Modeling Longitudinal Student Pathways with Explainable Generative Models

Luwen Huang
luwen@cs.utexas.edu
University of Texas at Austin
Austin, Texas, USA

Inderjit S. Dhillon
inderjit@cs.utexas.edu
University of Texas at Austin
Austin, Texas, USA

Karen E. Willcox
kwillcox@oden.utexas.edu
University of Texas at Austin
Austin, Texas, USA

Abstract

Student pathways — trajectories spanning academic readiness, course selections, grades, and enrollment outcomes — are critical for understanding educational progress, particularly in community college systems where pathways are highly diverse. Restricted access to student records and a sparsity of coherent pathway data pose barriers to modeling and broader research engagement. This paper contributes a scalable and explainable framework for generating synthetic student pathway data, along with a Bayesian network structure learning approach adapted to the sparsity and temporal complexity of educational trajectories. We present a generative modeling approach that produces realistic *synthetic student pathway data* by learning a Bayesian network trained on longitudinal student records across multiple tables linked across time. Our model captures complex conditional dependencies across hundreds of variables while remaining interpretable: each parameter encodes transparent relationships that can be inspected or adjusted. We show that our method outperforms an independent sampler in reproducing marginal, conditional, and higher-order patterns in the real data. Our analysis shows that existing k-anonymity rules are infeasible for real or synthetic data, motivating a shift toward model-aware approaches for privacy considerations in student pathway data.

CCS Concepts

• **Computing methodologies** → **Bayesian network models**; • **Applied computing** → **Education**.

Keywords

Educational student pathways, synthetic data, generative modeling, Bayesian network models

ACM Reference Format:
Luwen Huang, Inderjit S. Dhillon, and Karen E. Willcox. 2026. Modeling Longitudinal Student Pathways with Explainable Generative Models . In *LAK26: 16th International Learning Analytics and Knowledge Conference (LAK 2026), April 27-May 01, 2026, Bergen, Norway.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3785022.3785085

1 Introduction

Modeling how students progress through their academic careers is critical for understanding educational outcomes and for designing effective interventions such as advising and policy changes. Yet student pathways remain exceptionally challenging to model [4, 22], particularly in community colleges where trajectories are far more diverse than in four-year institutions. Many students follow unique sequences of entry, course selection, grades, and enrollment decisions over time [7], while prerequisites and institutional rules impose deterministic constraints or soft influences. Figure 1 illustrates one such notional pathway: Jennifer, a first-time community college student, fails College Algebra, pauses enrollment for a semester, and later returns to earn additional credits recommended at her community college before transferring to a four-year institution. Together, these factors define a domain that is high-dimensional, sparse, and temporally coupled.

In many educational settings, even institutional researchers may face restricted access to student-level data due to data-use agreements. Synthetic data therefore enable broader research, tool development, and reproducibility without exposing protected records. Generating synthetic student data has emerged as a promising solution in learning analytics [14, 15]. Although generative modeling is an active research area in the broader machine learning literature, a gap remains for educational settings: student records are multi-tabular, highly sparse due to unique individual behaviors, coupled across multiple dimensions of activity, and longitudinally dependent through time (see §2 for a detailed gap analysis). In addition, synthetic data for education must satisfy explainability. When generated data reveal unexpected or sensitive patterns (e.g., correlations between demographic attributes and performance), stakeholders must be able to trace and adjust the underlying parameters. Finally, educational privacy regulations impose strict requirements for privacy accounting, further compounding the modeling challenge and widening the gap between existing methods and educational needs.

In the Texas dataset used in this study, the current practice for producing synthetic data is a format-based approach applied independently to each table. Data stewards manually generate random values that match the type of each column (e.g., strings, integers), without accounting for probability distributions or consistency across linked tables. As a result, the synthetic records cannot be used to model student pathways, which require linking multiple tables and preserving temporal dependencies. In this paper, to evaluate our generated synthetic data, we adopt as a baseline an independent sampler that treats each column as independent but matches its empirical probability distribution. Despite its simplicity, it represents a meaningful step beyond the current practice and provides a clear point of reference that has not previously existed.

The contribution of this work is an explainable method for generating longitudinal synthetic student pathways across multiple linked tables. We learn a dynamic Bayesian network (DBN) that preserves conditional dependencies across student record tables and temporal dependencies across timesteps. Each parameter in

LAK 2026, Bergen, Norway

ACM ISBN 979-8-4007-2066-6/2026/04
https://doi.org/10.1145/3785022.3785085

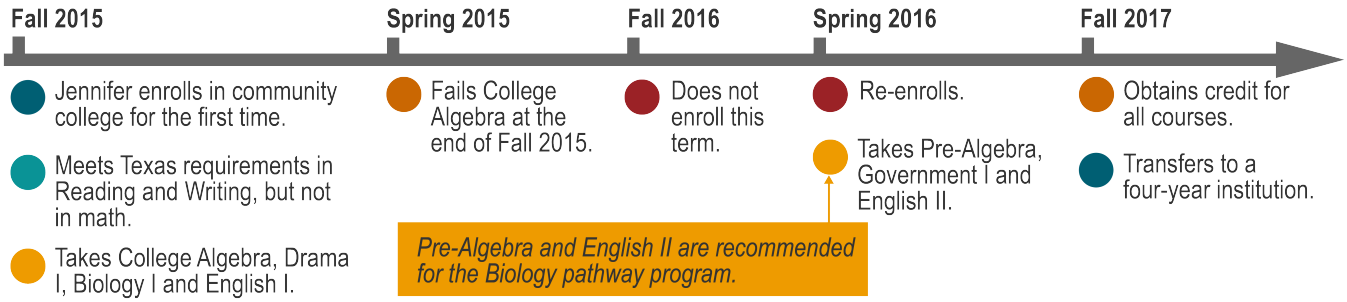


Figure 1: Illustration of a *student pathway*: a longitudinal sequence that combines demographics and academic readiness attributes with enrollment decisions and course outcomes. This notional example follows "Jennifer", who enrolls in community college, fails College Algebra, pauses enrollment, re-enrolls in subsequent courses, and ultimately transfers to a four-year institution. Pathways such as this highlight the diversity of transitions and temporal dependencies that make modeling student trajectories both complex and policy-relevant.

the DBN encodes interpretable relationships that can be inspected and adjusted to support policy-relevant simulation, such as evaluating the effects of changing prerequisites. We evaluate the generated synthetic data, showing that our method outperforms a baseline method that samples each variable independently. Finally, we analyze k-anonymity violations in both real and synthetic data, showing that current k-anonymity break down in student pathway contexts.

The remainder of the paper is organized as follows. §2 outlines the unique challenges of our problem and gaps in existing literature. §3 introduces the data tables used in this study. §4 details our approach, including variable formulation, Bayesian network design, structure learning, and synthetic data generation. §5 evaluates the generated data against real records. §6 concludes the paper.

2 Related Work and Problem Gaps

Generative modeling has been applied in educational contexts for a variety of purposes, including generating assessment items [13, 16] and simulating student responses [25, 28]. In the area of student pathways, prior studies have explored predictive tasks such as modeling enrollment decisions from demographics with Hidden Markov Models [3], and course enrollment dynamics at a single four-year university using latent-variable models [5]. Our work extends this line of research by modeling multiple aspects of student trajectories — demographics, readiness, course selections, grades, and enrollment decisions — at a statewide scale, with an emphasis on explainable synthesis from longitudinal, multi-table student records.

In the broader machine learning literature, generative modeling for synthetic data is an active area of research. One well-known technique is CTGAN [29], which uses a generative adversarial network to synthesize tabular data. Another is TabDDPM [11], which adapts denoising diffusion probabilistic models with a multi-layer perceptron for tabular domains. Most recently, TabDiff [24] introduced column-wise adaptive diffusion processes and demonstrated improved performance over both CTGAN and TabDDPM. However, all of these methods address a *single* table of i.i.d. rows. In contrast, our problem requires modeling longitudinal student pathways across multiple semesters, where records are explicitly *not* i.i.d. For example, a student who has obtained credit for a course is highly unlikely to repeat it in a subsequent semester. In theory, one could flatten these longitudinal records into a massive *wide-format* table with columns such as `Enrollment_t1`, `Algebra_t1`, `Algebra_t2`, `Enrollment_t2`, and so on. Yet this flattening comes with serious drawbacks: (1) It destroys temporal semantics. A flat model such as TabDiff only sees joint feature correlations, not state transitions. For instance, it cannot capture that taking `Freshman Orientation` in semester 1 makes it unavailable in semester 2; (2) The combinatorics yield infeasible sparsity. As a toy calculation, with $N = 1000$ students, $C = 500$ courses, $G = 5$ possible grades (including "not taken"), and $E = 3$ enrollment outcomes, one-hot encoding produces $T \times C \times G$ columns for T semesters. With $T = 4$, this is 10,004 columns. Assuming students take on average $k = 4$ courses per term, the table is approximately 99.8% sparse. This is orders of magnitude sparser than the public benchmarks used by recent generative methods, e.g., `Adult` (15 columns) [1], `Diabetes` (36 columns) [8], or `News` (48 columns) [18]; (3) It mismatches educational policy use-cases. Stakeholders often ask counterfactual questions such as: what if a required course is added or removed, or advising rules change? These are naturally expressed in a structured temporal model, but difficult to implement in a flattened black-box generator. Our approach addresses these challenges by building on Bayesian structure learning [6, 23, 27] and dynamic Bayesian networks (DBNs) [10, 19], which extend static Bayesian networks to temporal domains. While structure learning and DBNs are well established in the machine learning literature [9, 20, 21], their use for generative modeling of student pathways remains scarce.

Privacy in educational data is often enforced through the "small cell rule," formalized as k-anonymity, which requires each record to be indistinguishable from at least $k-1$ others with respect to quasi-identifiers (e.g., gender, race, grades). For instance, the Texas Dallas Education Research Center mandates k=5; thus, reporting on academically prepared students in Algebra is only allowed if five or more such students exist. An alternative approach is differentially private (DP) generative modeling, where privacy is guaranteed by perturbing either the training procedure or the synthetic release. Classical methods such as PrivBayes [30] inject noise during Bayesian network training, but more recent work has explored

approaches such as G-PATE for high-dimensional image data [17] and differentially private normalizing flows for tabular data [12]. Among these, the latter is the most relevant to our setting, yet its empirical evaluation is limited to the same public benchmarks commonly used for non-private generators (`Adult`, `Covertype` [2], etc.), single-table datasets with i.i.d records, no temporal dependencies, and no extreme sparsity. Moreover, such models also remain black-box in nature, offering little explainability for stakeholders. Thus, even in the non-private case, there is a gap in the literature on generative modeling for longitudinal student pathways; when privacy is required, this gap widens. Developing a fully private method is beyond the scope of this work. Instead, our contribution is to provide a strong non-private baseline that can serve as a foundation for future privacy-preserving methods tailored to sparse, temporal educational data.

3 Data: Multi-year Postsecondary Records across Texas Public Institutions

Our analysis is based on a multi-year dataset provided by the Texas Higher Education Coordinating Board (THECB), the governing agency overseeing all public post-secondary education in Texas. The dataset spans from Fall 2012 through Fall 2020 and captures key aspects of postsecondary student pathways across Texas public institutions. Each academic term is represented by a set of structured reports, with one table per report per time period — typically one per semester, though some span an entire academic year. In this paper, we focus on four particular aspects of student behavior: (1) Enrollment Report (`CBM001`), with demographic attributes for each student enrolled; (2) Readiness Report (`CBM002`), with academic readiness indicators for each student, such as Texas State Initiative (TSI) readiness[1] indicators; (3) Course Choice Report (`CBM00S`), with courses attempted for each student; and (4) Course Grade Report (`CBM00S`), with grades received for each student. Table 1 provides sample entries from each report for illustrative purposes. Each report is keyed by a de-identified student identifier (the `ID` columns in Table 1) and linked across time periods. Although student records are de-identified, the THECB retains control over data release, including any synthetic datasets derived from our methods. As a result, Table 1 shows only notional examples. Our goal, developed in collaboration with THECB stakeholders, is to model and generate full student pathways —from initial enrollment to exit — to support decisions around course requirements, policies, and interventions. Variables were selected in consultation with THECB domain experts to reflect real-world priorities rather than arbitrary design choices.

4 Methodology

We begin by defining the Bayesian network model, including variables, structure and parameter learning process. We then describe the structure learning strategy with feature engineering and synthetic data generation and validation.

[1]The Texas State Initiative Assessment determines whether students are ready for entry-level college coursework in the areas of mathematics (TSIM), reading (TSIR), and writing (TSIW).

4.1 Bayesian Network Model and Parameter Estimation

We first categorize the variables in our model into five categories: (1) **Demographics (D)**: Fixed characteristics, such as gender and ethnicity; (2) **Readiness (R^t)**: Academic preparedness indicators in math, reading, and writing at timestep t, assessed upon entry at the initial timestep. At subsequent timesteps, indicators are updated based on course outcomes (e.g., passing College Algebra satisfies math readiness); (3) **Courses Attempted (A^t)**: Binary vector indicating which of C course options were taken at timestep t; (4) **Grades Obtained (G^t)**: Ordinal-valued vector indicating the grade obtained for each course at timestep t, with $G_i^t = 0$ for course i not taken; and (5) **Enrollment Status (E^t)**: Ordinal variable representing whether the student continues to enroll, stops enrollment, or transfers to a four-year institution. These variable groups are not arbitrary; they reflect distinct decision points along the academic timeline and follow a natural temporal progression. For example, a student's demographic attributes are fixed prior to college entry. TSI readiness is assessed before the start of a semester. Next, students select courses, receive grades, and then make enrollment decisions for the following term. Although we encode this temporal ordering as a modeling constraint, it does not imply a causal claim between all upstream and downstream variables — for instance, between demographic traits and later academic outcomes. Rather, we use the ordering to guide the structure learning process (§4.2), helping to narrow the search space and ensure consistency with real-world academic timelines. Table 2 provides a complete description of the variables within each group.

Together, the variables in Table 2 specify the state space of student pathways, over which a joint probability distribution can be defined. Let $\boldsymbol{X} = [\boldsymbol{D}, \boldsymbol{R}^{0:T}, \boldsymbol{A}^{0:T}, \boldsymbol{G}^{0:T}, \boldsymbol{E}^{0:T}]$ denote the full collection of random variables in the system, where superscripts $0{:}\,T$ indicate sequences indexed over discrete time steps $t = 0, \ldots, T$. Here, T denotes the time horizon of interest in semesters (e.g., $T = 6$ corresponds to modeling 3 academic years). We model the joint distribution $P(\boldsymbol{X})$ over the student trajectory variables using a Bayesian network, a directed acyclic graph (DAG) $\mathcal{G}$ in which each random variable $X_i \in \boldsymbol{X}$ is a node, and directed edges $X_i \to X_j$ encode conditional dependency of X_j on X_i. In a Bayesian network representation, the DAG $\mathcal{G}$ states that each variable is conditionally independent of its non-descendants, given its parents. This conditional independence assumption allows the joint distribution to factor into local conditional distributions:

$$P(\boldsymbol{X}) = \prod_i P(X_i \mid \mathrm{Pa}(X_i)) \tag{1}$$

where $\mathrm{Pa}(X_i)$ denotes the parents of variable X_i in $\mathcal{G}$. The product factorization in Eqn. (1) reduces the complexity of modeling the full joint distribution. Rather than estimating a single, intractable distribution over all variables and timesteps, the Bayesian network decomposes the problem into a collection of lower-dimensional conditional distributions. This mathematical property is useful for modeling student pathways, where the variable space is high-dimensional but the dependency structure is often sparse and locally clustered. Figure 2 presents a notional example of our modeling approach, with nodes grouped by category and shown with illustrative edges. For instance, R_1^0 (TSIM) depends on D_{16} (4-Year Intent) and D_{14} (LEP); A_2^0 (College Algebra course selection) depends on R_1^0; and

id	intent	institution	first time	gender
1	2	10010	1	0
2	4	20141	1	0
3	4	33443	0	1

(a) Enrollment Report (CBM001): Fall 2015

id	intent	institution	first time	gender
1	4	10010	0	0
2	2	20141	0	0
3	4	33443	0	1

(b) Enrollment Report (CBM001): Spring 2015

id	TSI math	TSI read	TSI writ	SAT	ACT
1	1	1	2	1100	21
2	2	0	7	900	18
3	2	0	7	650	29

(c) Readiness Report (CBM002): Fall 2015

id	TSI math	TSI read	TSI writ	SAT	ACT
1	1	1	2	1100	21
2	2	1	2	900	18
3	2	2	2	650	29

(d) Readiness Report (CBM002): Spring 2015

id	course	grade	credit	developmental
1	math 1214	A	3	0
2	engl 0300	B	1	1
3	engl 0300	C	1	1
3	math 2314	C	1.5	0

(e) Course Schedule Report (CBM00S): Fall 2015

id	course	grade	credit	developmental
1	math 1314	B	3	0
2	engl 1300	A	3	0
3	engl 2301	F	3	0
3	math 1304	D	2	0

(f) Course Schedule Report (CBM00S): Spring 2015

Table 1: Selected rows and columns (notional examples) from Texas Higher Education Coordinating Board (THECB) data tables for two timesteps. Each subtable corresponds to a reporting file in a timestep: (a) demographics (CBM001), (b) TSI readiness indicators (CBM002), and (c) course-level outcomes (CBM00S). A full list of attribute definitions is available in the THECB Data Manual [26].

downstream decisions like A_5^2 (enrolling in Calculus I at $t = 2$) depend on earlier course choices. This schematic is for illustration only; actual structures are determined through a combination of expert input and structure learning, as described in §4.2.

Since the true joint distribution $P(\boldsymbol{X})$ is unknown, we treat the observed dataset $\boldsymbol{X}^* = [\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}]$ as being generated from a parametric model with unknown parameters $\boldsymbol{\theta}$. Here, each $\boldsymbol{X}^{(n)}$ denotes the full set of variables for student n, and N is the total number of students in the dataset. Our goal is to infer the parameters by computing the posterior distribution $P(\boldsymbol{\theta} \mid \boldsymbol{X}^*)$ — the probability of the parameters given the observed data. By Bayes' rule:

$$P(\boldsymbol{\theta}|\boldsymbol{X}^*) \propto P(\boldsymbol{\theta})P(\boldsymbol{X}^*|\boldsymbol{\theta})$$

where $P(\boldsymbol{\theta})$ encodes our prior belief over the parameters and $P(\boldsymbol{X}^* \mid \boldsymbol{\theta})$ is the likelihood of the data given those parameters. Using the joint factorization property of Bayesian networks (Eqn. 1), the likelihood of the observed data $\boldsymbol{X}^*$ under graph $\mathcal{G}$ and parameters $\boldsymbol{\theta}$ is given by:

$$P(\boldsymbol{X}^* \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{t=0}^{T} \prod_{i} P(X_i^{t,(n)} \mid \mathrm{Pa}(X_i^{t,(n)}); \boldsymbol{\theta})$$

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} \left[\log P(\boldsymbol{X}^* \mid \boldsymbol{\theta}) + \log P(\boldsymbol{\theta})\right]$$

Here, $X_i^{t,(n)}$ denotes the value of the i-th variable for student n at timestep t, and the product over i ranges over all random variables at time t. Each local conditional probability $P(X_i \mid \mathrm{Pa}(X_i); \boldsymbol{\theta})$ is defined by a conditional probability table (CPT), and the MAP estimate $\boldsymbol{\theta}_{\mathrm{MAP}}$ defines the final parameterization of all CPTs. To ensure tractability, we assume conjugate priors for $P(\boldsymbol{\theta})$, which allows each term in $\boldsymbol{\theta}_{\mathrm{MAP}}$ to be computed in closed form.

4.2 Structure Learning and Feature Engineering

To construct the Bayesian network structure $\mathcal{G}$, we adopt a hybrid approach that combines structure learning with manual design. Rather than learning a single global graph all at once, we construct the subgraph for each variable group (as listed in Table 2) independently. For each variable group, we learn the structure of its subgraph and, where appropriate, augment it with domain knowledge when doing so improves training performance. To learn the structure of a variable group, we use a score-based optimization method. We seek the subgraph G^* that maximizes a model selection score. Here, we use the Bayesian Information Criterion (BIC), which trades off model fit with complexity:

$$G^* = \arg\max_{G \in \boldsymbol{G}} \mathrm{BIC}(G), \quad \text{where}$$

$$\mathrm{BIC}(G) = \log P(\boldsymbol{X}^* \mid \boldsymbol{\theta}_{\mathrm{MLE}}, G) - \frac{d}{2} \log N$$

Here, $\boldsymbol{G}$ is the space of valid DAGs over the variable group, $\boldsymbol{\theta}_{\mathrm{MLE}}$ are the maximum likelihood parameters for G, N is the number of students, and d is the number of free parameters in the model. For each variable X_i with r_i values and q_i distinct parent configurations under G, the number of free parameters is $q_i \cdot (r_i - 1)$, so that $d = \sum_{i=1}^{|X|} q_i(r_i - 1)$. We search for the graph G^* that maximizes the BIC score using greedy hill climbing with local operations — edge additions, deletions, and reversals. At each step, a local modification is proposed and accepted if it increases the score; the search continues until a local maximum is reached. To avoid overfitting, we enforce a maximum in-degree constraint on each node. To support structure learning, we derive a set of features informed by domain expertise and institutional logic. These features encode temporal

Category	Variable	Description
Demographics	$D_1 \in \mathbb{Z}_{>0}$	Age
	$D_2 \in \{0, 1\}$	Gender
	$D_3 \in \{0, 1\}$	Hispanic
	$D_4 \in \{0, 1\}$	White
	$D_5 \in \{0, 1\}$	Black
	$D_6 \in \{0, 1\}$	Asian
	$D_7 \in \{0, 1\}$	Native American
	$D_8 \in \{0, 1\}$	International
	$D_9 \in \{0, 1\}$	Pacific-Islander
	$D_{10} \in \{0, 1\}$	Economically-disadvantaged
	$D_{11} \in \{0, 1\}$	Disability
	$D_{12} \in \{0, 1\}$	Homemaker
	$D_{13} \in \{0, 1\}$	Single-parent
	$D_{14} \in \{0, 1\}$	Limited English proficiency (LEP)
	$D_{15} \in \{0, 1\}$	Intent to earn a 2-year degree
	$D_{16} \in \{0, 1\}$	Intent to earn a 4-year degree
Readiness	$R_1^t \in \{0, 1\}$	TSI satisfied in math (TSIM)
	$R_2^t \in \{0, 1\}$	TSI satisfied in reading (TSIR)
	$R_3^t \in \{0, 1\}$	TSI satisfied in writing (TSIW)
Course(s) Attempted	$\boldsymbol{A}^t = [A_1^t, \ldots, A_C^t]$, where $A_i^t \in \{0, 1\}$	Course $i = \{1, \ldots, C\}$ taken at time $t \in \{0, \ldots, T\}$
Grade(s) Obtained	$\boldsymbol{G}^t = [G_1^t, \ldots, G_C^t]$, where $G_i^t \in \{0, 1, 2, 3, 4, 5\}$	Grade for course i at time t; where 0 = Not taken, 1 = F, 2 = D, 3 = Credit or C, 4 = B, 5 = A
Enrollment	$E^t \in \{0, 1, 2, 3\}$	Enrollment action at the end of time t; where 0 = Stop, 1 = Change major, 2 = Continue, 3 = Transfer

Table 2: Overview of the random variables used in the model. Variables are grouped by semantic category, each with a symbolic name and a discrete domain.

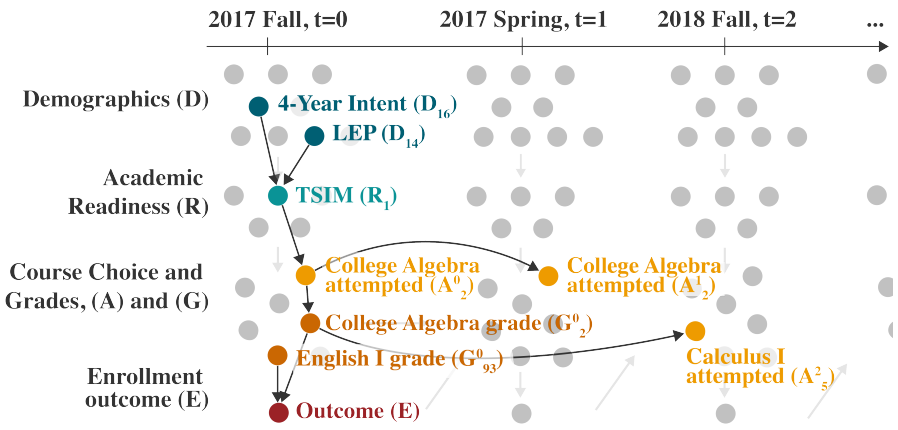


Figure 2: Illustrative Bayesian network structure. Nodes represent student pathway variables grouped by category. Edges encode conditional dependencies among variables.

and intra-variable group dependencies in a more informative and compact form than the original raw data, allowing us to reduce the dimensionality of the problem space. For example, *Course i Credit Obtained by t* ($\mathcal{F}_{1,i}^t$) is a binary feature indicating whether a student has earned credit for course i by timestep t. This eliminates the need to reference the entire sequence of past grades $G_i^{0:t-1}$, reducing the effective Markov order and simplifying dependency modeling. Another derived feature, *Max Math Difficulty Passed by t*, encodes

Feature Name	Description	Type
Course i Credit Obtained by t ($\mathcal{F}^t_{1,i}$)	Indicates whether student has received credit for course i by timestep t	Binary
Max Math Difficulty Passed by t ($\mathcal{F}^t_2$)	Highest difficulty of any math course passed by timestep t, based on prerequisite topology	Ordinal
Prerequisite Passed for Course i by t ($\mathcal{F}^t_{3,i}$)	True if all prerequisites for course i were passed by timestep t, e.g. if U.S. History I was passed for U.S. History II	Binary
# Courses Failed in semester t ($\mathcal{F}^t_4$)	Count of courses student failed during semester t	Integer
# Courses Passed in semester t ($\mathcal{F}^t_5$)	Count of courses student passed during semester t	Integer
Any Fail in semester t ($\mathcal{F}^t_6$)	True if student failed at least one course in semester t	Binary
Any Pass in semester t ($\mathcal{F}^t_7$)	True if student passed at least one course in semester t	Binary
Cumulative Credit Hours by t ($\mathcal{F}^t_8$)	Credit hours earned by timestep t	Integer

Table 3: Derived features $\mathcal{F}$ used as inputs to structure learning. These variables summarize student progress up to timestep t (e.g., cumulative credits, prerequisites satisfied, prior failures).

the most advanced math course a student has successfully completed by timestep t. To compute this, we define a directed acyclic graph $\mathbb{G}$ over the course catalog, where nodes represent courses and a directed edge indicates a prerequisite relationship. We assign each course a difficulty score based on its rank in a topological sort of $\mathbb{G}$. For example, `Calculus I` receives a difficulty score of 4 in the subgraph shown in Figure 3a. This allows the model to account for implausible backward progressions (e.g., taking College Algebra after passing Calculus II). Additional features track term-level academic behavior: total number of courses failed or passed in a given semester, binary indicators for whether any course was failed or passed, and cumulative credit hours earned to date. Table 3 summarizes the derived features used in model training.

Wherever appropriate, we incorporate prior domain knowledge by fixing a set of edges $\mathcal{E}_{\text{fixed}}$, e.g. $\mathcal{E}_{\text{fixed}} = \{R^0_1 \to A^0_2\}$, which remain unaltered during the search. These fixed edges reflect known relationships, such as the influence of academic readiness on course enrollment (e.g., TSIM → College Algebra), which arises from prerequisite requirements that are not always enforced in practice, but still strongly influence behavior. Given $\mathcal{E}_{\text{fixed}}$, the remainder of the graph is learned, constrained to ensure $\mathcal{E}_{\text{fixed}} \subseteq G^*$. Figure 3b illustrates an example of this process for variable group $\boldsymbol{A}^2$. In this example, we fix two edges, $\mathcal{F}^2_2 \to A^2_2$ and $R_1 \to A^2_2$, to encode the hypothesis that enrollment in College Algebra at $t = 2$ (A^2_2) depends on the maximum level of difficulty passed by the beginning of $t = 2$ ($\mathcal{F}^2_2$) and the student's math readiness at $t = 2$ (R^2_1).

We evaluate candidate graphs using held-out data to assess generalization. The dataset is partitioned into a training set $\mathcal{D}_{\text{train}}$, used to estimate parameters $\boldsymbol{\theta}_{\text{MAP}}$, and an evaluation set $\mathcal{D}_{\text{eval}}$, used to assess model fit. For each candidate graph structure, we select a set of target variables $\boldsymbol{Y} \subset \boldsymbol{X}$ and a corresponding set of observed variables $\boldsymbol{Z} \subset \boldsymbol{X} \setminus \boldsymbol{Y}$. Although our goal is to evaluate how well the model captures the conditional distribution $P(\boldsymbol{Y} \mid \boldsymbol{Z})$, we compute the joint divergence between $(\boldsymbol{Y}, \boldsymbol{Z})$ in the model and the evaluation set, which captures both conditional and marginal discrepancies:

$$\begin{aligned} &\mathrm{KL}(P_{\text{emp}}(\boldsymbol{Y}, \boldsymbol{Z}) \,||\, \widehat{P}(\boldsymbol{Y}, \boldsymbol{Z})) = \sum_{\boldsymbol{y}, \boldsymbol{z}} P_{\text{emp}}(\boldsymbol{y}, \boldsymbol{z}) \log \frac{P_{\text{emp}}(\boldsymbol{y}, \boldsymbol{z})}{\widehat{P}(\boldsymbol{y}, \boldsymbol{z})} \\ &\text{where } \widehat{P}(\boldsymbol{Y}, \boldsymbol{Z}) = \prod_{X_i \in \boldsymbol{Y} \cup \boldsymbol{Z}} P(X_i \mid \mathrm{Pa}_G(X_i)) \end{aligned} \tag{2}$$

In Eqn. (2), the first line defines the KL divergence between the empirical and model-implied joint distributions. The second line gives the model factorization used to compute $\widehat{P}(\boldsymbol{Y}, \boldsymbol{Z})$ using the CPTs in $\boldsymbol{\theta}_{\text{MAP}}$. The resulting learned Bayesian network comprises over 1,000 nodes due to repeated variables across time steps, making full visualization impractical. We present representative subgraphs in §5 to illustrate important learned structures.

4.3 Generating and Validating Synthetic Samples

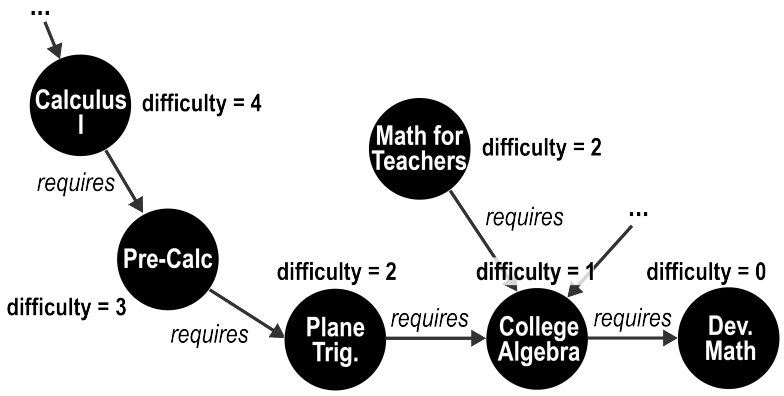
To generate a synthetic dataset, we sample each synthetic student record $\boldsymbol{x}^{(n)} = \{x^{(n)}_1, \ldots, x^{(n)}_{|\boldsymbol{X}|}\}$ by drawing values for each variable X_i from its conditional distribution given its parents in the learned graph G:

$$x^{(n)}_i \sim P(X_i \mid \mathrm{Pa}_G(X_i); \boldsymbol{\theta}_{\text{MAP}})$$

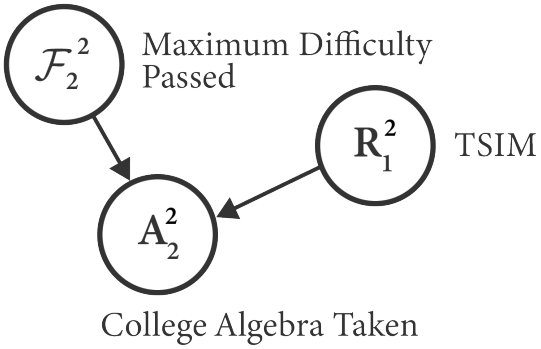
Sampling proceeds in topological order, starting with demographic variables $\boldsymbol{D}$, which are either unconditioned or depend only on other demographics. Given sampled demographics, we draw readiness variables $\boldsymbol{R}^0$ from their respective conditional distributions, followed by course-taking indicators $\boldsymbol{A}^0$ and grades $\boldsymbol{G}^0$, each sampled conditional on their respective parents. Finally, we sample the enrollment outcome E^0. This process repeats for subsequent timesteps $t = 1, 2, \ldots, T$, where $\boldsymbol{A}^t$, $\boldsymbol{G}^t$, and E^t are sampled, and $\boldsymbol{R}^t$ is deterministically updated using fixed rules based on completed coursework. Demographics $\boldsymbol{D}$ remain fixed throughout. The result is a synthetic dataset $\boldsymbol{X}^{\text{syn}}$ that preserves the structural dependencies and statistical properties of the original data $\boldsymbol{X}^*$. Algorithm 1 details this procedure.

We assess the realism of the generated synthetic dataset $\boldsymbol{X}^{\text{syn}}$ using a suite of validation techniques applied to a held-out evaluation set $\mathcal{D}_{\text{eval}}$. The dataset is partitioned into a training set $\mathcal{D}_{\text{train}}$, used to learn model parameters and define baselines, and an evaluation set $\mathcal{D}_{\text{eval}}$, reserved exclusively for validation. All comparisons to real data refer to this held-out evaluation set unless otherwise stated.

To contextualize results, we benchmark against two references. First, we compute divergence between $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{eval}}$ to quantify the degree of natural variability between samples — particularly important in high-dimensional, sparse datasets where many conditional patterns are under-observed. Second, we construct a naive baseline dataset $\boldsymbol{X}^{\text{baseline}}$ by sampling each variable independently



(a) Prerequisite subgraph used to assign difficulty scores. Each node represents a course, and edges represent prerequisite relationships. Difficulty scores are derived from topological rank in this graph (e.g., `Pre-Calc` has a rank of 3).



(b) Fixed edges for variable group A^2: $\mathcal{E} = \{\mathcal{F}_2^2 \to A_2^2,\ R_1 \to A_2^2\}$, encoding the hypothesis that College Algebra enrollment at $t = 2$ depends on the maximum math difficulty passed before $t = 2$ and on math readiness at $t = 2$.

Algorithm 1: Synthetic Student Trajectory Generation

Require: Graph G, CPTs $\boldsymbol{\theta}_{\text{MAP}}$, horizon T
  for each synthetic student $n = 1, \ldots, N$ **do**
    Sample demographics $\boldsymbol{D}^{(n)} \sim P(\boldsymbol{D}; \boldsymbol{\theta}_{\text{MAP}})$
    Sample readiness $\boldsymbol{R}^{t=0,(n)} \sim P(\boldsymbol{R}^0 \mid \boldsymbol{D}^{(n)}; \boldsymbol{\theta}_{\text{MAP}})$
    for $t = 0$ to T **do**
      if $t > 0$ **then**
        Update readiness $\boldsymbol{R}^{t,(n)} \leftarrow \text{Rules}(\boldsymbol{G}^{0:t-1,(n)})$
      end if
      Sample courses taken $\boldsymbol{A}^{t,(n)} \sim P(\boldsymbol{A}^t \mid \cdot;\ \boldsymbol{\theta}_{\text{MAP}})$
      Sample grades $\boldsymbol{G}^{t,(n)} \sim P(\boldsymbol{G}^t \mid \cdot;\ \boldsymbol{\theta}_{\text{MAP}})$
      Sample enrollment $E^{t,(n)} \sim P(E^t \mid \cdot;\ \boldsymbol{\theta}_{\text{MAP}})$
      if $E^{t,(n)}$ = exit or transfer **then**
        break
      end if
    end for
  end for
  return Synthetic dataset $\boldsymbol{X}^{\text{syn}} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$

from its empirical marginal distribution in $\mathcal{D}_{\text{train}}$:

$$x_i^{(n)} \sim P_{\text{emp}}(X_i), \quad \text{for } i = 1, \ldots, |\boldsymbol{X}|, \quad n = 1, \ldots, N$$
$$\text{where } P_{\text{marg}}(\boldsymbol{X}) = \prod_{i=1}^{|\boldsymbol{X}|} P_{\text{emp}}(X_i) \tag{3}$$

This ignores all inter-variable dependencies, yielding a product distribution in Eqn (3). The marginal baseline serves as a lower bound for realism, allowing us to isolate the value of modeling dependencies in $\boldsymbol{X}^{\text{syn}}$.

To assess whether the synthetic dataset preserves conditional dependencies, we examine conditional slices of the form $P(X \mid C)$, where C denotes a fixed condition on one or more variables (e.g., TSIM = 0), and X represents one or more downstream variables (e.g., course-taking indicators). For each selected condition, we compare the conditional distributions in both the real and synthetic datasets by computing the divergence $\text{KL}\big(P_{\text{emp}}(X \mid C) \,\|\, \hat{P}(X \mid C)\big)$ detailed in Eqn. (2).

5 Evaluation of Synthetic Student Pathways

Environment setup. Access to the educational dataset used in this study was governed by strict data use agreements and privacy requirements. Due to the sensitivity of student-level records, all data were stored and processed in a secure, HIPAA-compliant computing environment. Access was restricted to a small group of researchers who had received formal approval from the governing agency. No local copies of the data were permitted; all analysis was conducted via secure SSH connections to the remote server. While the original dataset cannot be shared due to privacy constraints, we provide examples and fully functional code to enable users to replicate the modeling pipeline end-to-end. All modeling and preprocessing code was written in Python and executed within this controlled environment. The primary libraries include `pgmpy` for Bayesian network modeling, `pandas` for data manipulation, and `numpy` for numerical computations, among others. A public version of the code is available in our GitHub repository to facilitate reproducibility.

Experimental setup. We evaluate the impact of both domain-informed fixed edge sets and modeling hyperparameters on the quality of generated synthetic data. As described in §4.2, we experimented with a subset of fixed edges $\mathcal{E}_{\text{fixed}}$ to incorporate known relationships between variables. Table 4 summarizes the best-performing fixed edge configurations identified for each variable group. For example, for math course selections in the very first timestep $t = 0$, TSIM emerges as the best performing predictor, while for non-math courses, TSIW is the strongest. In addition to edge constraints, we also varied parameters for structure learning, parameter estimation, and sampling. Table 5 lists the computational configurations explored, including choices of priors, pseudocounts, number of structure learning runs, sample sizes, and time horizon. Hyperparameters were tuned within practical ranges to assess sensitivity, and subsequent results report the configuration that achieved the best empirical fidelity among the values tried.

Category	Best Performing Fixed Edges
Demographics $\boldsymbol{D}$	$\epsilon_{\text{fixed}} = \emptyset$
Readiness $\boldsymbol{R}^{t=0}$	$\epsilon_{\text{fixed}} = \emptyset$
Course Choice $\boldsymbol{A}^{t=0}$	$\epsilon_{\text{fixed}} = \{R_1^{t=0} \rightarrow \boldsymbol{A}^{t=0}\}$ for math courses; $\epsilon_{\text{fixed}} = \{R_3^{t=0} \rightarrow \boldsymbol{A}^{t=0}\}$ for non-math courses
Course Choice $\boldsymbol{A}^{t>0}$	$\epsilon_{\text{fixed}} = \{\mathcal{F}_2^t \rightarrow \boldsymbol{A}^t\}$ for math courses; $\epsilon_{\text{fixed}} = \{(\mathcal{F}_{1,i}^t, \mathcal{F}_{3,i}^t), \rightarrow \boldsymbol{A}^t\}$ for non-math courses
Grades $\boldsymbol{G}^{t=0}$	$\epsilon_{\text{fixed}} = \{R_1^{t=0} \rightarrow \boldsymbol{G}^{t=0}\}$ for math courses; $\epsilon_{\text{fixed}} = \{R_3^{t=0} \rightarrow \boldsymbol{G}^{t=0}\}$ for non-math courses
Grades $\boldsymbol{G}^{t>0}$	$\epsilon_{\text{fixed}} = \{G_j^{t-1} \rightarrow G_i^t\}$ where G_j^{t-1} is the grade received in timestep $t-1$ for the most difficult prerequisite j for course i, if any.
Enrollment E^t	$\epsilon_{\text{fixed}} = \{\mathcal{F}_7^t \rightarrow E^t\}$

Table 4: Summary of best performing fixed edge configurations by variable group.

Table 6 reports the average KL divergence between predicted distributions and the held-out evaluation set $\mathcal{D}_{\text{eval}}$, evaluated at successive stages of the student pathway. Variables are accumulated at each step: the *Demographics* group contains 16 variables; *Readiness* adds three more (TSIM, TSIR, and TSIW); *Course Choice* introduces 117 binary indicators (one per course); and *Course Grade* expands to 253 variables. For clarity, we report results for one representative timestep. As expected, KL divergence increases as more variables are added, due to the growing dimensionality of the joint distribution in Eqn (2). The most pronounced jumps occur at the *Course Choice* and *Course Grade* steps, which reflect both a large increase in variable count and extreme sparsity — most students attempt only a small subset of the available courses. Despite this complexity, our model outperforms the baseline at all stages.

While both our model and the baseline method reproduce marginal distributions reasonably well, only our model captures dependencies across variables. This is expected: the baseline method samples each variable independently, so it cannot encode relationships. For instance, Figure 4 shows results for selected demographic features. Figure 4a shows the marginal distributions of individual demographic variables. Both methods produce similar distributions; however, Figure 4b reveals that only our model reproduces meaningful inter-variable correlations. For example, the variables LEP and Homemaker are strongly associated with Single Parent status in the real data.

We observe a similar trend in the course selection variables: both methods produce reasonable marginal distributions, but only our model captures conditional structure and higher-level patterns. Figure 5 presents results for enrollment inselected math courses at $t = 0$ and $t = 1$. On the left, Figure 5a shows conditional probabilities of enrollment in selected math courses at $t = 0$ given that the student is TSI-satisfied in math (TSIM). Our model correctly reflects that TSIM students are more likely to take College Algebra, and less likely to take Developmental Math or Math Prep, matching stated institutional requirements. On the right, Figure 5b shows the conditional distributions for enrollment in Plane Trigonometry, Developmental Math, and Math Prep, given that the student has already obtained credit for College Algebra by the end of the previous semester. As expected, Plane Trigonometry — one of the next courses in the math sequence — shows a markedly higher likelihood of being taken once College Algebra credit is earned. The model also captures the inverse relationship: students who have passed College Algebra are much less likely to move backward into lower-level coursework such as Developmental Math or Math Prep.

We next examine higher-order patterns, visualized in Figure 6. Figure 6a shows the predicted distribution of courses taken per student in a semester; our method closely matches the observed statistic, whereas the baseline does not. To provide a qualitative view of progression, Figure 6b presents a Sankey diagram illustrating pathways through selected variable groups. For clarity, we focus on a subset of variables. The first column shows degree intent, a stakeholder-relevant demographic attribute. The second column shows the flow of students who are TSIM-satisfied. The third column highlights the flow of students through two selected math courses: College Algebra and Dev Math. The final column shows enrollment outcomes. Colored flows represent the distribution in real data; overlaid crossed black outlines show corresponding flows from the synthetic data. In the real data, 49% of students who exited the system had previously received an F or incomplete in at least one of these math courses. The synthetic dataset captures a similar trend, with 53% of exited students exhibiting the same grade pattern. The close alignment shows that our model preserves key relationships. Such visualizations help stakeholders intuitively inspect synthetic progression patterns.

To support explainability and interpret model behavior, we visualize subgraphs of the learned Bayesian network in Figure 7. In Figure 7a, the green node represents the derived variable *maximum math difficulty passed* ($\mathcal{F}_2^t$ in Table 3). This feature consistently emerges as one of the strongest predictors of math course selection, surpassing even listed prerequisites. This reflects behavioral patterns observed in real life: students advance in difficulty rather than regress to easier courses. Figure 7b shows the parent nodes of the General Chemistry I enrollment decision at timestep t=2. Among the influencing variables — TSIW, TSIM, and the prior grade in College Algebra — the grade exerts the most influence, as evidenced by the thick edge. This was a surprising finding, given that TSIW and TSIM are the listed prerequisites and highlighted the importance of College Algebra as a gateway course.

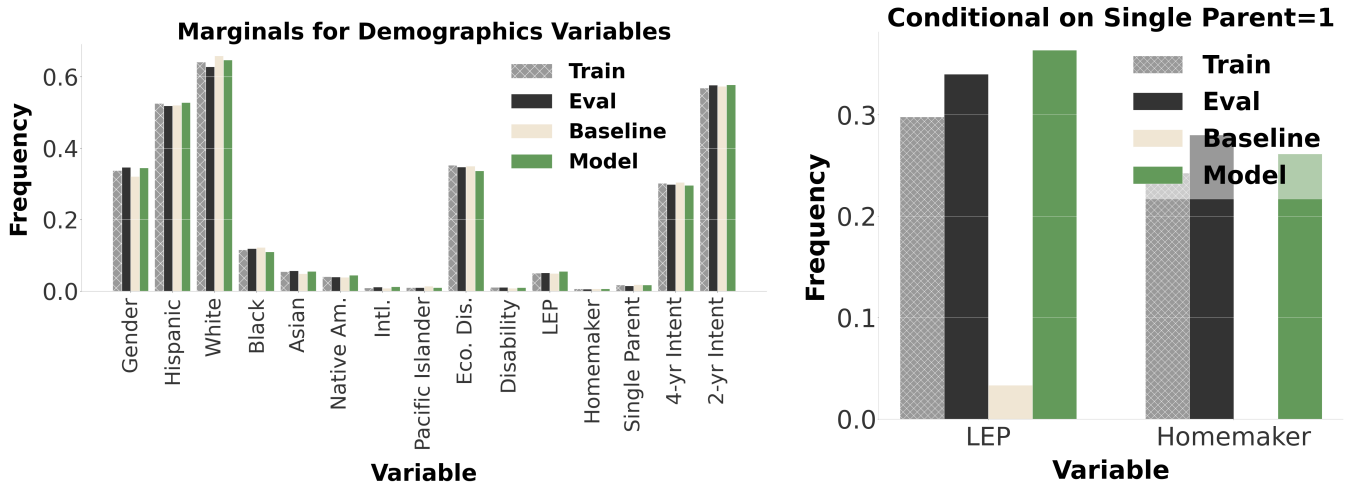
Privacy Considerations. Although synthetic data records are not directly traceable to any real individual, in practice, educational institutions often apply traditional disclosure thresholds, such as k-anonymity, to synthetic datasets before approving their release. To assess whether our method meets these current rules, we examine k-anonymity failure rates as the number of quasi-identifiers increases. Figure 8 shows the proportion of disclosive records — those with group size less than $k = 5$ — as we progressively accumulate variables from demographics to course grades. Even at the *Demographics* stage (16 variables), 10% of records fail. By the time 100 non-math course choice variables are included, nearly all records violate k-anonymity. Our synthetic data yield slightly fewer

Component	Parameter	Value(s) Used
Structure Learning	Score function used to evaluate graph structure	BIC
	Maximum number of parents allowed per node	2–5 (varied)
	Fixed edges $\mathcal{E}_{\text{fixed}}$ tried during search	See Table 4
	Python random seed	42
	Number of hill climb runs	1–5
Parameter Estimation	Prior types	Beta, Dirichlet
	α, β: Pseudocounts for Beta prior	1–10
	α: Pseudocounts per outcome for Dirichlet prior	1–10
	Prior strength	2–10
Sampling	Number of synthetic students generated	500–50000
	Sampling method from trained model	Forward sampling
Global	Train/eval split	80/20
	Number of terms modeled	6
	Course catalog size (after standardization)	113

Table 5: Modeling parameters and hyperparameters explored for structure learning, parameter estimation, and data generation.

Variable Group	# of Variables	Training Data	Predicted (Baseline)	Predicted (Model)
Demographics	16	0.312	1.565	**0.628**
Readiness	19	0.939	3.936	**1.702**
Course Choice	136	7.287	10.438	**8.934**
Course Grade	253	13.013	24.846	**18.911**
Enrollment Outcome	254	28.828	49.481	**39.232**

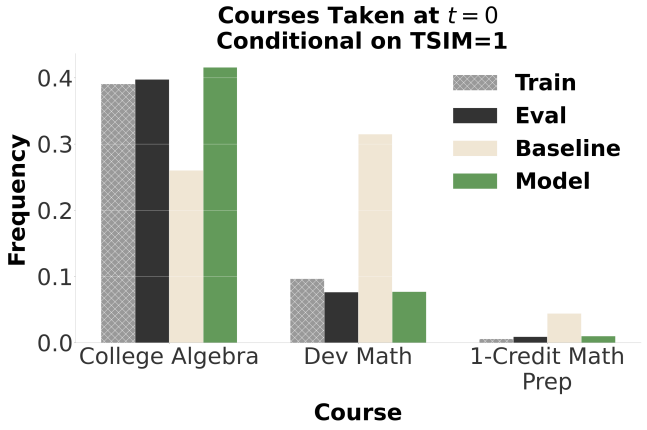
Table 6: Average KL divergence between predicted distributions and the evaluation set, grouped by variable stage. Variables accumulate at each stage. Our model consistently yields lower divergence than the baseline, indicating closer alignment with the evaluation data.



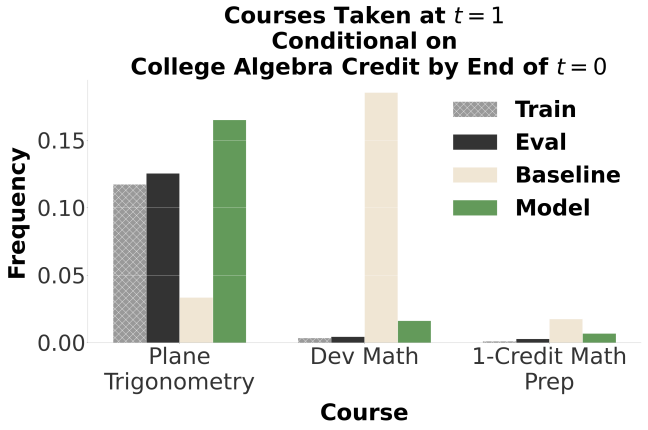
(a) Marginal distributions of demographic variables. Both our model and the baseline closely match observed frequencies.

(b) Conditional distributions of LEP and Homemaker given Single Parent. Our model better captures the increased likelihood of these variables being true.

Figure 4: Evaluation of selected demographic variables: marginals (top) and conditionals (bottom).

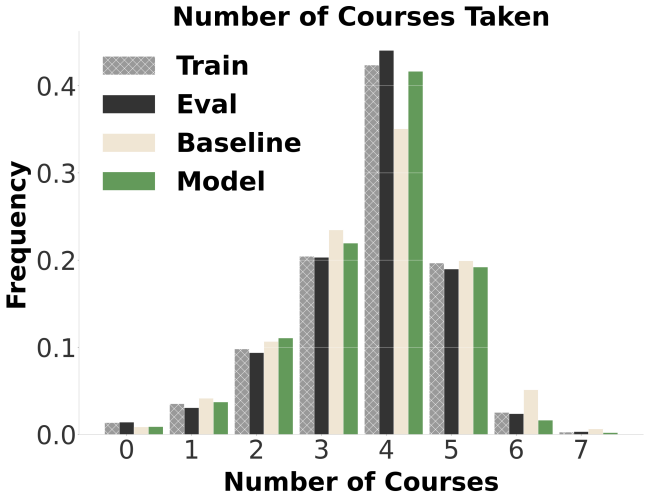


(a) Conditional distributions of enrollment in College Algebra, Developmental Math, and Math Prep given TSIM. Our model more accurately reflects course-taking patterns conditioned on readiness status.

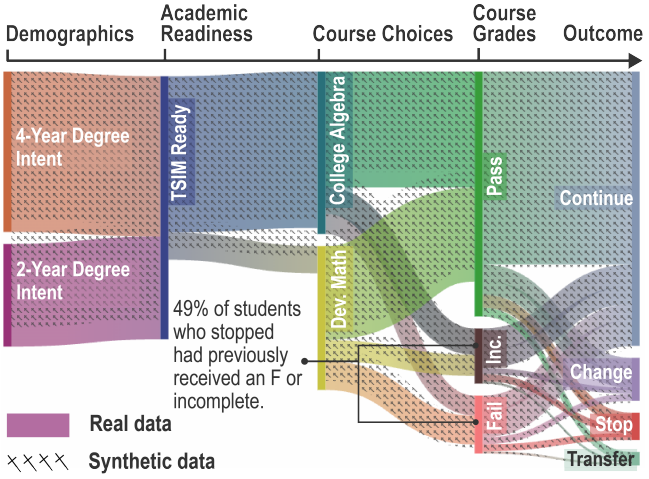


(b) Conditional distributions of enrollment in Plane Trigonometry at $t = 1$, given that credit has been obtained for College Algebra by the end of $t = 0$. Our model more accurately reflects temporal dependencies.

Figure 5: Comparison of course enrollment variables across conditional slices at $t = 0$ and $t = 1$.



(a) Distribution of the number of courses taken per student. Our model better captures higher-order structure, such as high correlations between certain courses (e.g., students rarely take both Dev Math and Calculus II), resulting in more realistic course load patterns compared to the baseline.



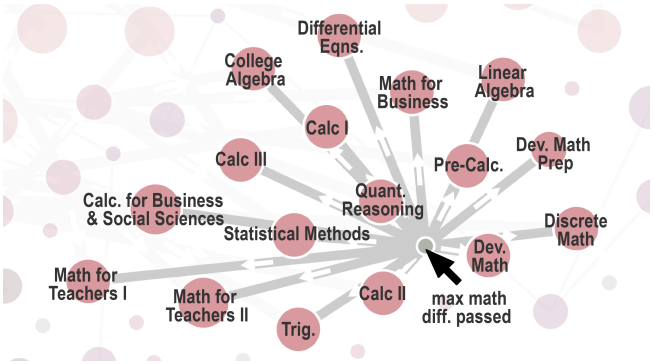
(b) Sankey diagram illustrating student flow through selected variables: degree intent, TSIM, College Algebra and Developmental Math, grades, and enrollment outcomes. Colored flows show real data; crossed black outlines show synthetic data flows generated by our model.

Figure 6: Visualization of two higher-order patterns: average course load in a semester and visual flow of synthetic students.

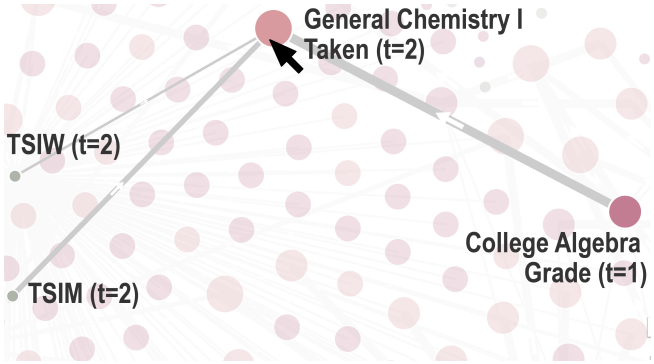
violations than the baseline, which samples variables independently and thus produces more implausible, unique records. However, the overall failure rate remains high and shows that k-anonymity is fundamentally ill-suited to the context of student pathways.

6 Conclusions

This paper introduces a generative method for producing longitudinal synthetic data of student pathways. Our approach uses structure learning to learn a dynamic Bayesian network (DBN) that captures dependencies across demographics, readiness, course choices, grades, and enrollment outcomes, while also modeling temporal dependence as the network expands with each additional semester. Empirical results show that our approach outperforms an independent sampler baseline in reproducing marginal and conditional distributions, particularly in later stages where sparsity and temporal dependencies increase. Our method offers explainability: each parameter in the model corresponds to interpretable relationships that stakeholders can inspect and adjust. The significance of



(a) Derived variable *maximum math difficulty passed* (green) is a key predictor of downstream math course choices (pink).



(b) Enrollment in General Chemistry I at $t = 2$ is predicted more strongly by College Algebra Grade at $t = 1$ than official TSIW and TSIM prerequsites.

Figure 7: Zoom-in portions of the learned Bayesian network visualizing strong dependencies for course choice.

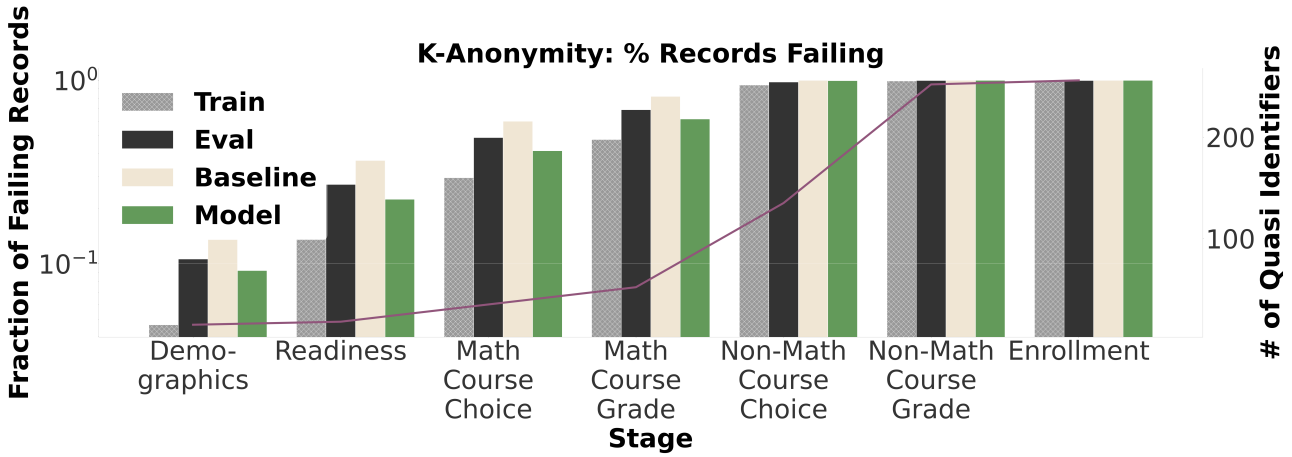


Figure 8: Proportion of records violating k-anonymity as quasi-identifiers accumulate. Our model reduces violations relative to the baseline, but both fail quickly, showing the limitations of k-anonymity for student pathway data.

our contribution lies in adapting DBNs to a domain characterized by sparse, longitudinal tabular records linked across multiple tables with strong temporal correlations. This provides a tractable and interpretable method for realistic educational data synthesis and for simulating policy questions. Finally, our k-anonymity analysis reveals that real student records become disclosive under even modest variable combinations, highlighting the impracticality of FERPA k-anonymization rules for pathway modeling. While our synthetic samples reduce disclosure rates, violations remain common because k-anonymity breaks down in sparse settings. Future work will explore privacy-aware methods for synthetic data generation of student pathways.

References

[1] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[2] Jock Blackard. 1998. Covertype. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C50K5N.

[3] Shahab Boumi and Adan Vela. 2019. Application of Hidden Markov Models to Quantify the Impact of Enrollment Patterns on Student Performance. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*.

[4] D. Chambliss and C. Takacs. 2018. *How College Works*. Harvard University Press, Chapter 1, 228.

[5] Nate Gruver, Ali Malik, Brahm Capoor, Chris Piech, Mitchell L. Stevens, and Andreas Paepcke. 2019. Using Latent Variable Models to Observe Academic Pathways. In *arXiv:1905.13383*.

[6] David Heckerman. 2008. *A Tutorial on Learning with Bayesian Networks*. Springer Berlin Heidelberg, Berlin, Heidelberg, 33–82. doi:10.1007/978-3-540-85066-3_3

[7] Luwen Huang and Karen E. Willcox. 2024. Educational Digital Twin: Tackling Complexity in Educational Big Data . In *2024 IEEE International Conference on Big Data (BigData)*. IEEE Computer Society, Los Alamitos, CA, USA, 1978–1985. doi:10.1109/BigData62323.2024.10825338

[8] Michael Kahn. [n. d.]. Diabetes. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5T59G.

[9] Neville Kenneth Kitson, Anthony C. Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. 2023. A survey of Bayesian Network structure learning. *Artificial Intelligence Review* 56, 8 (2023), 8721–8814.

[10] Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

[11] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. 2023. TabDDPM: modelling tabular data with diffusion models. In *Proceedings of the*

40th International Conference on Machine Learning (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 725, 16 pages.

[12] Jaewoo Lee, Minjung Kim, Yonghyun Jeong, and Youngmin Ro. 2022. Differentially Private Normalizing Flows for Synthetic Tabular Data Generation. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 7 (Jun. 2022), 7345–7353. doi:10.1609/aaai.v36i7.20697

[13] Naiming Liu, Zichao Wang, and Richard Baraniuk. 2024. Synthetic Context Generation for Question Generation. In *arXiv:2406.13188*. doi:10.48550/arXiv.2406.13188

[14] Qinyi Liu, Oscar Deho, Farhad Vadiee, Mohammad Khalil, Srecko Joksimovic, and George Siemens. 2025. Can Synthetic Data be Fair and Private? A Comparative Study of Synthetic Data Generation and Fairness Algorithms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*. Association for Computing Machinery, New York, NY, USA, 591–600. doi:10.1145/3706468.3706546

[15] Qinyi Liu, Mohammad Khalil, Jelena Jovanovic, and Ronas Shakya. 2024. Scaling While Privacy Preserving: A Comprehensive Synthetic Tabular Data Generation and Evaluation in Learning Analytics. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (Kyoto, Japan) *(LAK '24)*. Association for Computing Machinery, New York, NY, USA, 620–631. doi:10.1145/3636555.3636921

[16] Y. Liu, S. Bhandari, and Z. A. Pardos. 2025. Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology* (2025). doi:10.1111/bjet.13570

[17] Yunhui Long, Boxin Wang, Zhuolin Yang, Bhavya Kailkhura, Aston Zhang, Carl A. Gunter, and Bo Li. 2021. G-PATE: scalable differentially private data generator via private aggregation of teacher discriminators. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*.

[18] Rishabh Misra. 2022. News Category Dataset. *arXiv preprint arXiv:2209.11429* (2022).

[19] Kevin Patrick Murphy. 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph. D. Dissertation. University of California Berkeley.

[20] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy. 2002. Dynamic Bayesian Networks for Audio-Visual Speech Recognition. *EURASIP Journal on Advances in Signal Processing* 2002, 11 (2002), 783042.

[21] Mauro Scanagatta, Antonio Salmerón, and Fabio Stella. 2019. A survey on Bayesian network structure learning from data. *Progress in Artificial Intelligence* 8, 4 (2019), 425–439.

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

[23] Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. 2019. Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning* 115 (2019), 235–253. doi:10.1016/j.ijar.2019.10.003

[24] Juntong Shi, Minkai Xu, Hengrui Zhang Harper Hua, Stefano Ermon, and Jure Leskovec. 2025. TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation. In *arXiv:2410.20626*.

[25] Shashank Sonkar, Naiming Liu, Xinghe Chen, and Richard Baraniuk. 2025. Turing-Like Test for Personalized Educational AI. In *Artificial Intelligence in Education*. Springer Nature Switzerland, 405–412.

[26] THECB. 2017. *Reporting and Procedures Manual for Texas Community, Technical, and State Colleges. Fall 2017*. Technical Report. Texas Higher Education Coordinating Board.

[27] Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning* 65, 1 (2006), 31–78.

[28] Jill-Jênn Vie, Tomas Rigaux, and Sein Minn. 2022. Privacy-Preserving Synthetic Educational Data Generation. In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: 17th European Conference on Technology Enhanced Learning, EC-TEL 2022, Toulouse, France, September 12–16, 2022, Proceedings*. 393–406. doi:10.1007/978-3-031-16290-9_29

[29] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. *Modeling tabular data using conditional GAN*. Curran Associates Inc., Red Hook, NY, USA.

[30] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. PrivBayes: Private Data Release via Bayesian Networks. *ACM Trans. Database Syst.* 42, 4, Article 25 (2017), 41 pages. doi:10.1145/3134428