This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2020.2981331, IEEE Transactions on Learning Technologies

Characterizing Curriculum Prerequisite Networks by a Student Flow Approach

Roland Molontay, *Member, IEEE,* Noémi Horváth, Júlia Bergmann, Dóra Szekrényes and Mihály Szabó

***Abstract*—Curriculum prerequisite networks have a central role in shaping the course of university programs. The analysis of prerequisite networks has attracted a lot of research interest recently since designing an appropriate network is of great importance both academically and economically. It determines the learning goals of the program and also has a huge impact on completion time and dropping out. In this paper, we introduce a data-driven probabilistic student flow approach to characterize prerequisite networks and study the distribution of graduation time based on the network topology and on the completion rate of the courses. We also present a method to identify courses that have a significant impact on graduation time. Our student flow approach is also capable of simulating the effects of policy changes and modifications of the network. We compare our methods to other techniques from the literature that measure structural properties of prerequisite networks using the example of the electrical engineering program of Budapest University of Technology and Economics.**

***Index Terms*—curriculum design, discrete-event simulation, electrical engineering curriculum, degree completion time, higher education, prerequisite network, student flow**

I. Introduction

COLLEGE years are usually referred to as "*the best part of one's life*", although as far as these years go, a lot depends on the curriculum of the university program. In this paper, we analyze curriculum prerequisite networks based on a data-driven probabilistic student flow approach.

Graduating as soon as possible is not only the students' interest but it is also important from the institution's point of view. Delayed completion and dropping out are common academic problems – especially in STEM higher education – which should be minimized since they subsequently squander human and economic resources. An important question is how restrictive a university curriculum should be, how to set the prerequisite constraints. It cannot be too restrictive since the higher education directorate aims to increase the rate of completion and shorten the time needed to graduate, on the other hand, the program must train good specialists.

In this paper, we consider university programs where the curriculum is quite regulated, i.e. to fulfill the requirements students have to take well-specified courses in a sequential order controlled by the corresponding prerequisite network that is quite common in STEM programs. The prerequisite network is a directed graph (network) where the nodes correspond to the courses (also referred to as subjects or classes) and an edge goes from one course to another if the former course is a prerequisite of the latter one. The analysis of the prerequisite network is extremely important since it determines the learning goals of the program, moreover, the structure of the network has a huge impact on dropout rates and on graduation time [1], [2]. Here we characterize university curricula by a data-driven probabilistic student flow approach and determine the expected graduation time by considering both the topology of the prerequisite network and the completion rates of the courses. We also introduce a novel method to mathematically measure the *importance* of a course with respect to its relative impact on graduation time.

It was shown by several authors that curriculum organization has a high influence on study progress in higher education. Jansen studied the relationship between curriculum organization and first-year academic success and identified the key factors that affect students' study progress the most. Namely, spreading exams (i.e., the number and timetable of test periods), programming fewer parallel courses and not spreading re-test over the whole year have the highest positive contribution on academic success [3]. Robinson studied the patterns of individual pathways to monitor the process and outcomes of student progression [4].

Measuring curricular efficiency and analyzing the effects of curriculum organization on academic progress in higher education by a student flow approach have been in the focus of research interest for decades [5]–[7]. The most frequent approach is to model student flows by Markov chains to answer questions like what the mean time is that a student takes to complete a course or that a student spends in higher education [5], [8], [9]. Shah and Burke used Markov chains to model the movements of undergraduate students through the Australian higher education system [9]. Student characteristics such as age and gender are also taken into consideration by their model. Bessent and Bessent analyzed the progression of doctoral students using a Markov approach [5]. Brezavscek *et al.* studied the transition between different stages of a Slovenian study program based on Markov analysis [10].

Markov assumption may be too restrictive thus other authors

R. Molontay is with MTA-BME Stochastics Research Group and also with the Department of Stochastics, Budapest University of Technology and Economics, P.O. Box 91, 1521 Budapest, Hungary. (molontay@math.bme.hu)

N. Horváth is with SDA Informatika Zrt. and also with the Department of Stochastics, Budapest University of Technology and Economics, P.O. Box 91, 1521 Budapest, Hungary. (veruna@math.bme.hu)

J. Bergmann is with the Institute for Computer Science and Control, Hungarian Academy of Sciences and also with the Department of Stochastics, Budapest University of Technology and Economics, P.O. Box 91, 1521 Budapest, Hungary. (bjulia@math.bme.hu)

D. Szekrényes is with the Department of Stochastics, Budapest University of Technology and Economics, P.O. Box 91, 1521 Budapest, Hungary (szedola@math.bme.hu)

M. Szabó is with Central Academic Office, Budapest University of Technology and Economics, P.O. Box 91, 1521 Budapest, Hungary (szabo.mihaly@kth.bme.hu)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2020.2981331, IEEE Transactions on Learning Technologies

rather use a more flexible computer simulation approach to capture the complexity of the system [1], [6], [11]. Plotnicki and Garfinkel have proposed a simulation model to schedule courses in such a way that it allows as many students to flow smoothly through the curriculum as possible while keeping a feasible schedule for the department, too [6]. Mansmann and Scholl have introduced a decision support system to evaluate programs and curriculum modifications by simulation models [12]. Schellekens *et al.* presented a discrete-event simulation model for designing higher educational programs in the Netherlands [13]. Saltzman and Roeder have developed a discrete-event simulation model that allows for changes in curriculum policy and structure [1]. Saltzman *et al.* presented a model that simulates the flow of undergraduate students at a state university in California to test the potential impact of course sequencing, pass rates, retention rates, capacities, and enrollment [14]. Weber has developed a decision support system based on discrete-event simulation to help curriculum planners to achieve the maximal success of students [15].

Another line of research is to measure curricular complexity and the structure of curriculum prerequisite networks with the tools of network theory. Using complex network analysis and graph theory, Slim *et al.* have proposed a framework to study the structure of prerequisite networks and analyze the complexity of university curricula according to course criticality [16]–[19]. Slim *et al.* also used Markov networks to represent curriculum graphs to predict student performance [20]. Heileman *et al.* presented a curricular analytics approach to characterize and compare the curricular complexity of engineering programs at different institutions [21]. Heileman *et al.* summarized recent works related to curricular analytics and have introduced a framework to support curriculum-based improvement efforts [2].

Software applications have also been proposed to assist students to create personalized study plans and staff to maintain curriculum structure; such as Curriculum GPS by Akbas *et al.* [22] and STOPS (Software for Target-Oriented Personal Syllabus) developed by Auvnen *et al.* [23]. A curriculum analysis and simulation library has also been developed by Hickman [24].

Data-driven approaches in higher education have received a lot of attention recently from higher education researchers and policy-makers as well [25], [26]. Wigdahl has introduced a statistical model that can predict students graduation rate depending on institutional variables (e.g. semester grade point averages) and pre-institutional variables (e.g. high school performance data) [27]. Furthermore, student characteristics (e.g. gender, age) are also thought to play an important role and are taken into consideration by a number of papers [3], [4], [28]. Mendez *et al.* propose a data-based course difficulty estimation and measure the influence of a course on students overall academic performance to support curriculum designers in identifying the courses that should be revised due to their difficulty level [29].

Several approaches have also been proposed to visualize student flow patterns. Horváth *et al.* developed an efficient visualization tool to analyze student flow patterns by alluvial and Sankey diagrams that allows decision-makers to gain a better insight on how students are processing and it also makes easier to understand the effects of policy changes on retention and graduation rates [30]. Raji *et al.* present a data-driven system called eCamp that is able to model and visualize student flow patterns on three levels: on a campus level, where students flow through all degree programs; on a department level, where student flow through the curriculum structure within a degree program; on a classes level, where student flow through classes [31]. The main difference between the approach of Raji *et al.* and the one presented in this paper is that while they consider a rather flexible curriculum, our modeling framework is suitable for a context where students have a declared major from the very beginning of their studies and must follow a quite restrictive curriculum.

This paper combines curriculum prerequisite network analysis with discrete-event computer simulation modeling by introducing a data-driven probabilistic student flow approach to characterize prerequisite networks. Most of the related papers working with a student flow approach consider university programs with a quite flexible curriculum where students can choose from a variety of course options. However, our approach is rather developed for a strict curriculum where the path to earn the degree is rather strictly determined by the prerequisite network. This highly regulated aspect of the curriculum has enabled us to build a more analytical framework for curriculum analysis.

Besides the topological structure of the network, we also consider the completion rates of the courses based on real historical data. We introduce novel metrics to characterize prerequisite networks based on a data-driven probabilistic student flow approach. We present a model that can answer questions such as what the expected graduation time of the program is and which course has the greatest effect on the graduation time. Furthermore, the impact of policy changes and modification of the prerequisite network can be better analyzed and understood with the help of our framework. We also investigate the model analytically, however, computing the analytical solution is intractable, so we rely on discrete-event simulation. Using the example of the electrical engineering (EE) program of Budapest University of Technology and Economics (BME), we also compare our techniques to other methods from recent literature that characterize the topological structure of prerequisite networks. We present a software tool for analyzing prerequisite networks based on our proposed approach and we also discuss how it can support a wide range of educational stakeholders such as curriculum designers, administrators, and students.

II. Curriculum Prerequisite Network

In this section, numerous topological metrics are reviewed to characterize prerequisite networks, we also introduce tools for curriculum analysis based on a data-driven probabilistic student flow approach. Furthermore, we demonstrate these concepts in short examples.

A. Graph representation

A university curriculum is represented by directed a graph $G = (V, E)$. The graph can be given by the adjacency matrix

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2020.2981331, IEEE Transactions on Learning Technologies

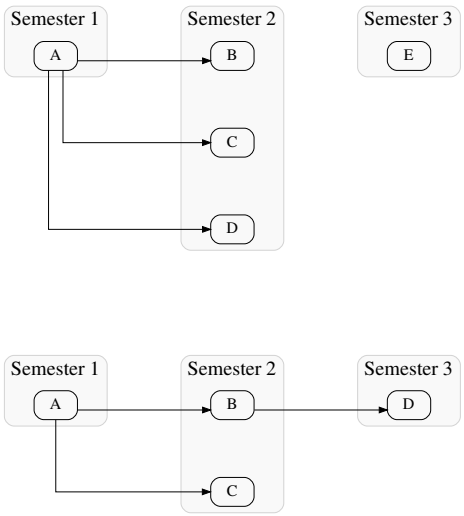

Fig. 1. Two simple sample curricula.

M for which $M_{ij} = 1$ if the ith course is a prerequisite of the jth course and $M_{ij} = 0$ otherwise. It is easy to see that G is a directed acyclic graph (DAG) as if it had a cycle it would mean that some i subject must have been completed to enroll some j course and vice versa.

In this paper, we use the simplifying assumption that prerequisite is a strict binary condition, i.e. weak prerequisites are not taken into consideration. Weak prerequisite refers to the fact that enrolling the course and its weak prerequisite in parallel is also permitted.

Next, notions are presented that measure the structural complexity of the curriculum and correspond to the roles of the courses in the prerequisite network based on the topology of the graph.

B. Topological indicators

Deferment factor: Some courses have more effect on completing the program on time than others. Motivated by the notion of delay factor by Slim *et al.* [16], we introduce the concept of deferment factor to represent if the failure of a certain course is necessarily followed by an increment of graduation time or not. Let us define the deferment factor of a compulsory course to be $1/(k+1)$ where k is the maximum number of possible enrollments in the course that does not increase the graduation time based on the curriculum graph, i.e. the student can fail the course k times without increasing the time of graduation, however with $k+1$ failures, one must have at least one extra semester. For example, if the deferment factor of a course is $1/3$ then students may fail that course twice but after the third failure, the graduation time necessarily increases.

Fig. 1 illustrates two simple curricula. It can be seen that in the first one even if the student fails course A (s)he has the chance to finish on time (i.e. in three semesters) while in the second curriculum the failure of all subjects except for course C causes at least one semester of delay. In this second case, one may say that courses A, B and D are more crucial than C, since the deferment factors of A, B and D are all 1 while the deferment factor of C is $1/2$.

Blocking factor: It is also natural to say that a course which is prerequisite to more courses is more crucial, this idea is reflected by the blocking factor introduced in [16]. Formally, we say that a given course has blocking factor $n \in \mathbb{N}$ if it has exactly n descendants in the graph, i.e., it is not equivalent of the out-degree of a vertex since it also counts all the descendants, not just the direct ones.

For example, take a look at Fig. 1. It is easy to see that course A blocks the same number of courses in both curricula, while B is more crucial in the bottom one.

Betweenness centrality: The betweenness centrality measure helps us find vertices that form a *bridge* between different parts of the networks. The application of betweenness centrality in a curriculum context was suggested in [32]. Although it can help identifying the key links between program tracks, the interpretation of betweenness centrality in the curriculum context is limited due to the special structure of prerequisite networks (namely they are DAG). It can be calculated as follows:

$$b_v = \sum_{\substack{s \neq v, t \neq v \\ s,t,v \in V(G)}} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

where σ_{st} is the total number of the shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v [32].

Connected components: A connected component is a maximal set of nodes such that each pair of nodes is connected by a path. While the prerequisite graph is directed, here we consider it as an undirected graph, i.e. we consider its weakly connected components.

Connected component analysis of a curriculum prerequisite network shows whether the curriculum is divided into independent, disconnected clusters of courses that do not serve as prerequisites of each other. A connected component in a curriculum graph may represent an independent knowledge community [33].

Long paths: The analysis of long paths in prerequisite networks was proposed in [17]. A path represents a chain of courses that must be taken in sequential order. Failing a course that is part of a long chain often implies falling behind by a semester (or a year). The definition of a *long* path may vary but for a seven-semester-long program, a path with a length of five (five edges and six nodes) or more can be definitely considered as a long path.

Bottlenecks: Considering the in- and out-degree (notation: $\deg^-(v)$ and $\deg^+(v)$) of the nodes, university programs have so-called *bottleneck* courses, a concept introduced in [17]. A course is said to be a bottleneck if it has in-degree or out-degree greater than or equal to a or b, respectively, or the total degree is greater than or equal to c where $a, b, c \in \mathbb{N}$, $a + b - 2 \geq c$ are fixed numbers. Formally, the number of bottlenecks that a program has is given as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2020.2981331, IEEE Transactions on Learning Technologies

$$b_n(G) = \sum_{v \in V(G)} \mathbb{1}\Big\{ \big[\deg^-(v) \geq a\big] \vee \big[\deg^+(v) \geq b\big] \vee \vee \big[\deg^+(v) + \deg^-(v) \geq c\big] \Big\}. \quad (2)$$

Other topological metrics have been proposed for curriculum analysis throughout the years [16], [17], [32], [33]. These concepts (including the presented ones) only take into account the topological structure of the graph but they miss out the fact that courses may have very different completion rates. To solve this deficiency, a novel framework is introduced which takes into account the rates of course completion as well as the topological structure of the graph.

C. Student flow based indicators

Here we present a student flow based simulation approach to characterize prerequisite networks considering both its structural topology and course completion rates estimated from historical data.

Expected graduation time: A university program can be characterized by the graduation time (i.e. the number of terms needed to complete all the required courses) as a discrete random variable X and by its expected value $\mathbb{E}(X)$. Let $p(x)$ be the probability mass function of the graduation time:

$$p(x) = \mathbb{P}(X = x). \quad (3)$$

The expected graduation time can be calculated as follows:

$$\mathbb{E}(X) = \sum_x x \cdot p(x). \quad (4)$$

The (expected) graduation time depends on the structure of the prerequisite network and the course completion probabilities. Suppose that the expected graduation time can be expressed in the following form: $\mathbb{E}(X) = f(p_1, ...p_n)$, where p_i denotes the completion rate (probability) of the ith course and the function f is determined by the structure of the prerequisite network.

Pass-through effect: The question naturally arises what impact it would have on the (expected) graduation time if the completion rate of a certain course was increased while the others remain unchanged. Mathematically, it can be represented by the following partial derivative:

$$D_i = \frac{\partial f(p_1, p_2, ..., p_n)}{\partial p_i}. \quad (5)$$

Similarly, the elasticity of a course completion rate is also a proper measure that can be defined as follows:

$$\varepsilon_i = \frac{\partial \log f(p_1, p_2, ..., p_n)}{\partial \log p_i}. \quad (6)$$

The main advantage of the previously defined concepts is the fact that they do not only rely on the topological structure of the graph but the course completion rates are also taken into consideration. In the next section, this approach is discussed in more detail. In real-life scenarios, function f is not known but both the expected graduation time and pass-through effects can be approximated by a discrete-event simulation.

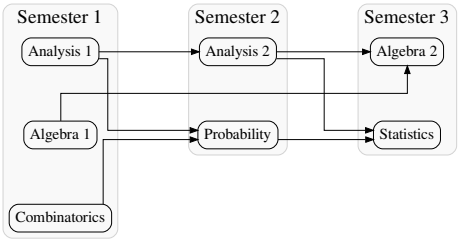

Fig. 2. Sample Curriculum.

III. Probabilistic student flow approach

In this section, we determine the distribution of the number of terms needed for successfully completing a university program, it is calculated analytically and approximated using a discrete-event simulation framework. We consider a virtual representative student, who is progressing in his/her studies according to the prerequisite network and if (s)he is enrolled in the ith course (s)he completes the course with probability p_i (a course-dependent constant probability independent from the past). This probability can be estimated based on historical data:

$$\hat{p}_i \approx \frac{\text{number of students accomplishing course } i}{\text{number of students enrolled in course } i}. \quad (7)$$

Our modeling approach uses the following assumptions:

- A course cannot be taken earlier than the recommended term.
- The representative student of our model attempts to fulfill all courses in the first possible term that is allowed by the prerequisite network.
- All courses are launched in every term. (This assumption will be relaxed later.)
- The completion probability of a course is assumed to be constant – i.e., it does not matter if one tries it for the first, second, etc. time.
- The success of courses are independent of each other, that is the fact of completing a course does not affect the success probability of another course.
- The representative student enrolls in a course until (s)he completes it no matter how many times (s)he fails it.
- There are no passive semesters[1].

A. Analytical solution

Let X_i denote a random variable that measures the number of attempts that one needs for the completion of the ith course. Assuming independent attempts, the X_i random variables are geometrically distributed (the number of Bernoulli trials needed to get the first success). If the success rate is p_i, i.e the random variable is distributed as $X_i \sim Geom(p_i)$ then the expected number of attempts of completing the course is $\mathbb{E}(X_i) = 1/p_i$.

[1] A passive semester is a semester when the students legal status is paused.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2020.2981331, IEEE Transactions on Learning Technologies

TABLE I
THE RANDOM VARIABLES CORRESPONDING TO THE ATTEMPTS AND THE NUMBER OF TERMS TO COMPLETE A COURSE COUNTED FROM THE TIME OF ENROLLMENT BASED ON THE SAMPLE CURRICULUM IN FIG. 2.

Course	Attempts	Terms to complete
Analysis 1	X_1	$Y_1 = X_1$
Algebra 1	X_2	$Y_2 = X_2$
Combinatorics	X_3	$Y_3 = X_3$
Analysis 2	X_4	$Y_4 = Y_1 + X_4$
Probability	X_5	$Y_5 = \max\{Y_1, Y_3\} + X_5$
Algebra 2	X_6	$Y_6 = \max\{Y_2, Y_4\} + X_6$
Statistics	X_7	$Y_7 = \max\{Y_4, Y_5\} + X_7$

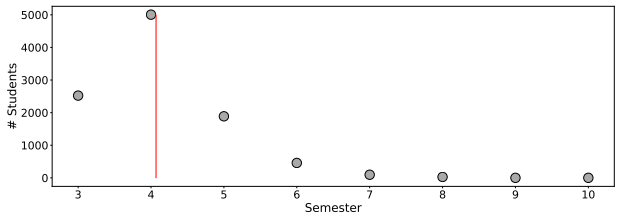

Fig. 3. Distribution of graduation time simulated on the sample curriculum from Fig. 2 with $p_i = 0.8$ for all courses. The mean graduation time is $\mu = 4.069$ (red line) and the standard deviation is $\sigma = 0.87$.

Let us consider the sample curriculum from Fig. 2. We determine the random variable Y_i corresponding to the number of terms that is needed to complete the ith course counted from the time of enrollment. It is clear that in the case of Analysis 1 Y_1 is equal to X_1. Regarding the course Probability, it is more complicated since its prerequisites, both Analysis 1 and Combinatorics must be completed before, that is: $Y_5 = \max\{Y_1, Y_3\} + X_5 = \max\{X_1, X_3\} + X_5$. For Statistics it is even more complex: $Y_7 = \max\{Y_4, Y_5\} + X_7 = \max\{X_1 + X_4, \max\{X_2, X_3\} + X_5\} + X_7$. All the other random variables can be seen in Table I.

Therefore, to calculate the expected value of these random variables, we have to calculate the expected value of the maximum of geometric random variables. If $X_1, X_2, \ldots, X_n$ are independent identically distributed geometric random variables with parameter p and M_n is the maximum of these random variables i.e. $M_n = \max\{X_1, X_2, ..., X_n\}$ then

$$\mathbb{E}(M_n) = \sum_{k=0}^{\infty} \left(1 - \left(1 - q^k\right)^n\right), \tag{8}$$

where $q = 1 - p$. The proof can be found in [34].

This sum is not easily countable, we can only approximate its value. Moreover, in our task the parameters of each X_i can be different that makes the calculation more difficult:

$$\mathbb{E}(M_n) = \sum_{k=0}^{\infty} \left(1 - \prod_{i=1}^{n} \left(1 - q_i^k\right)\right), \tag{9}$$

where $q_i = 1 - p_i$.

Going further, it becomes even more difficult to calculate the expected number of terms of Statistics since now we do not have the maximum of geometric random variables (since the maximum of geometric random variables is not geometric). Even though it is a well defined random variable, determining its distribution analytically requires more effort than it seems to. It implies that analytically calculating the distribution of the number of semesters needed for graduation is quite challenging. Hence we use a Monte Carlo method to simulate the distribution of the random variables and calculate their expected values.

B. Discrete-event simulation

Using the statistical software R, we simulate 10 000 representative students virtually attending the university program with the curriculum represented in Fig. 2. We set the completion probability to 0.8 for each course since it is a realistic average completion rate. We obtain the distribution of the graduation by simulating a student flow, i.e., the path of 10 000 representative students from the first semester until graduation. Fig. 3 shows the distribution of the graduation time given by our model.

The pass-through effect of a course can be also approximated with our simulation framework. The question is what effect it has on the (expected) graduation time if the success probability p_i of the ith course is increased while the other probabilities remain unchanged. The increase we consider can be additive or multiplicative and if multiplicative it can be proportional to the completion probability p_i or to the probability of failing the course $1 - p_i$. Formally, let h be a small positive number, the three approaches can be summarized as follows:

$$p_i^{(1)}(h) = \min\{p_i + h, 1\} \tag{10}$$

$$p_i^{(2)}(h) = \min\{p_i(1 + h), 1\} \tag{11}$$

$$p_i^{(3)}(h) = \min\{1 - (1 - p_i)(1 - h), 1\}. \tag{12}$$

To quantify the impact of increasing the course completion probability of the ith course, we approximate the pass-through effect of the ith course by:

$$d_i^{(j)}(h) = \frac{\hat{f}(p_1, \ldots, p_{i-1}, p_i^{(j)}(h), p_{i+1}, \ldots, p_n)}{\hat{f}(p_1, \ldots, p_{i-1}, p_i, p_{i_1}, \ldots, p_n)}, \tag{13}$$

where $\hat{f}(.)$ stands for the function that approximates the expected graduation time given the completion probabilities, $j \in \{1, 2, 3\}$ shows the type of probability increase and h is a small positive number.

To compare the pass-through effects of different courses, we believe that (12) is the most reasonable success rate modification approach since it measures the effect of letting a fixed h ratio of failing students pass the course. We measure the percentage change in the mean graduation time for each course with a fixed reasonably chosen h value.

For the sample curriculum from Fig. 2 the effects of increasing the success rate for each course separately using (12) with $h = 1$ can be found in Table II. It is clear that the mean graduation time decreased in all cases, while Statistics has the largest effect: if 100% of failing students manage to pass that course, the mean graduation time drops by 5.57%.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2020.2981331, IEEE Transactions on Learning Technologies

TABLE II
APPROXIMATIONS OF PASS-THROUGH EFFECTS FOR THE COURSES OF THE SAMPLE CURRICULUM FROM FIG. 2. THE PASS-THROUGH EFFECT IS APPROXIMATED BY $1 - d_i^{(3)}(1)$.

	Approximated pass-through effect
Analysis 1	0.0337
Algebra 1	0.0037
Combinatorics	0.0198
Analysis 2	0.0303
Probability	0.0246
Algebra 2	0.0418
Statistics	0.0557

IV. ANALYSIS OF THE PREREQUISITE NETWORK OF THE ELECTRICAL ENGINEERING PROGRAM AT BME

In this section, we study the curriculum of the Electrical Engineering (EE) BSc program (Embedded and Control Systems Specialization) at Budapest University of Technology and Economics (BME) [35]. The program is appropriate for the analysis since it has a high number of students (∼450 incoming students each year) which makes the estimation of the completion rates more reliable. We examine the operative curriculum that is in use since 2014. Fig. 4 shows the prerequisite network of this 7-semester-long program.

A. Topological metrics

We can see that a quite strict prerequisite network corresponds to the EE program. Regarding its topological metrics, we can observe the followings. It has some long paths and the longest path consists of six courses (Mathematics A1a - Calculus → Signals and Systems 1 → Signals and Systems 2 → Measurement Technology → Laboratory Exercises 1 → Laboratory 2). Since the program is designed for seven semesters, to graduate on time students may fail one of these courses only once. Using the definition of bottlenecks with parameters $a = 3$, $b = 3$ and $c = 4$, the following courses turned out to be bottlenecks: Mathematics A1a - Calculus, Signals and Systems 1 and Signals and Systems 2. Table III shows betweenness, pass-through effect, deferment and blocking factors for each obligatory course. We can observe that Signals and Systems 1 has the highest betweenness centrality thus it forms a bridge between many courses. The courses of the longest path and courses from semester 6 have the highest deferment factor; while Mathematics A1a - Calculus blocks the highest number of courses, namely it is the (not necessarily direct) prerequisite of 17 courses.

B. Student-flow based indicators

Fig. 5 shows the distribution of graduation time of the representative student regarding the EE program according to our model. The completion probabilities of obligatory courses are estimated using historical data from the educational administrative system, while the completion probabilities of elective courses are set to 1. At BME if the student fails to obtain the leaving certificate upon the expiry of twice the program duration, (s)he gets terminated by dismissal [36]. According to our model, 93.7% of students finish within 14 semesters (twice the program duration). It is important to note that the unrealistically long tail of the distribution is due to the fact that according to our model students enroll in a course until (s)he successfully completes it no matter how many times (s)he fails it. Failing Course 1 for at least 100 times has positive probability (namely $(1 - p_1)^{100}$), although in real life such students would give up earlier. This is the reason why we also get large values for graduation time.

We also measure how the mean graduation time decreases if the completion probability is increased for each course separately (pass-through effect) in a way described in the previous section with $h = 1$. The results are shown in Table III. By these results, we obtain that Introduction to Electromagnetic Fields has the highest effect on graduation time i.e. if we decrease the probability of failure of this course to zero then the expected graduation time decreases by 5.37%. The reason behind this is that this course has the lowest completion rate (as it is also illustrated in Fig. 9).

C. Credit point distribution over semesters

In Hungary, the European Credit Transfer and Accumulation System (ECTS) is used and one semester corresponds to 30 credit points. Using our modeling framework the worth of credit points that students attempt to complete and successfully fulfilled in each semester can be investigated. In Fig. 6, we can see the average enrolled and acquired credits of students per semester. It shows that the fourth and fifth semester can cause some hurdle for them since students attempt to catch up on failed courses.

V. EFFECT OF POLICY CHANGES

In this section, we demonstrate how the presented simulation framework can be used to gain a better understanding of the effects of policy changes and modifications of the curriculum. Namely, we investigate how the number of dismissed students vary if the regulation is modified and what effect the launching frequency of a course has on the graduation time. We note that our simulation framework can also be used to answer a wide variety of questions regarding curriculum design (such as the effect of success ratios of the courses and designing the prerequisite network).

A. Investigation of dismissed students

An important issue is to keep the standards of a university as high as possible. Several universities have a rule that indicates the maximum number of terms that a student can spend on studying in a program or the maximum times that (s)he can attempt to complete a course. At Budapest University of Technology and Economics, the number of terms is maximized in two times of the program duration time and the failed attempts of a course are maximized in 6. If students exceed one of the previous limits without graduation, they get dismissed [36]. Here we analyze the effect of varying the threshold of maximum failed attempts of a course on the number of dismissed students. Fig. 7 shows the histograms of the number of dismissed students corresponding to different threshold settings.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2020.2981331, IEEE Transactions on Learning Technologies

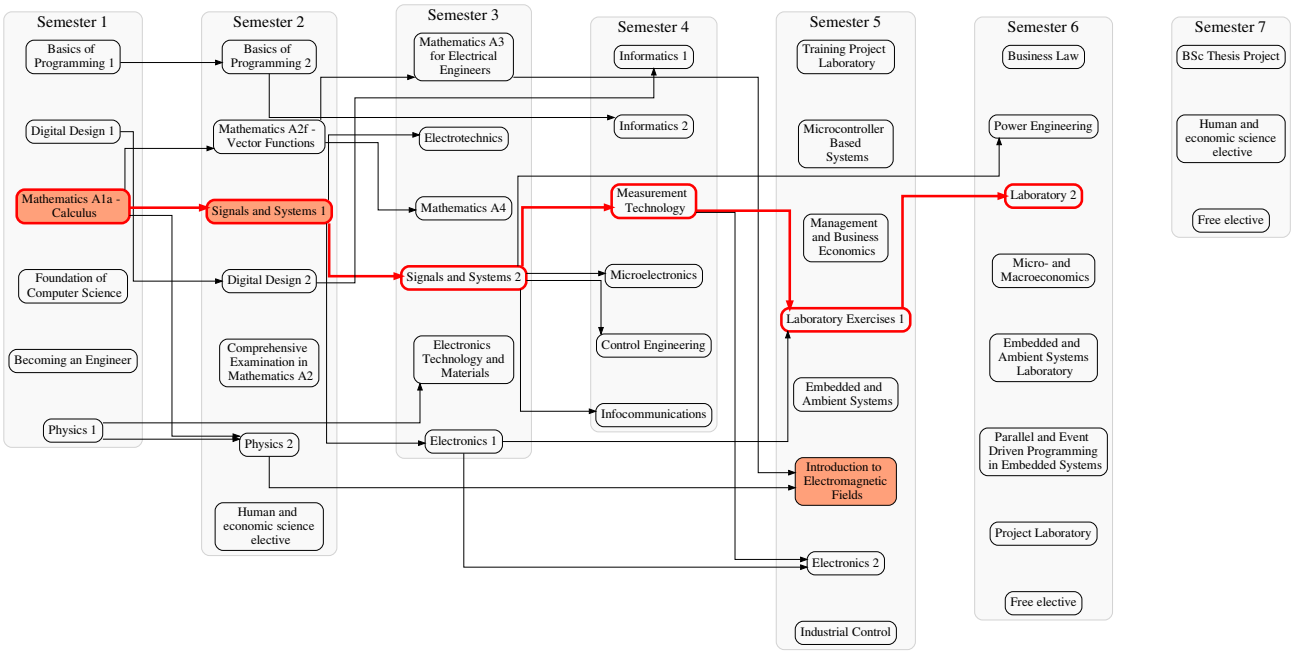

Fig. 4. Prerequisite network of Electrical Engineering Program (Embedded and Control Systems Specialization). The longest path is highlighted together with the courses having the highest betweenness, pass-through effect and blocking factor. For the exact values see Table III.

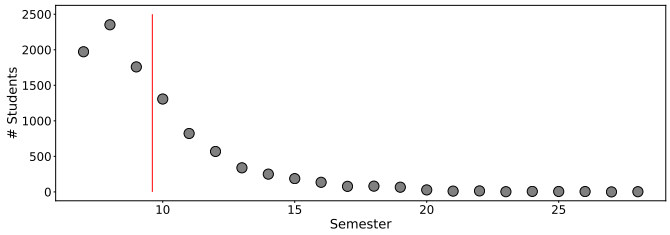

Fig. 5. Distribution of the graduation time at the EE program according to our model based on real course completion rates. The expected time of graduation is $\mu = 9.61$ (red line) and its standard deviation is $\sigma = 2.7$.

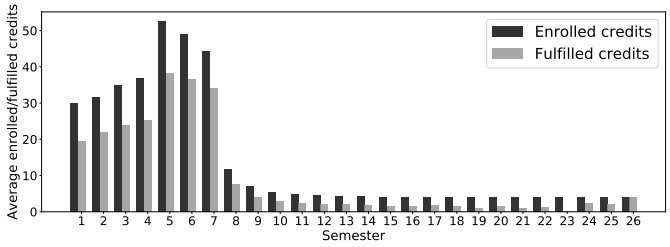

Fig. 6. Average enrolled and fulfilled credits per semester. The results are based on our simulation framework using the example of the EE program.

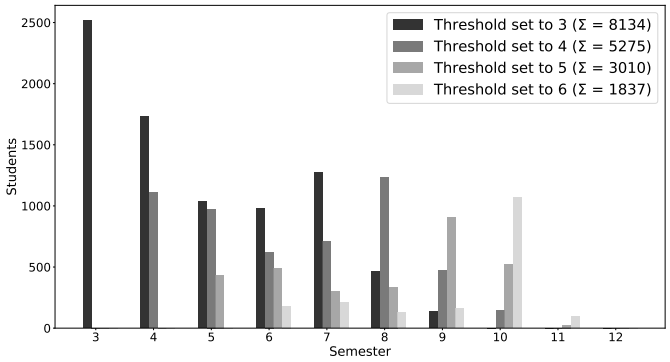

Fig. 7. Histograms of the number of dismissed students in each semester with different thresholds of maximum attempts to complete a course. The results are based on our simulation framework using the example of the EE program.

B. *Effect of launching courses in every semester*

Another important question regarding curriculum design is whether it is worth announcing a course every semester. Of course, this decision may depend on many factors, here we consider how the announcing frequency of the courses (i.e. launching in every semester as opposed to only once in a year) affects the graduation time. First, we consider how the distribution of the graduation time changes if every course is available only once in a year (autumn or spring term) instead of launching each course in every semester (see Fig. 8). Naturally, if the courses are announced only in every second semester then the mean graduation time significantly increases.

We also analyze what effect it has on the mean graduation time if a particular course is announced in every semester as opposed to every second semester. Introduction to Electromagnetic Fields has the greatest impact, not announcing the course in every semester but in every second semester, increases the mean graduation time by 13.84% according to our simulation framework.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2020.2981331, IEEE Transactions on Learning Technologies

TABLE III
CRUCIALITY METRICS FOR OBLIGATORY COURSES IN ELECTRIC ENGINEERING PROGRAM. THE PASS-THROUGH EFFECT IS APPROXIMATED BY $1 - d_i^{(3)}(1)$.

Course	Betweenness	Approximated pass-through effect	Deferment factor	Blocking factor
Digital Design 1	0	0.0034	0.2	2
Foundation of Computer Science	0	0.0034	0.143	0
Basics of Programming 1	0	0.019	0.2	2
Mathematics A1a - Calculus	0	0.0112	**0.5**	**17**
Becoming an Engineer	0	0	0.143	0
Physics 1	0	0.0123	0.2	3
Basics of Programming 2	0.00044	0.0045	0.2	1
Digital Design 2	0.00044	0.0045	0.2	1
Signals and Systems 1	**0.00443**	0.0134	**0.5**	11
Mathematics A2f - Vector Functions	0.00089	0.0011	0.25	3
Physics 2	0.00089	0.0101	0.2	1
Comprehensive Examination in Mathematics A2	0	0	0.167	0
Electrotechnics	0	0.0078	0.2	0
Signals and Systems 2	0.00399	0.0112	**0.5**	8
Mathematics A3 for Electrical Engineers	0.00044	0.0101	0.25	1
Mathematics A4	0	0.0045	0.2	0
Electronics 1	0.00266	0.078	0.333	3
Electronics Technology and Materials	0	0.0078	0.2	0
Informatics 1	0	0.0078	0.25	0
Informatics 2	0	0.0067	0.25	0
Measurement Technology	0.00133	0.0235	**0.5**	3
Microelectronics	0	0.0123	0.25	0
Control Engineering	0	0.0067	0.25	0
Infocommunications	0	0.0045	0.25	0
Training Project Laboratory	0	0.0034	0.333	0
Electronics 2	0	0	0.333	0
Microcontroller Based Systems	0	0.0067	0.333	0
Management and Business Economics	0	0.0089	0.333	0
Laboratory Exercises 1	0.00222	0	**0.5**	1
Embedded and Ambient Systems	0	0.0078	0.333	0
Introduction to Electromagnetic Fields	0	**0.0537**	0.333	0
Industrial Control	0	0.0022	0.333	0
Business Law	0	0.0045	**0.5**	0
Power Engineering	0	0.0179	**0.5**	0
Laboratory 2	0	0.0381	**0.5**	0
Micro- and Macroeconomics	0	0.0045	**0.5**	0
Embedded and Ambient Systems Laboratory	0	0.0123	**0.5**	0
Parallel and Event Driven Programming in Embedded Systems	0	0.0123	**0.5**	0
Project Laboratory	0	0.0012	**0.5**	0

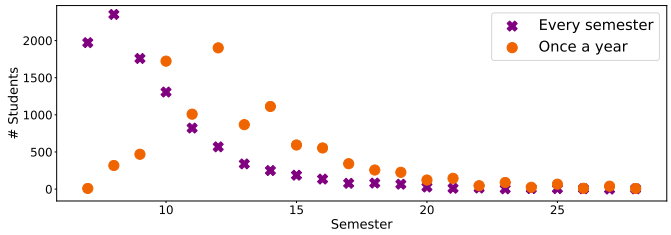

Fig. 8. The distribution of graduation time if every course is launched in every semester (purple) compared to the case when every course is launched only once in a year (orange). The results are based on our simulation framework using the example of the EE program.

VI. REFINING THE REPRESENTATIVE STUDENT MODEL

A drawback of the representative student model is that the course completion probabilities are fixed reflecting an average performance. On the other hand, in reality, the performance is highly student-dependent. To mitigate that, we modify the model, instead of one representative student we create three representative students corresponding to low-, medium- and high-performing students. The course completion rates are assigned to each group according to real historical data where the three groups are created according to the terciles along with the university entrance scores of the students. This is a reasonable choice since the university entrance score is a relatively good proxy of later academic performance [37]. The course completion rates in the three terciles can be seen in Fig. 9. We can observe that the performance difference among the groups are conspicuous in some courses (e.g. Physics 1, Basics of Programming 1), in other courses the difference is negligible (e.g. Micro- and macroeconomics, Laboratory 2).

Here we study the graduation time according to our model regarding the low-, medium- and high-performing representative student. Fig. 10 shows the distribution of graduation time by the three groups. It can be seen that the high-performing representative student is much more proficient than the medium- and the low-performing ones and low-performing students have the heaviest tail distribution.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2020.2981331, IEEE Transactions on Learning Technologies

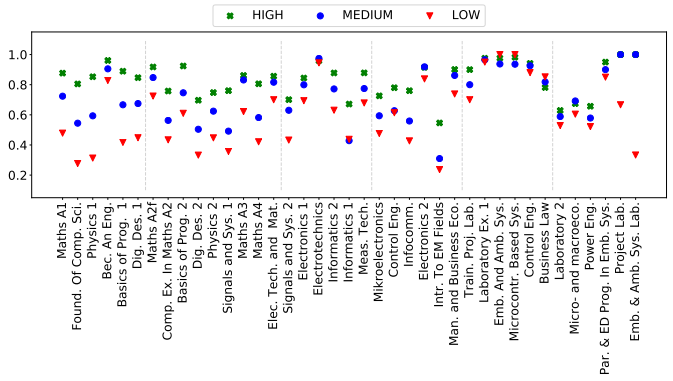

Fig. 9. Course completion rates for low-, medium- and high-performing students.

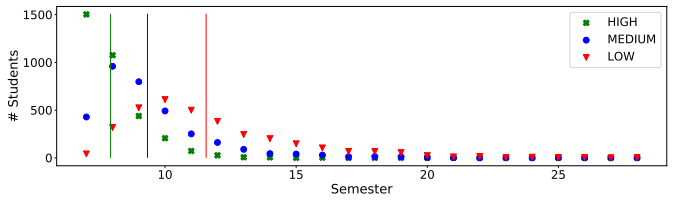

Fig. 10. Distribution of the graduation time for low-, medium- and high-performing students according to our model. The vertical lines represent the means of the distributions.

VII. A TOOL TO SUPPORT EDUCATIONAL STAKEHOLDERS

We developed a web application that can provide insights about curriculum prerequisite networks based on the approach presented in this paper. The software tool was developed in Python 3 using Dash web framework and Flask micro-framework.

Users can choose the curriculum (corresponding to the degree programs offered by our university) that they aim to analyze or users can also investigate their custom curriculum after uploading an Excel or CSV file containing the prerequisite structure in a matrix form together with the course completion ratios (if available). The application outputs the visual network representation of the prerequisite structure, calculates the topological indicators (e.g. deferment factor, betweenness centrality, blocking factor) and highlights the most critical courses.

Another function of the tool is that it can run simulations based on the student flow approach presented in this paper. Users can select the number of virtual representative students that the simulation is based on, the number of types of representative students (e.g. in Section VI this number is chosen to be 3 corresponding to low-, medium, and high-performing students). Users can also specify some policy rules such as the minimum number of credits that a student must complete in his/her last three active semesters, the maximum times that (s)he can attempt to complete a course, and whether or not a course is launched in every semester. The application returns a wide variety of outputs such as the approximated pass-through effect of the courses, the distribution of the graduation time, average enrolled and acquired credits per semester, the number of dismissed students in each semester.

Using the presented software, curriculum designers can find the answers to questions like how the structure of the prerequisite network together with the course completion rates affect the expected graduation time of students or what courses are the most critical concerning on-time graduation and long-term academic success. These are extremely important issues since delayed graduation and dropping out are serious problems all over the world, especially in STEM programs. Based on our framework, departments can not only identify the most critical ”bottleneck” courses but they can also simulate what-if scenarios mimicking either a change in the prerequisite structure or course completion ratios. Another advantage of our simulation model that the effect of various intended policy changes can also be evaluated before they are adopted, some examples of such policy changes are presented in more detail in Section V.

For assessing the utility and clarity of the application, we presented it to a body of university management including the rector, vice-rector, vice-deans, and other decision-makers. The overwhelming majority approved the tool and found it useful and easy to understand with high potential for further development as it answered the demands of the university to better design and support education, to understand the effect of curricular structure on graduation time and to evaluate curricular reform and policy changes.

While our framework was designed mainly for supporting curriculum designers and administrators, it can also help students. Using this tool, students can gain a better understanding of their time limits by which they need to complete certain courses to graduate on time, they can also realize what effect of failing certain courses have on their graduation time hence they can identify the courses that they should put more effort in. Moreover, our application also helps advisers provide more informed pieces of advice to students planning their semesters.

VIII. CONCLUSION AND FUTURE WORK

Analyzing university curricula is in high demand among policymakers and other stakeholders nowadays. In this paper, we presented a data-driven probabilistic student flow model to characterize prerequisite networks. We introduced a novel approach to characterize courses based on their effect on graduation time and illustrated the concept on the electrical engineering program of BME. We also developed a novel software tool based on our proposed approach and we demonstrated that our framework is suitable for evaluating curricular reform and policy changes, moreover it supports a wide range of stakeholders in education.

An interesting line of further research is to refine the model to account for the correlation between the completion of courses e.g. by revealing the Bayes structure of the prerequisite network since success/failure in a course clearly affect the success probability in follow-up courses. Another promising related future direction is to dynamically change the success probabilities for each student by a learning algorithm, based on their prior performance given by the model (instead of pre-categorizing them based on the university entrance score), that also takes into consideration how success or failure in a course

correlates with academic performance in other courses. If more data were available then the impact of student and instructor characteristics could be also taken into consideration.

ACKNOWLEDGEMENT

We are grateful for the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. We would like to thank Máté Baranyi and Marcell Nagy for their valuable insights and suggestions concerning this manuscript. We are also grateful for Bálint Csabay and István Bognár for their help in data collection.

The research reported in this paper was supported by the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Artificial Intelligence research area of Budapest University of Technology and Economics (BME FIKP-MI/SC). The publication is also supported by the EFOP-3.6.2-16-2017-00015 project entitled "Deepening the activities of HU-MATHS-IN, the Hungarian Service Network for Mathematics in Industry and Innovations". The research of R. Molontay was supported by NKFIH K123782 research grant and by the NKP-19-3-III New National Excellence Program of the Ministry for Innovation and Technology.

REFERENCES

[1] R. M. Saltzman and T. M. Roeder, "Simulating student flow through a college of business for policy and structural change analysis," *Journal of the Operational Research Society*, vol. 63, no. 4, pp. 511–523, 2012.

[2] G. L. Heileman, C. T. Abdallah, A. Slim, and M. Hickman, "Curricular analytics: A framework for quantifying the impact of curricular reforms and pedagogical innovations," *arXiv preprint arXiv:1811.09676*, 2018.

[3] E. P. Jansen, "The influence of the curriculum organization on study progress in higher education," *Higher Education*, vol. 47, no. 4, pp. 411–435, 2004.

[4] R. Robinson, "Pathways to completion: Patterns of progression through a university degree," *Higher Education*, vol. 47, no. 1, pp. 1–20, 2004.

[5] E. W. Bessent and A. M. Bessent, "Student flow in a university department: Results of a Markov analysis," *Interfaces*, vol. 10, no. 2, pp. 52–59, 1980.

[6] W. Plotnicki and R. Garfinkel, "Scheduling academic courses to maximize student flow: A simulation approach," *Socio-Economic Planning Sciences*, vol. 20, no. 4, pp. 193–199, 1986.

[7] B. M. Tallman and R. D. Newton, "A student flow model for projection of enrollment in a multi-campus university," 1973, Office of Budget and Planning, Pennsylvania State Univ.

[8] A. Bairagi and S. C. Kakaty, "A stochastic process approach to analyse students performance in higher education institutions," *International Journal of Statistics and Systems*, vol. 12, no. 2, pp. 323–342, 2017.

[9] C. Shah and G. Burke, "An undergraduate student flow model: Australian higher education," *Higher Education*, vol. 37, no. 4, pp. 359–375, 1999.

[10] A. Brezavšček, M. P. Bach, and A. Baggia, "Markov analysis of students performance and academic progress in higher education," *Organizacija*, vol. 50, no. 2, pp. 83–95, 2017.

[11] A. Fiallos and X. Ochoa, "Discrete event simulation for student flow in academic study periods," in *2017 12th IEEE Latin American Conference on Learning Technologies*, pp. 1–7.

[12] S. Mansmann and M. H. Scholl, "Decision support system for managing educational capacity utilization," *IEEE Transactions on Education*, vol. 50, no. 2, pp. 143–150, 2007.

[13] A. Schellekens, F. Paas, A. Verbraeck, and J. J. van Merriënboer, "Designing a flexible approach for higher professional education by means of simulation modelling," *Journal of the Operational Research Society*, vol. 61, no. 2, pp. 202–210, 2010.

[14] R. Saltzman, S. Liu, and T. Roeder, "Simulating student flow through a universitys general education curriculum," *Journal of Supply Chain and Operations Management*, vol. 17, no. 1, p. 14, 2019.

[15] A. C. Weber, "Simulating the flow of students through Cal Poly's undergraduate industrial engineering program for policy analysis," M.S. thesis, Ind. and Manuf. Eng. Dept., California Polytechnic State Univ., San Luis Obispo, California, USA, 2013.

[16] A. Slim, J. Kozlick, G. L. Heileman, and C. T. Abdallah, "The complexity of university curricula according to course cruciality," in *2014 8th IEEE International Conference on Complex, Intelligent and Software Intensive Systems*, pp. 242–248.

[17] J. Wigdahl, G. L. Heileman, A. Slim, and C. Abdallah, "Curricular efficiency: What role does it play in student success," in *Proceedings of the 121st ASEE Annual Conference and Exposition*, 2014, pp. 24.344.1–24.344.12.

[18] A. Slim, J. Kozlick, G. L. Heileman, J. Wigdahl, and C. T. Abdallah, "Network analysis of university courses," in *Proceedings of the 23rd ACM International Conference on World Wide Web*, 2014, pp. 713–718.

[19] A. Slim, "Curricular analytics in higher education," Ph.D. dissertation, Elect. and Comp. Eng., Univ. of New Mexico, Albuquerque, New Mexico, USA, 2016.

[20] A. Slim, G. L. Heileman, J. Kozlick, and C. T. Abdallah, "Employing markov networks on curriculum graphs to predict student performance," in *2014 13th IEEE International Conference on Machine Learning and Applications*, 2014, pp. 415–418.

[21] G. L. Heileman, M. Hickman, A. Slim, and C. T. Abdallah, "Characterizing the complexity of curricular patterns in engineering programs," in *ASEE Annual Conference & Exposition, Columbus, Ohio. https://peer. asee. org/28029*, 2017.

[22] M. İ. Akbaş, P. Basavaraj, and M. Georgiopoulos, "Curriculum GPS: an adaptive curriculum generation and planning system," in *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, 2015.

[23] T. Auvinen, J. Paavola, and J. Hartikainen, "STOPS: a graph-based study planning and curriculum development tool," in *Proceedings of the 14th ACM Koli Calling International Conference on Computing Education Research*, 2014, pp. 25–34.

[24] M. S. Hickman, "Development of a curriculum analysis and simulation library with applications in curricular analytics," Master's thesis, University of New Mexico, 2017.

[25] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning analytics*. Springer, 2014, pp. 61–75.

[26] R. C. Valle, S. Normandeau, and G. R. González, *Education at a glance interim report: update of employment and educational attainment indicators*. Organisation for Economic Co-operation and Development (OECD), 2015.

[27] J. Wigdahl, "Assesment of curriculum graphs with resppect to student flow and graduation rates," M.S. thesis, University of New Mexico, 2013.

[28] M. Van der Hulst and E. Jansen, "Effects of curriculum organisation on study progress in engineering studies," *Higher Education*, vol. 43, no. 4, pp. 489–506, 2002.

[29] G. Mendez, X. Ochoa, K. Chiluiza, and B. De Wever, "Curricular design analysis: a data-driven perspective," *Journal of Learning Analytics*, vol. 1, no. 3, pp. 84–119, 2014.

[30] D. M. Horváth, R. Molontay, and M. Szabó, "Visualizing student flows to track retention and graduation rates," in *2018 22nd IEEE International Conference Information Visualisation (IV)*, pp. 338–343.

[31] M. Raji, J. Duggan, B. DeCotes, J. Huang, and B. T. Vander Zanden, "Modeling and visualizing student flow," *IEEE Transactions on Big Data*, 2018, early access.

[32] J. M. Lightfoot, "A graph-theoretic approach to improved curriculum structure and assessment placement," *Communications of the IIMA*, vol. 10, no. 2, p. 5, 2010.

[33] P. R. Aldrich, "The curriculum prerequisite network: Modeling the curriculum as a complex system," *Biochemistry and Molecular Biology Education*, vol. 43, no. 3, pp. 168–180, 2015.

[34] B. Eisenberg, "On the expectation of the maximum of iid geometric random variables," *Statistics & Probability Letters*, vol. 78, no. 2, pp. 135–143, 2008.

[35] Budapest University of Technology and Economics, Faculty of Electrical Engineering and Informatics, "Bachelor of science degree program electrical engineering curriculum," 2019. [Online]. Available: https://www.vik.bme.hu/document/2432/original/BSC_EE_final.pdf

[36] Budapest University of Technology and Economics, "Rector's Order No. 7 (Code of Studies)," 2017.

[37] M. Nagy, R. Molontay, and B. Csabay, "Predictive power of admission point score and its variants on academic performance," presented at the 2nd Danube Conf. for Higher Education Management, Budapest, Hungary, Nov. 22-23, 2018.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2020.2981331, IEEE Transactions on Learning Technologies

Roland Molontay is a junior research fellow at the MTA-BME Stochastics Research Group operating at the Department of Stochastics at Budapest University of Technology and Economics. He earned his MSc degree in applied mathematics with highest honors. His PhD thesis focuses on network theory. He is also interested in data science and applied probability. He leads a small research group conducting data-driven educational research.

Noémi Horváth earned her MSc degree in applied mathematics from Budapest University of Technology and Economics. She is a software developer at SDA Informatika Zrt. developing education dashboards. She is also a research assistant at BME conducting data-driven educational research.

Júlia Bergmann earned her MSc degree in mathematics from Budapest University of Technology and Economics. She is a data analyst at the Institute of Computer Science and Control, Hungarian Academy of Sciences.

Dóra Szekrényes earned her MSc degree in applied mathematics from Budapest University of Technology and Economics. She works as a machine learning developer.

Mihály Szabó is an assistant professor of chemical engineering at Budapest University of Technology and Economics. He has gained expertise in higher education management and technologies by holding many administrative roles. He has been serving as the director of the Central Academic Office for more than 15 years.