm presents an opportunity for educational big data to support improved decision-making and student outcomes. The approach introduced in this paper goes beyond traditional educational data analytics, which tend to focus on isolated snapshots of data, typically confined to a single institution and frozen in time. Instead, our construct integrates diverse, large-scale and evolving data to represent the changing physical system, capturing evolving student pathways within larger educational contexts. Our implementation integrating millions of student records across the Texas postsecondary domain demonstrates how the digital twin paradigm provides valuable insights to educational decision-makers.

Several aspects of the Educational Digital Twin go beyond this paper's scope. First, we are limited in the specific demonstrations that can be published, due to privacy restrictions on the underlying datasets. For example, while our construct has the potential to generate personalized insights for advising or course selection—similar to digital twins in personalized medicine—we are not permitted to show such an analysis. However, the graphical formulations of our methodology make clear how such analyses could be done. Second, privacy and security are major concerns for digital twins. Our research is conducted within a secure data environment and our results shared only with authorized stakeholders. There are clear benefits to making an Educational Digital Twin more widely accessible (for example, to students or faculty advisors) but doing so would require guarantees that the datasets underlying the digital twin could not be reverse-engineered. This remains an open and pressing research area for all digital twin applications. Future directions include expanding the model to include additional data facets and larger slices of the student population, such as extending coverage to K-12 levels.

ACKNOWLEDGMENT

This work was supported in part by Texas Higher Education Coordinating Board (THECB) and Department of Energy grant DE-SC0021239. The opinions and conclusions expressed in this document are those of the authors and do not necessarily represent the opinions or policy of the THECB. We thank Lori Fey for her support throughout the course of the project.

REFERENCES

[1] J. Juszkiewicz, "Trends in community college enrollment and completion data, issue 6," in *American Association of Community Colleges*, 2020.

[2] J. Brand, F. T. Pfeffer, and S. Goldrick-Rab, "The community college effect revisited: The importance of attending to heterogeneity and complex counterfactuals," *Sociol Sci.*, 2014.

[3] J. Labov, "Changing and evolving relationships between two- and four-year colleges and universities: They're not your parents' community colleges anymore," *CBE Life Sci Educ.*, 2012.

[4] A. Varty, "Promoting achievement for community college STEM students through equity-minded practices," *CBE Life Sci. Educ.*, 2022.

[5] G. Crisp and A. Nunez, "Understanding the racial transfer gap: Modeling underrepresented minority and nonminority students' pathways from two-to four-year institutions," *Review of Higher Education*, vol. 37, no. 3, 2014.

[6] P. R. Bahr, E. Jones, and J. Skiles, "Investigating the viability of transfer pathways to STEM degrees: Do community colleges prepare students for success in university STEM courses?" *Community College Review*, vol. 51, no. 4, 2023.

[7] H. Jabbar and W. Edwards, "Choosing transfer institutions: examining the decisions of texas community college students transferring to four-year institutions," *Education Economics*, vol. 28, no. 2, pp. 156–178.

[8] D. Chambliss and C. Takacs, *How College Works*. Harvard University Press, 2018, ch. 1, p. 228.

[9] J. Scott-Clayton, *The Shapeless River: Does a Lack of Structure Inhibit Students' Progress at Community Colleges?* Routledge, 2015, ch. 6.

[10] J. Juszkiewicz, "Trends in community college enrollment and completion data," in *American Association of Community Colleges*, ser. 6, 2019.

[11] National Academies of Science, Engineering and Medicine, *Foundational Research Gaps and Future Directions for Digital Twins*. Washington, DC: The National Academies Press, 2024.

[12] M. Grieves and J. Vickers, "Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems," *Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches*, 2017.

[13] A. Rasheed, O. San, and T. Kvamsdal, "Digital twin: Values, challenges and enablers from a modeling perspective," *IEEE Access*, vol. 8, 2020.

[14] M. W. Grieves, "Digital twins: past, present, and future," in *The Digital Twin*. Springer, 2023, pp. 97–121.

[15] T. Hernandez-Boussard, P. Macklin, E. J. Greenspan, A. L. Gryshuk, E. Stahlberg, T. Syeda-Mahmood, and I. Shmulevich, "Digital twins for predictive oncology will be a paradigm shift for precision cancer care," *Nature Medicine*, vol. 27, no. 12, pp. 2065–2066, 2021.

[16] R. Laubenbacher, B. Mehrad, I. Shmulevich, and N. Trayanova, "Digital twins in medicine," *Nature Computational Science*, vol. 4, no. 3, pp. 184–191, 2024.

[17] F. Stadtmann, A. Rasheed, T. Kvamsdal, K. A. Johannessen, O. San, K. Kölle, J. O. Tande, I. Barstad, A. Benhamou, T. Brathaug *et al.*, "Digital twins in wind energy: Emerging technologies and industry-informed future directions," *IEEE Access*, 2023.

[18] E. Glaessgen and D. Stargel, "The digital twin paradigm for future NASA and US Air Force vehicles," in *Proceedings of the 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, 2012, pp. AIAA Paper 2012–1818.

[19] R. Auras, L. Bix, D. Xu, C. Weir, M. Daum, E. Almenar, A. Brann, R. Iwaszkiewicz, D. P. Kamdem, M. Mahmoudi, J. Marshall, M. Mohiuddin, M. Rabnawaz, A. Joodaky, and E. Lee, "Mapping class learning outcomes of the core curriculum to university learning goals at Michigan State University's School of Packaging," *Packaging Technology and Science*, vol. 36, no. 4, pp. 293–305, 2023.

[20] L. Huang, K. Bicol, and K. Willcox, "Modeling COVID-19 disruptions via network mapping of the Common Core Mathematics Standards," *Computers in Education Journal*, vol. 13, no. 2, 2023.

[21] V. Kivimäki, J. Pesonen, J. Romanoff, H. Remes, and P. Ihantola, "Curricular concept maps as structured learning diaries: Collecting data on self-regulated learning and conceptual thinking for learning analytics applications," *Journal of Learning Analytics*, vol. 6, no. 3, 2019.

[22] J. Seering, L. Huang, and K. Willcox, "Mapping outcomes in an undergraduate aerospace engineering program," in *ASEE Annual Conference & Exposition, 14-17 June, Seattle, Washington.*, 2015.

[23] K. Arnold and M. Pistilli, "Course signals at Purdue: using learning analytics to increase student success," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. New York, NY, USA: Association for Computing Machinery, 2012, p. 267–270.

[24] S. Atalla, M. Daradkeh, A. Gawanmeh, H. Khalil, W. Mansoor, S. Miniaoui, and Y. Himeur, "An intelligent recommendation system for automating academic advising based on curriculum analysis and performance modeling," *Mathematics*, vol. 11, no. 5, 2023.

[25] T. Cavanagh, B. Chen, R. A. M. Lahcen, and J. R. Paradiso, "Constructing a design framework and pedagogical approach for adaptive learning in higher education: A practitioner's perspective," *International Review of Research in Open and Distributed Learning*, vol. 21, no. 1, 2020.

[26] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," *Journal of Medical Systems*, vol. 43, no. 6, p. 162, 2019.

[27] K. R. Koedinger, S. D'Mello, E. A. McLaughlin, Z. A. Pardos, and C. P. Rosé, "Data mining and education," *WIREs Cognitive Science*, vol. 6, no. 4, pp. 333–353, 2015.

[28] R. Duarte, A. Lacerda Nobre, F. Pimentel, and M. Jacquinet, "Broader terms curriculum mapping: Using natural language processing and visual-supported communication to create representative program planning experiences," *Applied System Innovation*, vol. 7, no. 1, 2024.

[29] K. E. Willcox and L. Huang, "Network models for mapping educational data," *Design Science*, vol. 3, p. e18, 2017.

[30] R. Ghannam and I. S. Ansari, "Interactive tree map for visualising transnational engineering curricula," in *2020 Transnational Engineering Education using Technology (TREET)*, 2020, pp. 1–4.

[31] J. Guerra, Y. Huang, R. Hosseini, and P. Brusilovsky, "Graph analysis of student model networks," in *CEUR Workshop Proceedings*, vol. 1446, 06 2015.

[32] V. Sheshadri, C. Lynch, and T. Barnes, "InVis: An EDM tool for graphical rendering and analysis of student interaction data," *CEUR Workshop Proceedings*, vol. 1183, pp. 65–69, 01 2014.

[33] C. Simon, D. Gomez, and R. Herrero, "Network analysis: An indispensable tool for curricula design. a real case-study of the degree on mathematics at the URJC in Spain," *PLOS ONE*, vol. 16, 03 2021.

[34] C. Vieira, P. Parsons, and V. Byrd, "Visual learning analytics of educational data: A systematic literature review and research agenda," *Computers and Education*, vol. 122, pp. 119–135, 2018.

[35] M. Raji, J. Duggan, B. DeCotes, J. Huang, and B. V. Zanden, "Modeling and visualizing student flow," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 510–523, 2021.

[36] Y. Chen, A. Fu, J. J.-L. Lee, I. W. Tomasik, and R. F. Kizilcec, "Pathways: Exploring academic interests with historical course enrollment records," in *Proceedings of the Ninth ACM Conference on Learning @ Scale*, ser. L@S '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 222–233.

[37] G. Angus, R. D. Martinez, M. L. Stevens, and A. Paepcke, "Via: Illuminating academic pathways at scale," in *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*, ser. L@S '19. New York, NY, USA: Association for Computing Machinery, 2019.