Fall 12-17-2016

Curricular Analytics in Higher Education

Ahmad Slim 3589498

Ahmad Slim

Candidate

Engineering

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Gregory Heileman , Chairperson

Chaouki T. Abdallah

Terry Babbitt

Christos Christodoulou

Curricular Analytics in Higher Education

by

Ahmad Slim

B.E., Electrical Engineering, Lebanese American University, 2008
M.S., Computer Engineering, Lebanese American University, 2010

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctorate of Philosophy
Engineering

The University of New Mexico

Albuquerque, New Mexico

December, 2016

Dedication

To my parents

Acknowledgments

I would like to thank my advisor, Professor Gregory Heileman, for his support and guidance. His patience helped me to overcome many hard times during my PhD journey. It was a great opportunity to work under his supervision. He is a real life example of leadership.

I would also like to deeply thank Professor Chaouki Abdallah. All the credits go to him for facilitating the way for me to enroll in the PhD program at UNM—much appreciated!

I'm also grateful to Professor Samer Saab for his encouragement and advisement. He insisted that I continue my graduate studies. I never regret it.

Curricular Analytics in Higher Education

by

Ahmad Slim

B.E., Electrical Engineering, Lebanese American University, 2008
M.S., Computer Engineering, Lebanese American University, 2010

Ph.D., Engineering, University of New Mexico, 2016

Abstract

The dissertation addresses different aspects of student success in higher education. Numerous factors may impact a student's ability to succeed and ultimately graduate, including pre-university preparation, as well as the student support services provided by a university. However, even the best efforts to improve in these areas may fail if other institutional factors overwhelm their ability to facilitate student progress. This dissertation addresses this issue from the perspective of curriculum structure. The structural properties of individual curricula are studied, and the extent to which this structure impacts student progress is explored. The structure of curricula are studied using actual university data and analyzed by applying different data mining techniques, machine learning methods and graph theory. These techniques and methods provide a mathematical tool to quantify the complexity of a curriculum structure. The results presented in this work show that there is an inverse correlation between the complexity of a curriculum and the graduation rate of students attempting that curriculum. To make it more practical, this study was extended further to implement a number of predictive models that give colleges and universities the ability to track

the progress of their students in order to improve retention and graduation rates. These models accurately predict the performance of students in subsequent terms and accordingly could be used to provide early intervention alerts. The dissertation addresses another important aspect related to curricula. Specifically, how course enrollment sequences in a curriculum impact student progress. Thus, graduation rates could be improved by directing students to follow better course sequences. The novelty of the models presented in this dissertation is characterized in introducing graduation rate, for the first time in literature, from the perspective of curricular complexity. This provides the faculty and staff the ability to better advise students earlier in their academic careers.

Contents

List of Figures

List of Tables

Glossary

GPA	Grade Point Average is a standardized measure to average all grades from all current classes.
SAT	Scholastic Aptitude Test is a standardized test widely used for college admissions in the United States.
$CIRP$	Cooperative Institutional Research Program is a national longitudinal study of the American higher education system.
$STEM$	It is an acronym for Science, Technology, Engineering and Math education..
$CBCB$	Crucial Based Curriculum Balancing.
MSE	Mean Square Error.
BBN	Bayesian Belief Network.
MN	Markov Network.
EG	Expected Grade.
MAP	Maximum a Posteriori Probability .
SPM	Sequential Pattern Mining.
DAG	Directed Acyclic Graph .

Chapter 1

Introduction

1.1 Overview

Many definitions of student success exist in the literature. While these vary from grades and persistence to self-improvement, most studies consider graduation the ultimate measure of student success [59]. From the university's perspective, and especially for public universities, the definition of student success broadens from graduation into student retention rates and time-to-degree. These factors are important because many States have tied a percentage of the university's funding directly to such student success metrics [3]. This so-called "performance-based funding" has become a popular way to incentivize universities to help students graduate in a timely fashion. Whether a causal relationship exists between performance-based funding and graduation rates remains to be seen, but studies have clearly shown a rise in graduation rates as state appropriations per student increase [62].

From the state and federal levels, graduation rates are under increasing scrutiny [3]. This is driven by numerous factors, including the desire to improve institutional characteristics for rating purposes, the increasing trend of states tying institutional

funding to student outcomes through performance-based funding, as well as the fact that a bachelor's degree has become an increasingly necessary prerequisite for success in the work place creating a moral imperative for colleges and universities to graduate the students they admit. "If we want America to lead in the 21st century, nothing is more important than giving everyone the best education possible from the day they start preschool to the day they start their career," said President Barack Obama [4]. This is driven by the fact that higher educational attainment leads to healthier economic outcomes [1]. Thus earning a post-secondary degree is not considered a marginal achievement anymore or just an opportunity to fulfill personal ambitious; rather, it is a critical factor directly effecting the progress of the new economy [1]. The market that requires bachelor's degrees or higher is growing faster than those that do not; among the 30 fastest growing jobs, more than half require a bachelor's degree or higher. With the fact that the average salary of a university graduate is double that of a high school graduate, the middle class are seeking post-secondary degrees in ever increasing numbers.

Despite the value of a bachelor's degree, only 32.5% of the adult population in the United States has completed college [48]. Moreover, the degree completion of those students is widely disparate by race/ethnicity and gender. Only 22.5% of African American and 15.5% of Hispanics have a bachelor's degree compared to 36.2% of Whites. Given these pressures, universities are collecting unprecedented amounts of information related to student performance and progress, and applying ever more sophisticated analytical techniques in efforts to determine the most important factors that contribute to attrition and persistence [59,62]. Perhaps the most common guiding framework used by these universities to analyze factors contributing to student success is presented in Fig. 1.1.

These factors can be partitioned into three main paths: pre-institutional experiences, institutional conditions and student behaviors [35, 57]. The former include

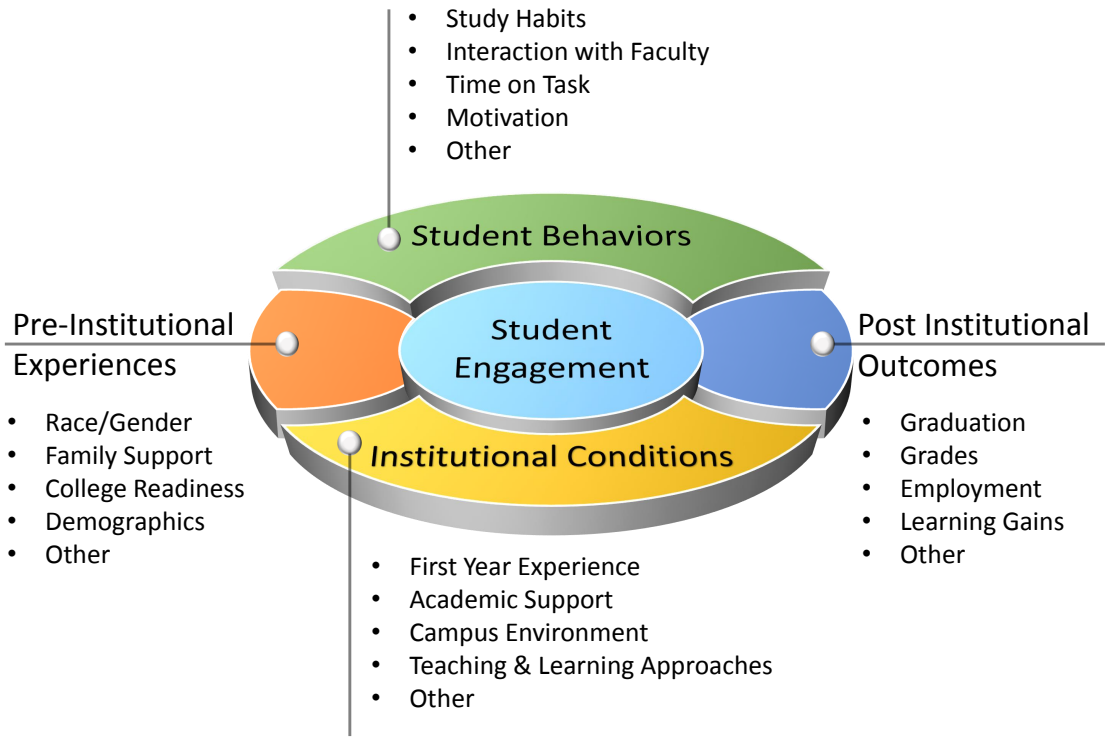


Figure 1.1: Student success framework.

such factors as pre-university preparation and socio-economic status, while the latter two include the interactions that take place while a student is enrolled at the university, these include institutional conditions and student behaviors. In this work we exclude any attempt to enhance pre-institutional experiences factors because they are typically beyond the direct control of the university. Our main contribution in this work is to present and discuss a novel institutional factor framework that can be easily employed by universities at a minimum cost and used to improve student outcomes.

A number of researchers have worked to identify the institutional conditions, e.g., the policies, programs, practices and cultural characteristics, that lead to stu-

dent success [35, 57]. They found that one of the most important factors is *student engagement*, which sits at the intersection of student behaviors and the aforementioned institutional conditions (Fig. 1.1). Furthermore, unlike most of the other factors that determine student success (e.g., previous preparation, socioeconomic status, etc.), student engagement is a factor that can be influenced by the institution. In efforts to improve student success, many institutions took these lessons to heart and worked to increase the amount and quality of the student support services they provide [33, 58]. For instance, many schools began to more rigorously and intentionally track the academic progress of their students, the extent to which they participate in educationally purposeful activities, the level of satisfaction with their campus experiences, and the added value (in terms of knowledge and skills acquired) of the entire undergraduate experience [40]. Some institutions reported significant increases in student success as a result of their efforts, but with others the benefits were much more limited.

The most fundamental measure of student success is *degree attainment*, and it is not uncommon to find accounts of students that earn a degree in spite of the fact that multiple indicators gave them little chance of success. They succeed in spite of the odds. For these students, indeed for any student, the simple facts are these: if they are able to successfully navigate all of the various requirements associated with a degree program, they earn the degree. Thus, at a very basic level it makes sense to think of all of the success-driven interventions mentioned above in terms of their ability to facilitate the movement of students through the individual requirements associated with degree programs. Indeed, the efficiency with which a student may progress through these requirements is what matters most in the end. Certainly, creating institutional conditions "that matter" will facilitate student progression, but there may also exist structural conditions within the curricula itself that limit progress independent of any success initiatives. Thus in this work we address student progress from the perspective of curriculum structure. This is an institutional con-

dition that is often overlooked. This dissertation presents a framework for analyzing student progress from this perspective.

First, this chapter will cover some of the most commonly discussed topics in the literature related to pre-institutional and institutional conditions contributing to student success. Then, Chapter 2 will present our proposed framework that addresses student progress at the most basic level, by investigating the structural properties of individual curricula. Chapter 3 extends this work by showing how to design curricula that reduce complexity by moving the courses with relatively higher "crucial values" to the earliest possible terms while meeting the prerequisite conditions and balancing the workloads of terms. We argue that this has a direct impact on student success and graduation rates. Chapters 4 and 5 introduce new applications for Bayesian Belief Networks (BBNs) and Markov Networks (MNs) to predict the performance of students early in their academic careers. These applications may prove useful in tracking the progress of students in order to provide early interventions aimed at improving student outcomes. In Chapter 6 we propose a model for analyzing university course enrollment networks at the program level. The analyses we provide are based on quantifying the importance of course enrollment sequences on a student's final grade point average (GPA), a metric that is highly correlated with graduation rates. In particular, we investigate the orderings of courses enrollment sequences that best contribute to student performance and achievement. In the last chapter we provide some concluding remarks and give some perspectives for further research.

The remainder of this chapter provides a snapshot of nationally effective pre-institutional and institutional conditions that are commonly cited and have shown a potential for increasing student success, retention, and graduation rates. Note that these factors and conditions will not be studied or analyzed further in this work as they have been extensively studied elsewhere. The main purpose of presenting them

here is to provide the "big picture" that will help formulate the problem that this dissertation addresses.

1.2 Pre-institutional Factors

This section provides a brief overview about the most common pre-instituational factors that were proven to have direct effect on graduation rates. In particular this section presents statistical results showing how characteristics, such as gender, ethnicity, academic background, and first-generation status, can influence graduation rates. The results presented here show the variation in correlation between these factors and graduation rates starting from 1967 until 2016. Thus this section is intended to give a literature review of the most studied factors that have clearly shown to be correlated with graduation rates. Although these factors will not be addressed further in the following chapters, they constitute a gateway that help better understand the model we are proposing in this dissertation.

1.2.1 Degree Attainment by Gender

Gender is a major factor that may be correlated to retention and graduation rates. Statistics show that, on average, women tend to graduate earlier than men. In the United States, for example, degree attainment for both genders has witnessed remarkable fluctuations throughout the years. Fig. 1.2 shows that men used to have higher college attainment compared to women up until 2014 [48]. From 2013 back to 1967 , the gap in degree attainment between men and women who are 25 years and older ranged between 1% and 8% with a peak in 1983. In 2013 the gap went down to 1% with degree attainment at approximately 30% for the two genders. In 2015, the picture changed. At that time 33% of women 25 years and older held a bachelor's

degree or higher compared to 32% of men. This increase in degree attainment is driven by the increased involvement of women in higher education. The year 1991 was a turning point in the history of US women aged 25 to 29. Starting from that year and up untill the present (2016), women have higher college attainment than men within the same age range (Fig. 1.3). Between 1967 and 1990, men aged 25 to 29 held more bachelor's degree compared to women with a peak of 27% in 1976. After that time period the percentage went slightly down and did not rise above 27% for 35 years. In 2012 degree attainment for men crossed 27% to reach 31% in 2015. However this is not the case with women aged 25 to 29. Fig. 1.3 shows that the growth in degree attainment for women is almost monotonically increasing. Between 1976 and 2011, the percentage of young men (25–29) with bachelor's degrees was 27% or below. However, the percentage of women (25–29) with bachelor's degrees went up from 20% to 36%. This indicates that women currently tend to graduate at a higher rate compared to men. This fact is reflected by a number of models in the literature that predict graduation rates where universities with higher women populations have higher graduation rates [15].

1.2.2 Degree Attainment by Race/Ethnicity

Race and ethnicity are also major predictive factors of retention and graduation rates. For example, in the United State, statistics show that diversity in race and ethnicity tends to significantly impact university outcomes. Fig. 1.4 shows the degree attainment variations among groups of different races. Asians recorded the highest degree completion percentage among all other groups in all years. For example in 1988, the percentage of Asians aged 25 years and older holding a bachelor's degree or higher is 38% compared to 21% of Whites, 11% of Blacks and 10% of Hispanics. Excluding the gap between Blacks and Hispanics, the degree completion gap among the rest of the groups remained almost the same over time. In 1988 the percentage

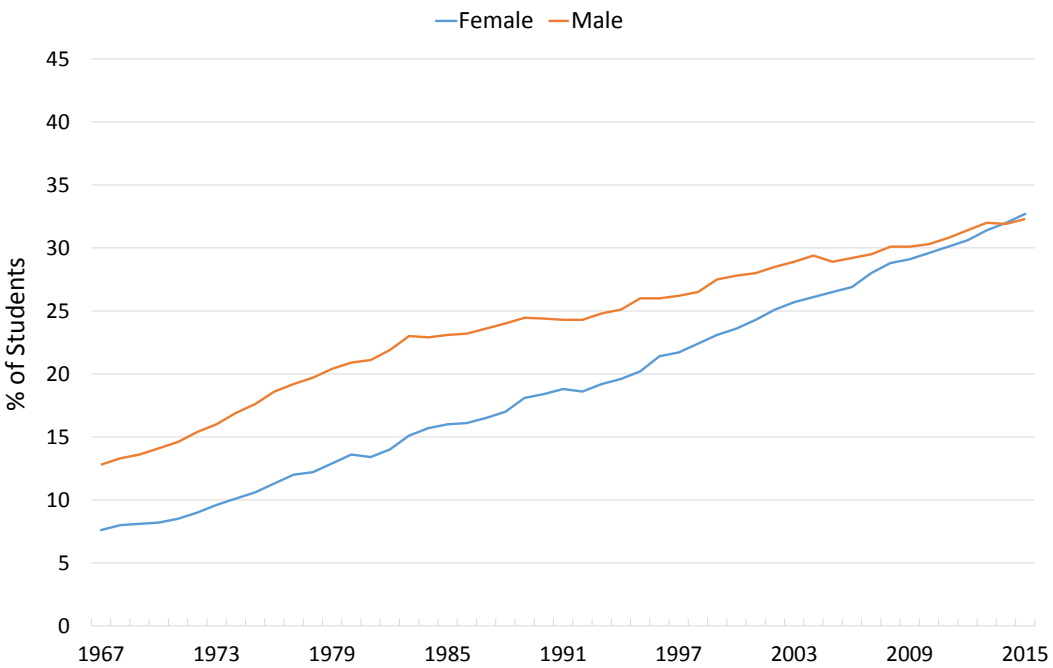


Figure 1.2: Percentage of the population 25 years and older with a bachelor's degree or higher by gender: 1967 to 2015

of degree attainment for both Blacks and Hispanics is around 11%. In 2015, however, the gap in degree completion between these two groups increased with Blacks reaching 22% compared to 15% for Hispanics. Fig. 1.4 shows an important fact. There is an increasing trend in degree attainment for all the races: Asians, Whites, Blacks and Hispanics starting with 38%, 21%, 11% and 10% completion rates in 1988, respectively, and reaching 54%, 36%, 22% and 15% in 2015.

1.2.3 Degree Attainment by Academic Background

Studies also show that pre-institutional academic backgrounds have direct influences on degree attainment [15]. Two of the most common measures used to examine

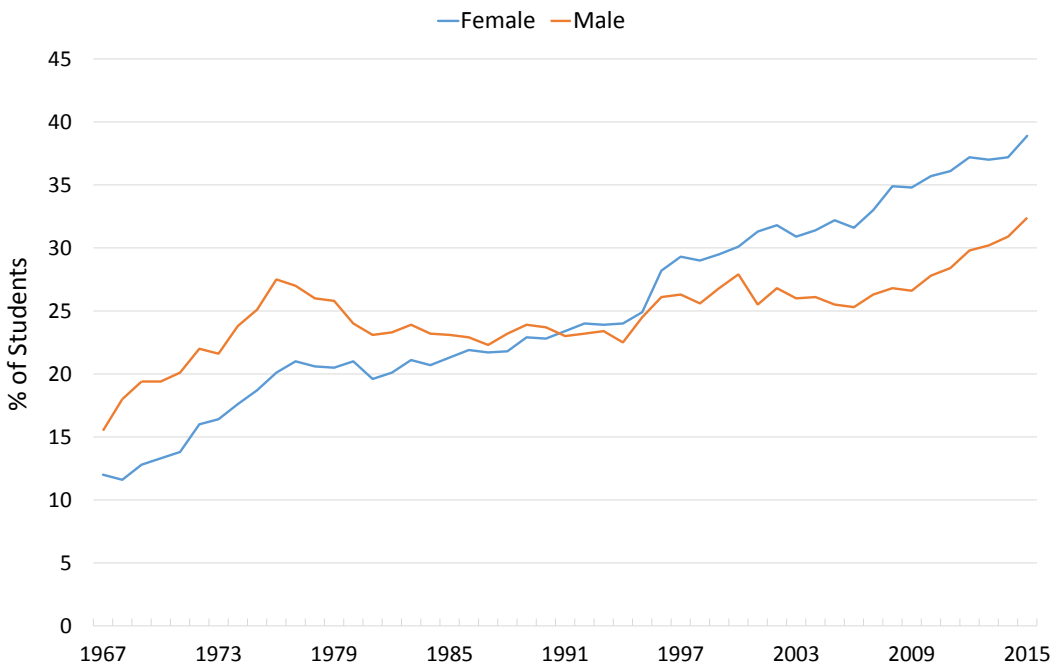


Figure 1.3: Percentage of the population aged 25 to 29 with a bachelor's or higher degree, by gender: 1967 to 2015

academic backgrounds are high school Grade Point Average (GPA) and Scholastic Aptitude Test (SAT) scores. Tables 1.1 and 1.2 show the results reported by the Cooperative Institutional Research Program (CIRP) Freshman Survey for the entering cohorts of 1994 and 2004, show this monotonically increasing relationship between degree attainment and high school GPA and SAT scores, respectively. In particular, Table 1.1 shows that students with higher high school GPAs graduate sooner than those with lower high school GPAs. For example, students with A/A+ high school GPA are twice as likely to graduate in four years compared to their B grade colleagues. Note that this gap decreases as students proceed in time. By the end of the sixth year the difference is approximately one third. The same applies for SAT scores. Table 1.2 shows that students with SAT scores of 1300 or higher have better

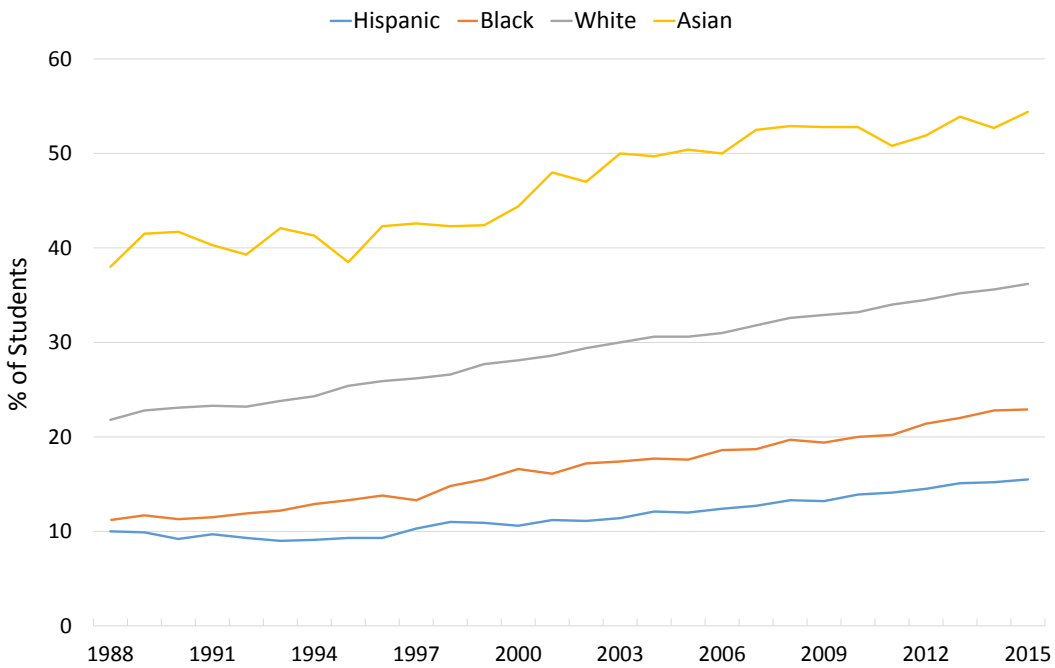


Figure 1.4: Percentage of the population 25 years and older with a bachelor's degree or higher by race: 1988 to 2015

four, five and six year graduation rate compared to those with lower scores. This result lines up with the rest of the SAT scores in Table 1.2.

	% of Students holding Bachelor's degrees Within		
HSGPA	4 Years	5 Years	6 Years
A/A+	58.2	75.6	79.3
A-	47.8	66.3	70.6
B+	35.9	54.7	59.8
B	25.2	43.3	48.7
B-	15.5	30.5	36.6
C+	9.8	22.4	27.7
C or less	6.3	16.0	21.2

Table 1.1: Four, five and six year graduation rates by high school GPA.

SAT score	**% of Students holding Bachelor's degrees Within**		
	4 Years	5 Years	6 Years
1300+	62.2	78.2	81.6
1200—1299	51.9	69.5	73.3
1100—1199	42.9	61.2	65.6
1000—1099	34.8	53.7	58.6
900—999	24.6	44.0	49.9
800-899	17.2	34.1	40.5
Less than 800	10.5	23.9	30.4

Table 1.2: Four, five and six year graduation rate by SAT score.

1.2.4 Degree attainment by First-Generation Status

Another factor that prove to affect degree attainment is the academic background of students' parents [15]. Fig. 1.5 shows that students whose parents attended college earn college degrees at a higher rate than those whose parents did not have higher education experience. The gap in degree attainment for these two groups of students remained almost the same for four, five, and six-year graduation rate with a difference of 14%.

1.3 Current Institutional Conditions

In this section we summarize the literature related to the most common institutional factors employed by universities to boost student outcomes. These include structural factors and academic factors. The former tries to insure a suitable campus environment for students by offering a combination of institutional physical features and students' demographic characteristics (this is explained in details in the following section). The latter tries to improve student outcomes by offering different academic support programs and teaching approaches. Fig. 1.1 shows a sample of these factors.

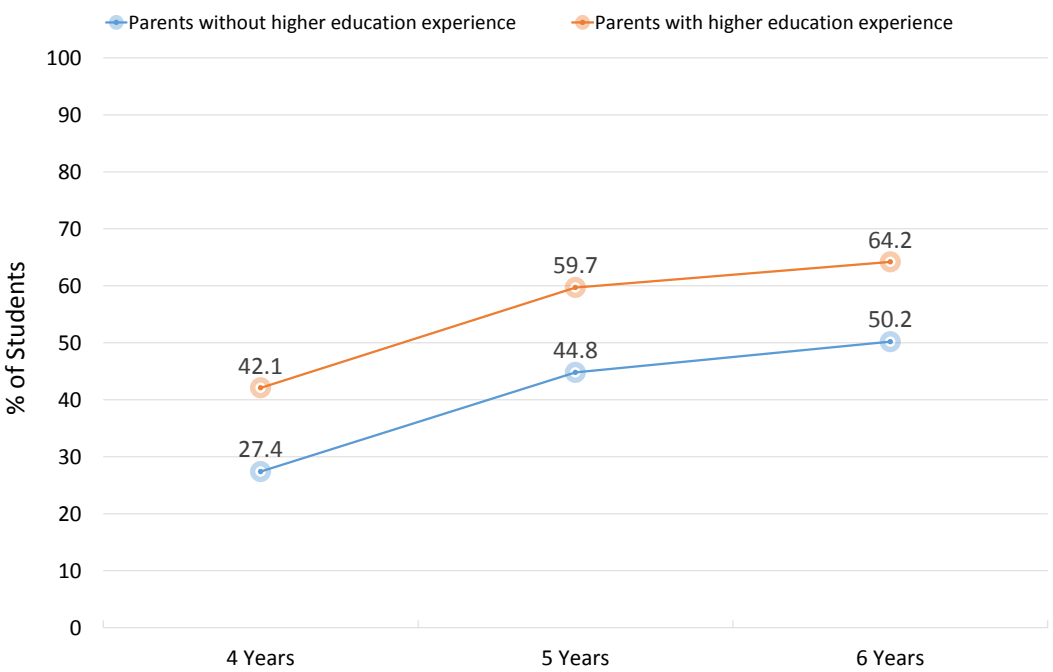


Figure 1.5: Four, five and six year graduation rate by parents' college experience [15].

1.3.1 Structural Factors

Structural factors include all non-academic institutional features related to the campus environment. These features range from the university's physical structure to student demographics. Features such as university size, architecture, design, buildings, residential character and student-faculty ratio have a direct impact on student outcomes by encouraging or discouraging the learning process [55]. The diversity of students in a university creates a supportive learning environment [2]. For example, students in liberal arts majors tend to have better diversity experiences because they simply are more likely to interact with students from different racial and ethnic backgrounds in their classes [36].

1.3.2 Academic Factors

As opposed to structural factors, academic factors include all the student services offered by the university that directly influence the academic performance of the students. This include the support programs services that facilitate a smooth traversal of the students through the requirements associated with their respective degree programs [31,34]. These services include advising, tutoring, seminars, remedial courses, intensive courses, study groups, etc. Another important academic aspect that has a crucial role in improving student outcomes is the learning approach or the pedagogical practice pursued by the faculty. This has been under extensive consideration and research for its direct influence on student performance [45]. The trend to increase academic standards, and hence student competence, is to switch from the traditional formal learning, which is a teacher-centered learning, to new informal learning methods that basically give the students the ability to acquire knowledge by observing and participating in social activities [44].

1.4 Curricula Structure and Graduation Rates

A number of data-driven tools have been developed for institutions to help predict graduation rates [5,15]. These tools mainly use methods such as traditional statistics, data mining and machine learning. A major factor that influences the accuracy of these predictive models is the choice of the independent variables. Some of these variables, such as ACT/SAT score and high school GPA, are very informative and hence they can be useful in predicting graduation rates; however, others might not be that informative. In the literature, most of these models use the pre-institutional and institutional conditions, discussed in previous section, as independent variables to predict graduation rates. Although the results of these models show a remarkable accuracy in predicting graduation rates, more work could be done in this area. That

is more informative variables should be integrated to these models. We claim that curriculum structure is one of these variables that may be used in order to better improve the accuracy of these models. The results shown in the following Chapters support this claim.

Chapter 2

Complexity Analysis of University Curricula

In Chapter 1 we provided a detailed background about the practices associated with improving graduation and retention rates in higher education. We discussed broadly the most efficient models and features documented in the literature to study and analyze the factors influencing student success metrics. In particular, we went over the pre-institutional and institutional conditions that have a direct impact on student success, retention, and graduation rates. However, none of these studies explore student progress from the perspective of curriculum structure. In this chapter we formulate a mathematical model that analyzes curricula structure and relate it to graduation rates. First, we determine the components of the curriculum that form the basis of our analysis. These are the courses and their respective dependency relationships. Then, using these components, we introduce two factors that measure the structural characteristics of the curriculum, we refer to these as blocking and delay factors. Using these two factors we define the complexity of the curriculum and accordingly, we study the correlation between the complexity of the curriculum and the graduation rate of students attempting that curriculum.

This chapter represents the core of this dissertation, along with one of the most important concepts we have derived, the *structural complexity* of a curriculum. Briefly defined, the *structural complexity* of a curriculum is determined by the manner in which the courses in the curriculum are arranged, e.g., prerequisites, number of courses, etc. On the other hand, the *Instructional complexity* of a curriculum is determined by the inherent difficulty of the courses in the curriculum, the quality of instruction, academic support, etc. These two components together define the complexity of a curriculum. In this dissertation, however, we focus our study only on the *structural complexity* of a curriculum and we analyze its impact on student progress.

Due to the nature of course interactions in curricula, we use graph theory and complex network analysis to provide a mathematical foundation for detecting crucial courses, which may help the university make decisions on when to offer certain classes, who should teach them, and what is truly necessary for a degree in a certain field. This work is important as it presents a robust framework to ensure the ease of flow of students through curricula with the goal of improving a university's graduation rates. Crucial courses have a high impact on student progress at universities and ultimately on graduation rates. Detecting such courses should therefore be a major focus of decision-makers at universities. The proposed cruciality measure is then further extended to study the complexity of curricula. In particular the cruciality measure is used to quantify the complexity of curricula and hence study the relation between curricular complexity and graduation rates.

2.1 Curriculum Graph

Using graph theory as the basic method to study curricular complexity, we build a model for the curriculum graph structure by abstracting the courses into nodes

and connecting two nodes with a directed edge if there is a pre-requisite relationship between the courses associated with the nodes. Fig. 2.1 is an example of the electrical engineering curriculum at the University of New Mexico (UNM)[1].

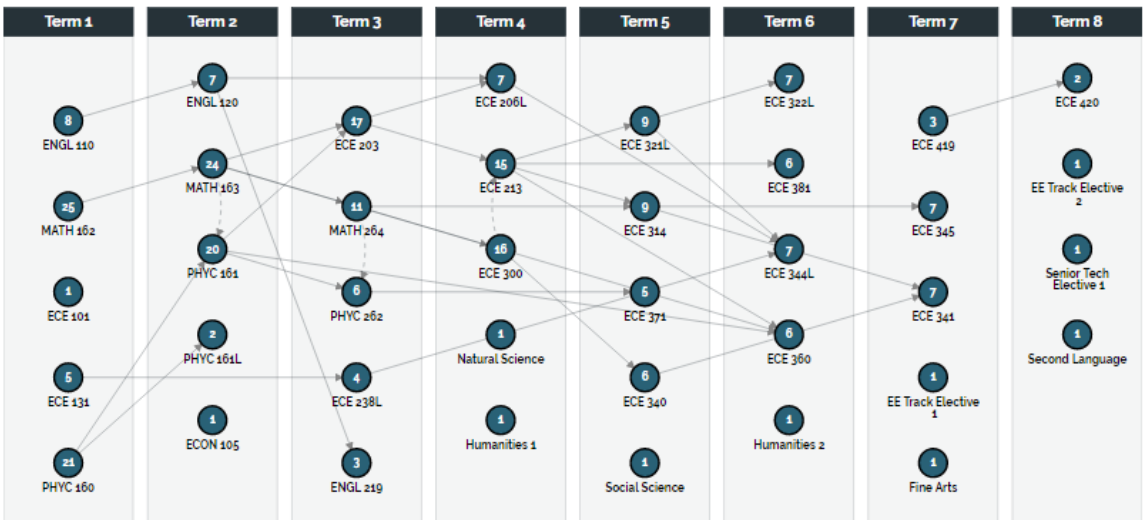


Figure 2.1: The electrical engineering curriculum at UNM.

By observing the graph structure of university curricula, we propose course cruciality as a major factor that impact students' ability to complete the curricula. Specifically, the cruciality of a course within a network is related to two main features, its delay factor and its blocking factor, and these two factors are characterized by two additional parameters, the longest path and the connectivity. The longest path L_i of node i is defined as the length of the longest path passing through node i. The connectivity, V_i of a node i is defined as the total number of nodes connected to i. That is, let n_{ij} be 1 if there is a path from i to j and 0 if no such a path exists. Then the connectivity V_i is given by

$$V_i = \sum_j n_{ij} \tag{2.1}$$

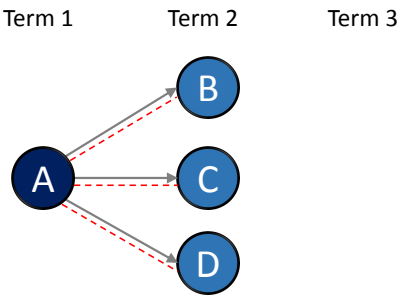
[1]https://curricula.academicdashboards.org

The following sections illustrate in details the significance of these parameters in quantifying the cruciality of courses and accordingly curricular complexity.

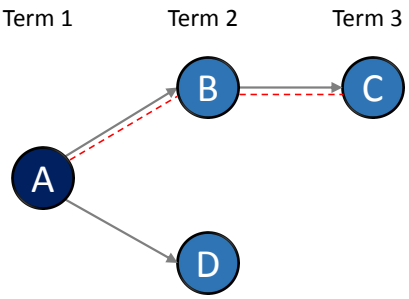
2.1.1 Delay Factor

Some courses have a critical impact on the academic progress of a student in the sense that any failure in these courses (or delays in taking them at the appropriate time) subjects the student to the risk of not finishing on time. It is therefore essential to detect these courses. The following example illustrates a process for detecting them using the longest path length parameter.

Given four nodes A, B, C and D representing four different courses, possible relationships between courses are shown in two different scenarios in Fig. 2.2. In Fig. 2.2(a) course A is the pre-requisite of B, C and D, while Fig. 2.2(b) shows the same courses, but with different prerequisite relationships between them. In the latter, A is the prerequisite of B and D whereas B is the prerequisite of C. Comparing these two figures, it is clear that A in Fig. 2.2(b) is more "crucial" than it is in Fig. 2.2(a). This may be explained as follows: assuming a three-term curriculum, a student who fails course A in Fig. 2.2(a) still have the chance of finishing on time, whereas one who fails course A in Fig. 2.2(b) ends up requiring more than three terms and is thus delayed. This phenomenon is reflected by the length of the longest path, L_A, shown by the red dashed lines. In Fig. 2.2(a), the longest path value of A is one whereas in Fig. 2.2(b) it is two. However, the value of the connectivity of A, V_A, is three in both scenarios.



(a) Node A has a longest path value, L_A, of 1



(b) Node A has a longest path value, L_A, of 2

Figure 2.2: The two graphs illustrate the cruciality of node A using using the longest path length factor.

2.1.2 Blocking Factor

In addition to the delay factor, it is natural to conclude that a course that is a prerequisite to a large number of other courses is more crucial. If a student fails such a course or does not attempt and pass it at the right time, the student may be blocked from attempting follow-on courses, leading to a negative impact on progress. This is illustrated by the following example. Nodes in Fig. 2.3 represent three different courses. In Fig. 2.3(a) the nodes are linked differently from those in Fig. 2.3(b). Node A in Fig. 2.3(a) is a prerequisite to node B whereas in Fig. 2.3(b) it is a prerequisite to nodes B and C. Comparing these two figures, it would be reasonable to consider node A in Fig. 2.3(b) more crucial than it is in Fig. 2.3(a). In the case of failure or delay, node A in Fig. 2.3(b) will block more courses. This result is reflected by the value of the connectivity, V_A, shown by the yellow dashed circles. In Fig. 2.3(a), the connectivity of A is one whereas in Fig. 2.3(b) it is two. However, the value of the longest path length for A, L_A, is one in both scenarios.

Based on the foregoing discussion, the cruciality of course i, denoted C_i, is defined as follows:

(a) Node A has a connectivity value, V_A, of 1

(b) Node A has a connectivity value, V_A, of 2

Figure 2.3: The two graphs illustrate the cruciality of node A using connectivity factor.

$$C_i = V_i + L_i \tag{2.2}$$

Note that course cruciality, C_i, may be defined using different forms of Eq. (2.2). For example, different weights may be assigned to V_i and L_i, that is $C_i = \alpha V_i + \beta L_i$ where α and β are constants. However, in the absence of training data that would better correlate these two factors to graduation rates, we assume $\alpha = \beta = 1$, that is blocking and delay factors are equally likely to influence graduation rate.

Note that other parameters such as in-degree and out-degree measures are not as suitable as the longest path and connectivity parameters. For example, if we consider the in-degree and out-degree parameters instead of the longest path length parameter to compute the crucaility of node C in Fig. 2.2(a) and Fig. 2.2(b), both scenarios would lead to the same crucaility value which does not differentiate between the two scenarios despite the fact that node C in Fig. 2.2(b) is more crucial than it is in Fig. 2.2(a) taking into consideration the delay factor discussed previously.

2.1.3 Curricular Complexity

Accordingly, we define the complexity, S, of a curriculum as the sum of the crucialities of all courses in the curriculum:

$$S = \sum_{i}^{n} C_i \tag{2.3}$$

where n is the number of courses in a curriculum

To better illustrate the definition of curricular complexity, consider the curriculum shown in Fig. 2.4(a). In this example, the curriculum complexity is $8+7+3+5+4 = 27$ which is simply the sum of the crucialities of all courses in the curriculum. On the other hand, the cruciality value of each course in this curriculum is the summation of its respective delay factor (Fig. 2.4(b)) and blocking factor (Fig. 2.4(c)).

In the following section we analyze the influence of curricular complexity on graduation rates. In particular we show how these two variables are correlated. This can be exploited to improve the accuracy of models predicting graduation rates. In other words, curricular complexity together with other pre-institutional factors, such as gender, ethnicity, ACT/SAT scores, and first-generation status, would constitute independent variables for the models (i.e., regression, support vector machines, Bayesian networks, etc.) that predict graduation rates. We claim that variables, such as curricular complexity, improve the accuracy of such predictive models. This summarizes, to an extent, one of our main contributions in this dissertation. Our claim is supported by a number of simulations shown in the following section.

2.2 Simulation

In this section we present the results for a number of Monte Carlo simulations [32]. These simulations show empirically the type of correlation between the complexity of a curriculum and the graduation rate of students attempting that curriculum. In particular we design a number of curricula made up of four courses each (Fig. 2.5). Each curriculum has a complexity value representing its structural layout. Then we run a Monte Carlo simulation of students flowing through each of these curricula and accordingly compute the graduation rate. The results for each curriculum are shown in Tables 2.1, 2.2, 2.3, and 2.4.

Course	**Term**			
	1	**2**	**3**	**4**
1	51.2%	75.5%	87.5%	93.8%
2	49.8%	74.7%	87.5%	93.7%
3	50.2%	75.4%	87.7%	94.1%
4	0	44.3%	71.6%	85.7%
Grad. rate	0	20.6%	49.3%	71.3%

Table 2.1: Simulated graduation rate for curriculum 2.5(a).

Course	**Term**			
	1	**2**	**3**	**4**
1	49.5%	75.0%	87.8%	94.1%
2	49.9%	74.8%	87.2%	9347%
3	0	24.0%	49.4%	69.0%
4	49.9%	74.6%	87.4%	93.6%
Grad. rate	0	13.4%	37.9%	60.4%

Table 2.2: Simulated graduation rate for curriculum 2.5(b).

Table 2.1 shows the simulated graduation rates for the curriculum shown in Fig. 2.5(a). The layout of this curriculum has no prerequisite relationships. Thus the total complexity value -using Eq. (2.3)-sums up to a relatively low number of four. The

Course	**Term**			
	1	**2**	**3**	**4**
1	50.0%	75.2%	87.2%	93.5%
2	49.5%	74.7%	87.6%	93.7%
3	0	12.3%	34.0%	54.7%
4	50.2%	75.5%	87.6%	93.7%
Grad. rate	0	9.3%	29.7%	51.2%

Table 2.3: Simulated graduation rate for curriculum 2.5(c).

Course	**Term**			
	1	**2**	**3**	**4**
1	50.1%	74.9%	87.3%	93.6%
2	49.7%	75.2%	87.8%	93.7%
3	0	25.1%	49.8%	68.2%
4	0	25.2%	50.2%	69.1%
Grad. rate	0	6.4%	25.2%	47.3%

Table 2.4: Simulated graduation rate for curriculum 2.5(d).

graduation rate for the students attempting this curriculum after 4 terms is 71.3%. This result is relatively higher than that shown in Table 2.2 for the curriculum shown in Fig. 2.5(b). This curriculum (Fig. 2.5(b)) has only one prerequisite relationship going from course A to course C. This one prerequisite relationship increases the complexity value to seven. In return, the graduation rate after four terms decreases to 60.4%. Thus there is a drop of 11% in the graduation rate after adding only one prerequisite relationship. This simple simulation reveals the type of correlation between curricular complexity and graduation rates. As we increase curricular complexity the graduation rates decrease. The following additional simulations confirm this claim: the curriculum shown in Fig. 2.5(c) has two prerequisite relationships going from course A to course C and from course B to course C. These two prerequisites increase the complexity of the curriculum to nine. This increase in complexity imposes an additional drop of 9% (Table 2.3) in the graduation rate compared to the

curriculum shown in Fig. 2.5(b). This inverse correlation continues to show up in the simulation done for the curriculum shown in Fig. 2.5(d). This curriculum, similar to the curriculum shown in Fig. 2.5(c), has two prerequisite relationships. However the layout structure is different. The difference in the structure is reflected a difference in the complexity value. The curriculum of Fig. 2.5(d) is more complex than that of Fig. 2.5(c) with a complexity value of ten. This increase in complexity adds an additional 4% drop in the graduation rate (Table 2.4).

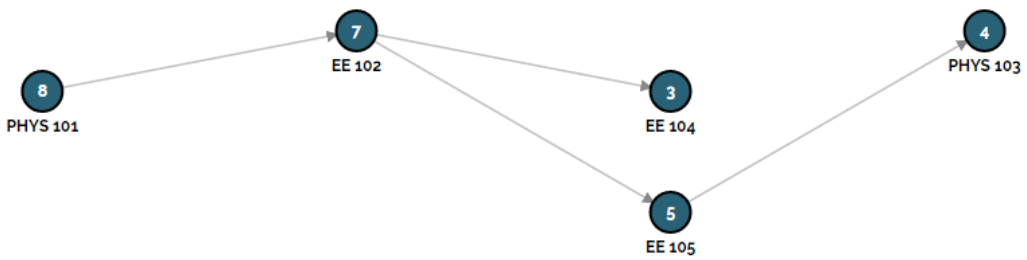
The following section presents a real life scenario showing the improvement in the graduation rate after modifying the structural layout of a common curricular pattern in the school of engineering.

Fig. 2.6(a) shows a common curricular pattern in electrical, computer and mechanical engineering. Recently a number of universities (i.e., Wright State University (WSU), UNM, etc.) have investigated changes in the structure of this curricular pattern. They realized that the current pattern imposes unnecessary complexity to students attempting it. Students must complete a total of six courses before they can take *circuits I* with the longest chain being four courses. If a student fails to pass one of these courses, they are delayed a whole term. Thus any effort to reduce the complexity of this pattern would facilitate the traversal of students through its degree requirements. In return, this would indeed improve the graduation rate. Thus these universities managed to design a less complex layout for this pattern and at the same time achieve the same learning outcomes. The new proposed curricular pattern is shown in Fig. 2.6(b). *Precalc* is replaced with an *Engineering* 101 course that prepares students for *Calc I* and *Circuits I*. The number of courses remains the same, but the curriculum is less complex.

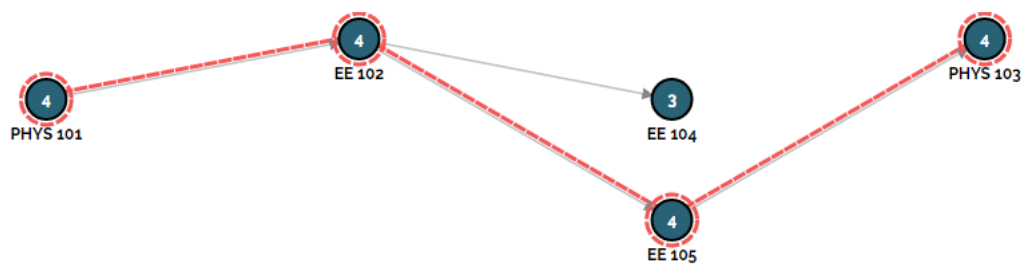
In this section we used our proposed model in order to quantify the changes in the complexity value after modifying the original curricular pattern (Fig. 2.6(a)). We also showed the difference in the graduation rates imposed by this modification

using the Monte Carlo simulation.

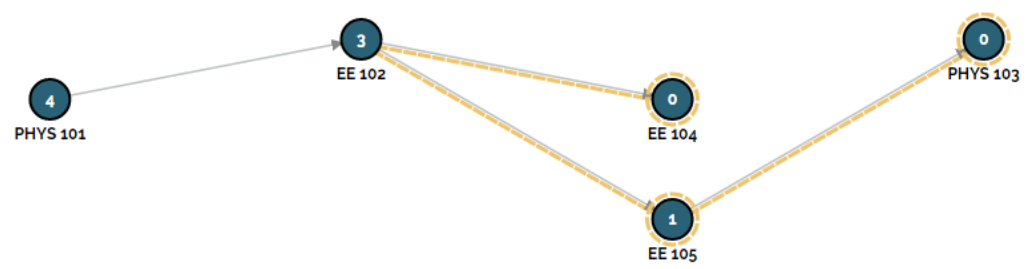
We computed the complexity of both patterns using Eq. (2.3). It is clear that the complexity value dropped significantly after modifying the original curricular pattern. The complexity went down from 56 to 42. The drop in complexity is reflected as an increase in graduation rate. The Monte Carlo simulation shows that the graduation rate after 7 terms went up from 72% to 89%. The results shown here support the claim of the universities modifying their curricula. It is obvious that these kinds of efforts would positively influence the educational sector at least at the level of graduation rates.



(a) Curriculum Complexity.

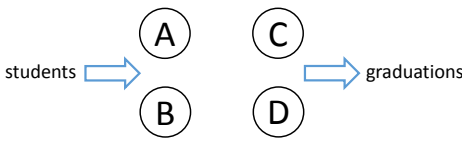


(b) Delay factor.

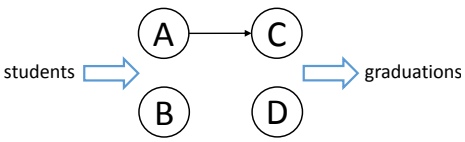


(c) Blocking factor.

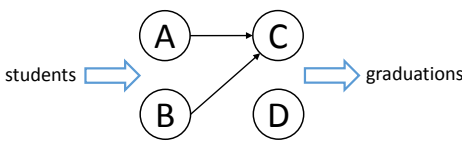
Figure 2.4: (a) In this example, the curricular complexity is $8 + 7 + 3 + 5 + 4 = 27$. (b) *EE 102* has a delay factor of 4. This can be seen by the dashed line connecting *PHYS 101, EE 102, EE 105*, and *PHYS 103*. (c) *EE 102* has a blocking factor of 3. This can be seen by the dashed line connecting *EE 102* to the three other courses *EE 104, EE 105*, and *PHYS 103*.



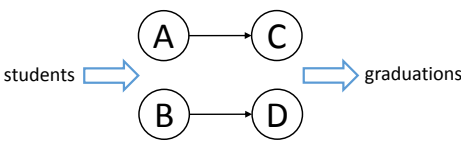
(a) This curriculum has a complexity value of four. There exists no prerequisite relationships in this layout.



(b) This curriculum has a complexity value of seven. There exists one prerequisite relationships in this layout going from course A to course C.

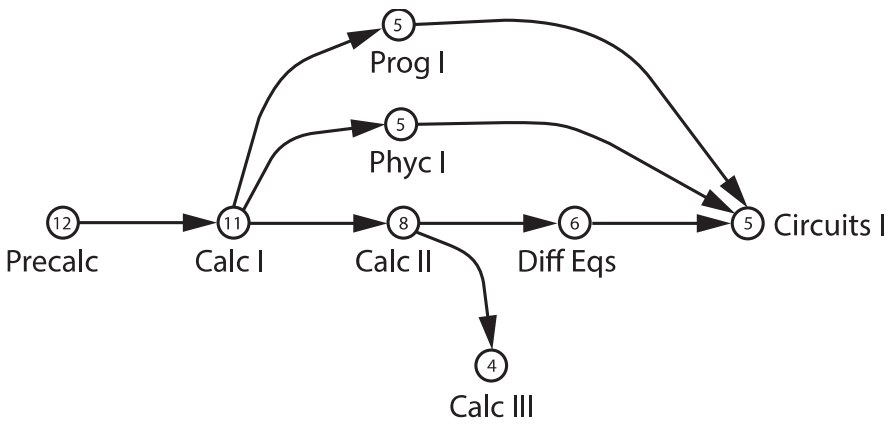


(c) This curriculum has a complexity value of nine. There exists two prerequisite relationships in this layout going from course A to course C and from course B to course C.

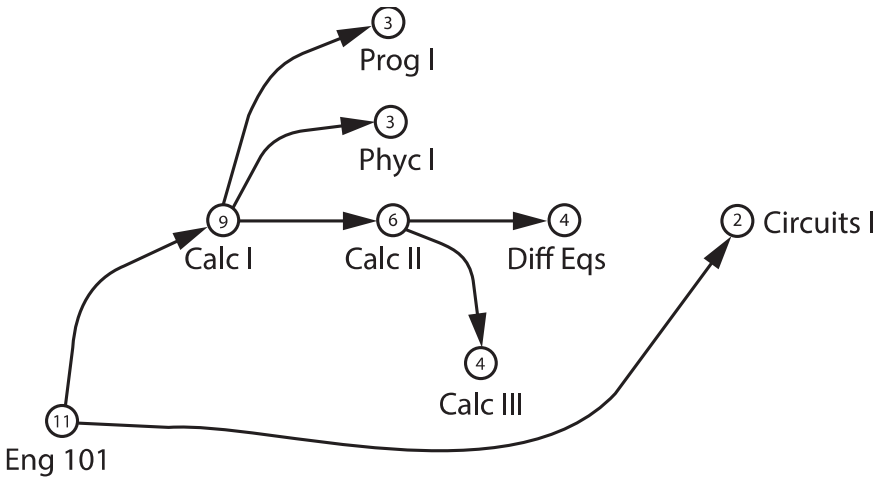


(d) This curriculum has a complexity value of ten. There exists two prerequisite relationships in this layout going from course A to course C and from course B to course D.

Figure 2.5: Four course curricula.



(a) Original curricular pattern.



(b) Modified curricular pattern.

Figure 2.6: A common curricular pattern in electrical, computer and mechanical engineering.

Chapter 3

Cruciality-Based Curriculum Balancing

In the previous chapter we defined the cruciality of a course in a curriculum based on its respective blocking and delay factors. A course is crucial in the sense that any failure or delay in taking it at the appropriate time subjects the student to the risk of not finishing on time. So it would be essential to move these courses to the earliest possible terms while meeting all the constraints related to prerequisite relationships, maximum and minimum amount of academic load per term, maximum and minimum number of credit hours per term, etc. In this chapter we introduce a new optimization model called Cruciality-Based Curriculum Balancing (CBCB) that achieves this goal using Integer Linear Programing (ILP) and Constraint-Based (CB) techniques. The novelty of this model is characterized by its ability to outperform other models by utilizing a number of objective functions in a single framework.

3.1 Introduction

The Balanced Academic Curriculum Problem (BACP) aims to schedule all courses within a curriculum to specific terms while satisfying the prerequisite dependency relationships and maintaining a balanced workload across all terms [11, 39]. The motivation for this work was mainly to reduce the total weekly lecture hours for a student [13, 49]. Accordingly, a large number of variants of the BACP have been proposed in the literature in an attempt to improve the performance and solution quality.

In 2001 and 2002, constraint and integer programming techniques were used to solve different BACP models [11, 24]. In 2006, a hybrid technique utilizing genetic algorithms and constraint programming was developed to solve the BACP [37]. In 2008, a new parameter related to the lecturer preferences was added to the BACP extending it to the Generalized BACP (GBACP) model [19]. In 2012 an integer programming model was introduced for the GBACP based on hybrid local search techniques [13].

In the previous studies, BACP was formulated to assign the courses to terms while meeting prerequisite conditions [39]. But there was no special precaution for assigning a specific course and its prerequisite as close as possible. For instance, the prerequisite of a course in the seventh term may be assigned to the first, second or third terms. But of course, it would be much better to locate the prerequisite course just before its latter course (i.e. the 6th term in this case). To achieve this goal, curriculum balancing was modeled as a Generalized Quadratic Assignment Problem (GQAP), which is a totally new approach for curriculum design [42]. This work developed a model called the Relevance Based Curriculum Balancing (RBCB) that assigns relevant courses to closest possible terms while meeting all the constraints of BACP. However, designating the pair to the"term 5–6" instead of "term 6–7" would

still be another major improvement, considering the impact this approach imposes on student success and hence graduation rates [51, 52, 60]. The RBCB model and the rest of the above mentioned studies did not take into account implementing this improvement in their works.

In this chapter, we design a curriculum that will better fit real life situations by not only minimizing the distance between relevant courses but also moving them to the earliest possible terms while meeting all the constraints of BACP (Fig. 3.1). To achieve this goal, we propose CBCB as a multi-objective optimization problem using linear objective functions which is another advantage over the proposed RBCB model implemented using a non-linear function—nonlinear optimization problems are considered to be harder than linear problems [25].

3.2 Problem definition

According to Castro and Manzano, the BACP should encapsulate a number of regulations and constraints [11]. These constraints define the limits of the optimization problem we are solving. For example, an *academic curriculum* is defined as a set of courses and a set of precedence relationships among them. An *academic curriculum* should have a specified *number of terms*. Each term requires a *minimum* and a *maximum number of courses*. This is required in order to consider students as full time and in order to avoid overload. Each course, in returns, is associated with a number of credit hours that represent the academic effort required to successfully follow it. Table 3.1 presents detailed definitions for the regulations and constraints of the BACP.

However, in real life, balancing the academic workload per term and satisfying

Academic Curriculum	A set of courses and a set of precedence relationships among them.
Number of terms	Courses must be assigned within a maximum number of academic terms.
Academic load	Each course has a number of credit hours.
Prerequisites	Some courses can have other courses as prerequisites.
Minimum academic load	A minimum amount of credits per term is required to consider a student as full time.
Maximum academic load	A maximum amount of credits per term is allowed in order to avoid overload.
Minimum number of courses	A minimum number of courses per term is required to consider a student as full time.
Maximum number of courses	A maximum number of courses per term is allowed in order to avoid overload.

Table 3.1: The regulations and constraints of BACP.

prerequisite conditions are not the only criteria for curriculum design. The proposed model in this chapter considers criteria distinct from other models in literature by moving courses with relatively higher crucial values to the earliest possible terms (Fig. 3.1). This summarizes the main objective of this chapter.

3.3 Lexicographic Optimization

Different researchers have defined the term "solving a multi-objective optimization problem" in various ways. Therefore, in the literature multiple methods were proposed to address this problem. Many of these methods try to convert the original problem with multiple objectives into a single-objective optimization problem. This is called a scalarized problem. The lexicographic technique is one of these methods which will be used in this chapter to solve our proposed CBCB model.

With the lexicographic method, the objective functions are arranged in order of importance [38]. Then, the following optimization problems are solved one at a time:

$$\min_{x \in X} F_i(x) \tag{3.1}$$

$$\text{subject to } F_j(x) \leq F_j(x_j^*),$$

$$j = 1, 2, \ldots, i-1, i > 1; i = 1, 2, \ldots, k.$$

Here, i represents a function's position in the preferred sequence, and $F_j(x_j^*)$ represents the optimum of the jth objective function, found in the jth iteration. After the first iteration ($j = 1$), $F_j(x_j^*)$ is not necessarily the same as the independent minimum of $F_j(x)$, because new constraints have been introduced. The constraints in (3.1) are sometimes replaced with equalities [53].

3.4 Integer Linear Programming (ILP) Model

An integer programming problem is a mathematical optimization program in which some or all of the variables are restricted to be integers. In many settings the term refers to ILP, in which the objective function and the constraints are linear. In this section, we present an ILP model for the CBCB.

3.4.1 Parameters

In the previous section we defined an academic curriculum as a set of m courses related with a set of prerequisite relationships. We also defined the work load of course i by the total number of credit hours α_i. We then assigned the curriculum a specified number of terms n. Each term, in returns, is assigned a minimum number of courses δ and a minimum number of credit hours β. This is essential in order to consider students full time. On the other side, each term is assigned a maximum number of courses ϵ and a maximum number of credit hours γ. This is essential to avoid overload. The main objective is to move course i with cruciality c_i to the earliest possible term. This could minimize the risk for students of not finishing on time. Table 3.2 shows the parameters of the ILP model in more details.

This section defines the variables we are optimizing. Basically we are moving the highly crucial courses to the earliest possible terms while maintaining a balanced

m	Number of courses
n	Number of academic terms
α_i	Number of credits of course i; $\forall\ i = 1 \ldots m$
c_i	Cruciality of course i; $\forall i = 1 \ldots m$
β	Minimum academic load allowed per term
γ	Maximum academic load allowed per term
δ	Minimum amount of courses per term
ϵ	Maximum amount of courses per term

Table 3.2: The parameters of the ILP model.

curriculum. This may be achieved in two steps. In the first step we lay out a balanced curriculum by applying the constraints of the BACP (i.e., maximum/minimum load, prerequisite dependency, etc.). This is done by minimizing the academic load l defined as following:

$$l = \max\{l_1 \ldots l_n\}$$

where l_j is the academic load of term j defined as:

$$l_j = \sum_{i=1}^{m} \alpha_i * x_{ij}; \forall j = 1 \ldots n$$

where

$$x_{ij} = \begin{cases} 1 & \text{if course } i \text{ is assigned to term } j \\ 0 & \text{otherwise} \end{cases}$$

This step will give us more than one layout since the constraints of the BACP may be achieved in many different ways.

In the second step we simply move high crucial courses to the earliest terms while maintaining the same value of the academic load l obtained in the first step. This could be done by minimizing the total weighted summation of courses' cruciality C defined as:

$$C = \sum_{j=1}^{n} \sum_{i=1}^{m} j * c_i * x_{ij}; \forall j = 1 \ldots n$$

Again, this step will give us more than layout because many courses might have the same cruciality values. Thus switching these courses will not change the value of C. Note that minimizing the value of C is achieved with smaller values of j. Thus minimizing C will guarantee that the courses with relatively higher crucial values are assigned to earliest possible terms.

It was previously mentioned that the BACP is formulated using a number of constraints. These constraints restrict the maximum and the minimum number of courses in a term, the maximum and the minimum number of credit hours in a term, and they assign a set of prerequisite dependency among these courses as well. Mathematically, these constraints are defined as following:

- All courses i must be assigned to some term j:

$$\sum_{j=1}^{n} x_{ij} = 1; \forall i = 1 \ldots m$$

- Course b has course a as prerequisite:

$$\sum_{j=1}^{n} j * x_{aj} < \sum_{k=1}^{n} k * x_{bj}$$

- The academic load of term j must be greater than or equal to the minimum required:

$$l_j \geq \beta; \forall j = 1 \ldots n$$

- The academic load of term j must be less than or equal to the maximum allowed:

$$l_j \leq \gamma; \forall j = 1 \ldots n$$

- The number of courses of term j must be greater than or equal to the minimum allowed:

$$\sum_{i=1}^{m} x_{ij} \geq \delta; \forall j = 1 \ldots n$$

- The number of courses of term j must be less than or equal to the maximum allowed:

$$\sum_{i=1}^{m} x_{ij} \leq \epsilon; \forall j = 1 \ldots n$$

3.5 Constraint-Based (CB) Model

Constraint Programming deals with optimization problems using the same basic idea of verifying the satisfiability of a set of constraints. Asuming one is dealing with a minimization problem, the idea is to use an upper bound that represents the best possible solution obtained so far. Then we solve a sequence of constraint satisfaction problems (CSPs) each one giving a better solution with respect to the optimization function [11]. In this section, we present our CB model for the CBCB.

The decision variables are the same ones defined in the ILP model; however the total weighted summation of courses' cruciality C that we are minimizing is defined differently:

$$C = \sum_{i=1}^{m} P_i * c_i$$

where P_i is the term number of course i; $\forall\ i = 1 \ldots \mathrm{m}$ The constraints in this case are similarly the maximum and the minimum number of courses in a term, the maximum and the minimum number of credit hours in a term, and they assign a set of prerequisite dependency among these courses as well. Mathematically, these constraints are defined as following:

- The academic load of term j is defined by:

$$l_j = \sum_{i=1:P_i=j}^{m} \alpha_i * x_{ij}; \forall j = 1 \ldots n$$

ILP model	
Advantage	*Disadvantage*
All constraints are linear	Difficulty to state the prerequisite constraint
Ease of statement of the academic load constraint	

CB model	
Advantage	*Disadvantage*
Ease of statement of the prerequisite constraint	Inefficient statement of the academic load constraint
Use of global constraint (better propagation)	

Table 3.3: Advantages and disadvantages of using the ILP model over that using the CB model.

- Course b has course a as prerequisite:

  $$P_a < P_b$$

- The academic load of term j must be greater than or equal to the minimum required:

  $$l_j \geq \beta; \forall j = 1 \ldots n$$

- The academic load of term j must be less than or equal to the maximum allowed:

  $$l_j \leq \gamma; \forall j = 1 \ldots n$$

- Global constraints to restrict the number of courses for each term

  $$atleast(j, P, \delta) \text{ and } atmost(j, P, \gamma); \forall j = 1 \ldots n$$

The main advantages and disadvantages of using the ILP and the CB models are presented in Table 3.3.

3.6 Experimental Results

Our intention in this work is not to evaluate the performance of the proposed integer and constraint programming. The literature already contains a large number of methods each having a different performance measure which is not the scope of this work. The main goal of this chapter is to present a framework that extends the previous models by adding the course cruciality criterion. This is the novelty of the model that makes it different from the other models in literature. On the other hand, the new contribution achieved in this work is the ability to model the CBCB problem using *linear* objective functions which is another improvement compared to that of the RBCB model implemented using non-linear functions.

In order to empirically validate our proposed CBCB model, we created a simple five-term curriculum (Fig.3.4) with actual university[1] requirements along with their respective prerequisite relationships. In the curriculum, there can be several technical elective (TE) and general elective (GE) courses. However, BACP models in literature do not consider TEs and GEs. TEs are generally selected from a specific list of courses, while GEs can be selected among all courses that are neither in the "Compulsory Courses" nor "Technical Elective Courses" lists. Since students are free to select the elective courses among a number of alternatives, different combination of elective courses will generate different curriculum plans. However, for most of the cases, suggesting different type of curricula is not practically feasible; therefore departments offer standard curricula to their students.

For the elective courses, the proposed model has the following assumptions:

- **Assumption 1:** since the contents of GEs are quite different than the compul-

[1] http://degrees.unm.edu/undergrad_programs/by_college/22

sory courses and TEs, it is assumed that the prerequisite conditions between the GEs and all other courses are zero.

- **Assumption 2:** TEs are generally interrelated with one another and the compulsory courses. To be able to offer a standard curriculum, mostly selected TEs are used.

The prerequisite relationships and the cruciality values for all the courses within the curriculum presented in Fig.3.4 are given in Table 3.4 and Table 3.5 respectively.

Fig. 3.4(b) shows how courses with relatively higher crucial values are assigned to the earliest possible terms compared to Fig. 3.4(a). Using the CBCB model, courses with relatively higher cruciality values, such as *ENGL 110, MATH 162, MATH 163, MATH 314, MATH 316* and others, are moved one term closer compared to the BACP model. As mentioned previously, our proposed model is a twofold goal: it moves relevant courses to the earliest possible terms and it minimizes the distance between them. For example, the distance between *ENGL 110* $\rightarrow$ *ENGL 120, MATH 162* $\rightarrow$ *MATH 163* and *PHYC 160* $\rightarrow$ *PHYC 161* is one term. However, using the BACP models, the distance may be greater.

3.7 Student Progress

As mentioned in previous sections, time-to-degree is a critical factor in the academic life of both students and universities. Students normally want to obtain their degrees as soon as possible (subject to financial and work-life constraints) while universities want their graduation rate to be as high as possible. Usually grades (e.g., GPA) are the main criteria to measure the student progress throughout a curriculum. Grades

do not however take the time factor into consideration. Theoretically, a student may have a high GPA while progressing slowly through the curriculum. Engineering student A who takes crucial courses in the first semester earning a GPA of 4.0 is therefore in better shape than student B who takes non-crucial courses while earning the same GPA of 4.0. Obviously, the probability that student B may be delayed in a program is higher than that of student A based on the definition of crucial courses. Hence and based on the time factor mentioned, crucial courses must be included in studying the progress of students through out their respective academic life.

3.7.1 Framework

To achieve this, we propose a framework that makes use of the CBCB model and the earned letter grade. We thus create an "efficient" curriculum (using the CBCB framework) for every department within the university and accordingly monitor a student's progress every semester based upon the type of the courses (i.e., crucial or noncrucial) taken, along with respective letter grades. Students having more courses matching the cruciality values of the "efficient" curriculum courses per term are in a better shape assuming all students have the same GPA. Fig. 3.2 shows a three-term "efficient" curriculum. Next to each course there are two numbers. The number on top represents the cruciality value whereas the one below represents the earned letter grade. Note that in this work the highest grade value of a course is 4.0. Assuming students X and Y have the same letter grades for all the courses as shown in Fig. 3.2, student X is less likely to get delayed throughout her academic program. Numerically, this may be quantified by summing the product of both cruciality value and letter grade of all the courses taken up to that term, that is:

$$P_j = \frac{\sum_{ij} c'_{ij} G_i}{\sum_{in} c'_{in}} \tag{3.2}$$

where P_j is the student progress score (SPS) at term j, c'_{ij} is the cruciality value of course i taken in term j, G_i is the letter grade of course i and n is the total number of terms in a curriculum. Note that a course must match one of the "efficient" curriculum courses, otherwise the cruciality value is zero, that is

$$c'_{ij} = \max\{c_i, 0\} \tag{3.3}$$

The denominator in Eq. (3.2) normalizes the SPS so that P_j is always less than or equal to 4.0. In fact the value of the SPS in the last semester is equivalent to the GPA value. However, the advantage of the SPS over GPA is its ability to quantify student progress, taking into consideration the time factor mentioned previously. Its cruciality is even more evident in the first couple of semesters where students at this time need more advisement than at other points in their academic careers. Examples shown in the next section illustrate how to analyze the SPS.

As a first step, the student progress framework discussed above is to an extent idealistic. Normally, curricula are not as simple as it appears in Fig. 3.2. For example, degree requirements for many curricula are technical elective courses, social science courses, humanity courses, etc. It may therefore be hard to create one "efficient" curriculum for a particular program. Some technical elective or social science courses might have different cruciality values. Thus the SPS would not reflect the true progress value unless some further assumptions are made. First it should be clear that there must be one curriculum for each program. Accordingly, this means there must be one reference to which students can refer to and hence student progress framework would be feasible then to apply. To achieve this, it is assumed that all degree requirements that are unspecified within a curriculum (e.g. technical elective courses and social science courses) do not have pre-requisites. This will provide a curriculum with a minimum bound above which the SPS wouldn't exceed. So all the

unnamed courses taken by a student in which they match her respective department curriculum are assumed to have no pre-requisites and thus the cruciality values for such courses are 1. Note that the cruciality values of the courses in each term in the "efficient" curriculum are computed excluding the pre-requisite edges emerging from courses of previous terms. Hence, for example, the cruciality values of the courses D, E and F in Fig. 3.2 are one instead of two. Thus, once students pass courses A, B and C, it makes no difference if they take D, E and F before G, H and I or vice-versa the next term. The example shown in Fig. 3.3 illustrates the main idea of the student progress framework. The optimal SPS P_1^o, P_2^o and P_3^o are $\frac{48}{17}$, $\frac{60}{17}$ and $\frac{68}{17}$ respectively.

3.7.2 Student Progress Ratio

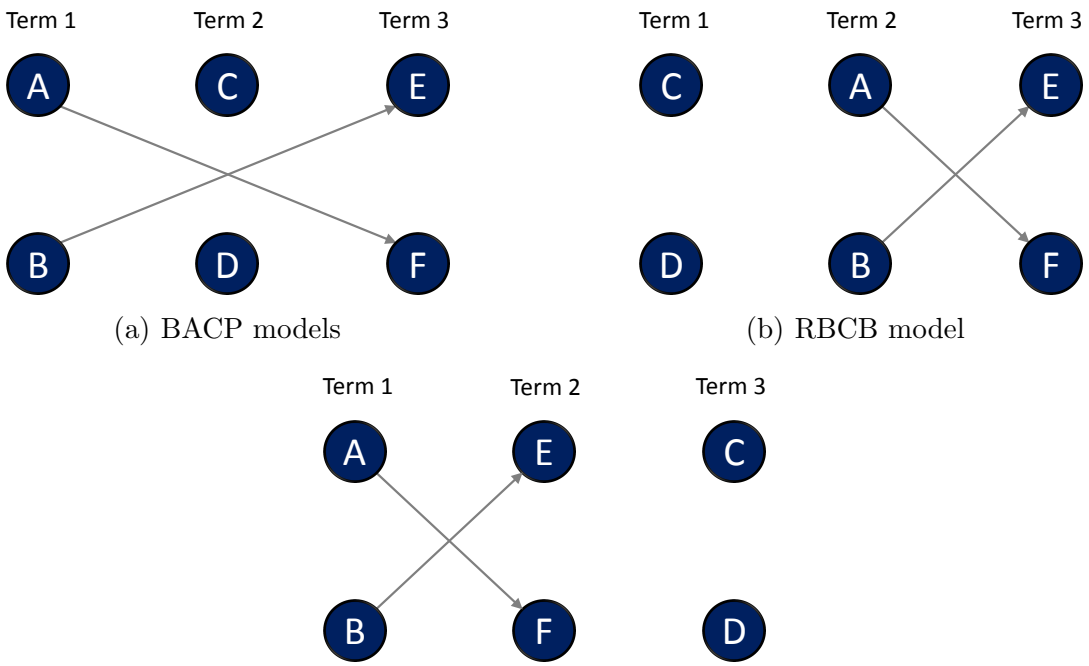
Analyzing student progress is achieved by considering the ratio of P_j^s from the student curriculum over that of P_j^o from the "efficient" one each term, that is:

$$I_j = \frac{P_j^s}{P_j^o} \tag{3.4}$$

where I_j is the student progress ratio (SPR) at term j. If the value of the SPR is greater than or equal to 1, then the student is on track. Otherwise, special attention must be taken depending on how far below 1 the SPR is. For example, in Fig. 3.3, $I_1 = \frac{24}{48}$ which is 0.5. This might be a sign that student X is at risk of being delayed, because she did not take crucial courses in her second term or because she earned bad grades. Note that SPR must be less than or equal to 1 in the last term meaning that the student has finished all the curriculum requirements.

Course name	prerequisite	Course name
ENGL 110	→	ENGL 120
ENGL 110	→	ENGL 219
MATH 162	→	MATH 163
MATH 162	→	CHEM 121
ECE 131	→	ECE 203
PHYC 160	→	PHYC 161
ENGL 120	→	ECE 206L
MATH 163	→	ECE 203
MATH 163	→	MATH 264
MATH 163	→	MATH 314
MATH 163	→	MATH 316
PHYC 161	→	PHYC 262
ECE 203	→	ECE 206L
ECE 203	→	ECE 213
MATH 316	→	ECE 213
Technical Elective 1	→	Technical Elective 2

Table 3.4: The prerequisite relationships for all the courses within the curriculum.



(a) BACP models

(b) RBCB model

(c) CBCB model

Figure 3.1: This figure shows a three-term curriculum. The courses in the curriculum are scheduled using three different models: BACP, RBCB and CBCB. Using the BACP model, the distance between relevant courses (A—F; B—E) is not optimal or close enough. The RBCB model overcomes this limitation by implementing a non-linear framework that minimizes the distance between these relevant courses. However these courses are not assigned to the closest terms (i.e., Term 1). The CBCB model overcomes the limitations in BACP and RBCB models by using a linear framework which minimizes the distance between relevant courses and assigns them to the closest possible terms.

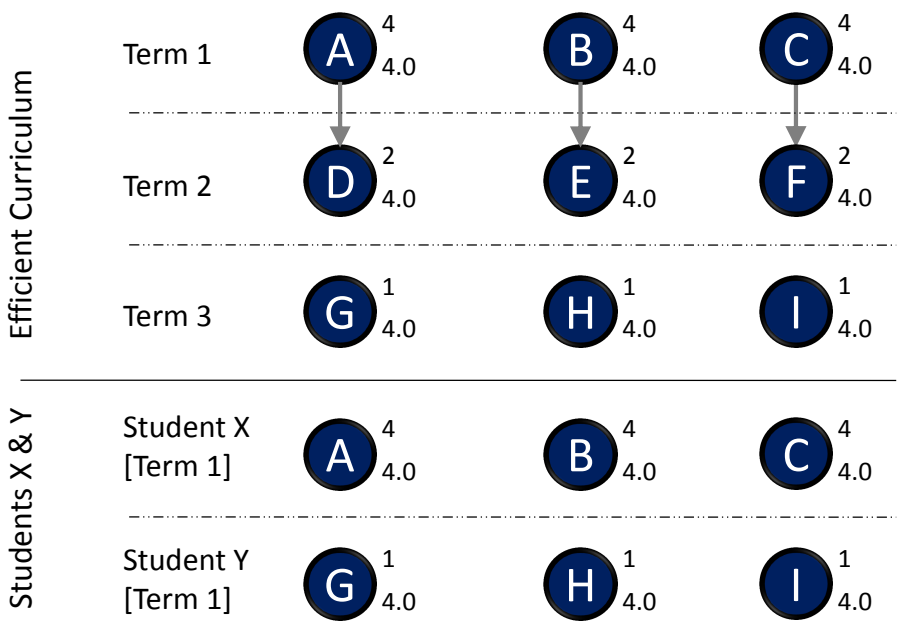


Figure 3.2: Progress of students X and Y with respect to the "efficient" curriculum. SPS of X is $\frac{48}{18}$ whereas that of Y is $\frac{12}{18}$.

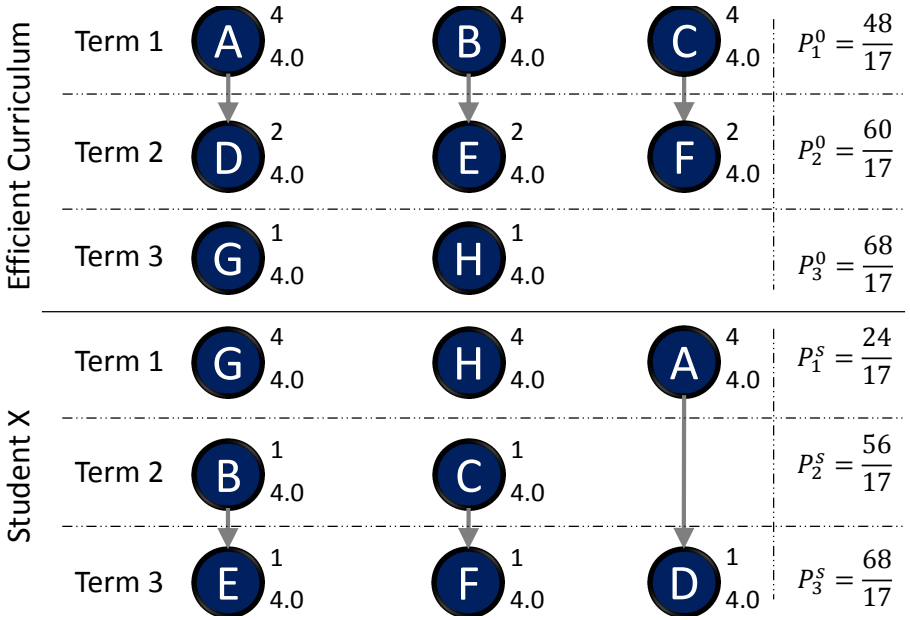


Figure 3.3: SPR of student X. $I_1 = \frac{24}{48}$; $I_2 = \frac{56}{60}$; $I_3 = \frac{68}{68}$.

Course no	Course name	Cruciality value
01	ENGL 110: Accelerated Composition	5
02	MATH 162: Calculus I	11
03	ECE 131: Introduction to Programming	5
04	PHYC 160: General Physics I	4
05	ENGL 120: Composition III	3
06	MATH 163: Calculus II	10
07	PHYC 161: General Physics II	3
08	CHEM 121: General Chemistry I	1
09	ECE 203: Circuit Analysis I	5
10	MATH 264: Calculus III	3
11	MATH 316: Differential Equations	4
12	PHYC 262: General Physics III	2
13	ECE 213: Circuit Analysis II	3
14	MATH 314: Linear Algebra	2
15	ECE 206L: EE Lab I	3
16	ENGL 219: Technical Writing	1
17	Humanities	0
18	Technical Elective 1	2
19	Technical Elective 2	1
20	Social Science	0

Table 3.5: The cruciality values for all the courses within the curriculum.

i \ j	Term 1	Term 2	Term 3	Term 4	Term 5
ENGL 110	☐	☑	☐	☐	☐
ENGL 120	☐	☐	☑	☐	☐
ENGL 219	☐	☐	☐	☐	☑
MATH 162	☐	☑	☐	☐	☐
MATH 163	☐	☐	☑	☐	☐
CHEM 121	☐	☐	☑	☐	☐
ECE 131	☑	☐	☐	☐	☐
ECE 203	☐	☐	☐	☑	☐
PHYC 160	☑	☐	☐	☐	☐
PHYC 161	☐	☑	☐	☐	☐
ECE 206L	☐	☐	☐	☐	☑
MATH 264	☐	☐	☐	☑	☐
MATH 314	☐	☐	☐	☑	☐
MATH 316	☐	☐	☐	☑	☐
PHYC 262	☐	☐	☑	☐	☐
ECE 213	☐	☐	☐	☐	☑
Technical Elective 1	☑	☐	☐	☐	☐
Technical Elective 2	☐	☑	☐	☐	☐
Humanities	☐	☐	☐	☐	☑
Social Behavior	☑	☐	☐	☐	☐

(a) BACP model

i \ j	Term 1	Term 2	Term 3	Term 4	Term 5
ENGL 110	☑	☐	☐	☐	☐
MATH 162	☑	☐	☐	☐	☐
ECE 131	☑	☐	☐	☐	☐
PHYC 160	☑	☐	☐	☐	☐
ENGL 120	☐	☑	☐	☐	☐
MATH 163	☐	☑	☐	☐	☐
PHYC 161	☐	☑	☐	☐	☐
CHEM 121	☐	☐	☐	☐	☑
ECE 203	☐	☐	☑	☐	☐
MATH 264	☐	☐	☐	☑	☐
MATH 316	☐	☐	☑	☐	☐
PHYC 262	☐	☐	☑	☐	☐
ENGL 219	☐	☐	☐	☐	☑
ECE 206L	☐	☐	☐	☑	☐
ECE 213	☐	☐	☐	☑	☐
MATH 314	☐	☐	☑	☐	☐
Humanities	☐	☐	☐	☐	☑
Social Behavior	☐	☐	☐	☐	☑
Technical Elective 1	☐	☑	☐	☐	☐
Technical Elective 2	☐	☐	☐	☑	☐

(b) CBCB model

Figure 3.4: The two figures represent a five-term curriculum with actual university courses. (a) The curriculum designed using the BACP model whereas (b) The same curriculum using the CBCB model. This shows the improvement achieved using the CBCB by assigning courses with relatively higher crucial values to closest terms while maintaining a balanced workloads of the terms. This layout outperform that of the RBCB model by not only assigning relevant courses to closest terms but also moving them to closest terms.

Chapter 4

Predicting Student Success Based on Prior Performance

In Chapter 3 we introduced an optimization model that schedules the crucial courses within a curriculum to the earliest possible terms while satisfying the prerequisite dependency relationships and maintaining a balanced workload across all terms. The main objective of this model is to offer students "efficient" curricula that could allow them to graduate on time even if they fail some of these courses. Now that we have an "efficient" curriculum, it would be essential to follow the progress of the students attempting this particular curriculum. Perhaps the best way to track the student progress would be to predict their academic performance in advance and accordingly give them suitable advice. This chapter presents a machine learning model that achieves this goal. The results show that, by presenting curricula as BBNs, we can predict student performance with high accuracy.

4.1 Introduction

Obviously, students progress within a degree program has a direct influence on graduation rates. Hence any effort aimed at enhancing or predicting the progress of a student in order to provide earlier advisement and/or interventions has the potential to positively impact graduation rates. Thus, in this chapter we propose a probabilistic graphical model that allows us to reason about a student performance and progress. In particular, we use a BBN model to represent the curriculum graphs of specific degree programs. Based upon the performance of a student in a given semester, we hypothesize that the BBN model can predict the future performance of the student in subsequent semesters. The model developed in this chapter was applied to a number of different degree programs at the UNM, and was able to predict the final GPA of the students with small error.

4.2 Background and Related Work

In Chapter 2, we developed various metrics related to curricular complexity that correspond to the ease with which a student may satisfy the degree requirements associated with a given degree. These metrics were intended to measure the role that the structure of a curriculum plays in student academic success and accordingly suggest enhancement to the curriculum structure in an attempt to help students perform better in their respective academic programs. However this work did not take into account the performance of a student in progress. In Chapter 3, we proposed the SPR in order to study the progress of a student in a curriculum in each semester by investigating the structural properties of individual curricula, taking into account the degree to which individual courses in a curriculum may impact student progress. This previous work takes into account the cruciality of particular courses in a curriculum,

as well as the grades that students earn, in order to measure student progress. In this chapter, we take into account this student progress in order to predict future performance.

4.3 Bayesian Belief Networks

A BBN is a graphical structure that allows one to represent and reason about an uncertain domain. For a set of variables $X = (X_1, \ldots, X_n)$, a Bayesian network consists of a network structure S that encodes a set of conditional independence assertions about variables in X, and a set P of local probability distributions associated with each variable [22]. An example of a BBN which represents a subset of network behavior through variables namely, *Ahmad Oversleeps*, *Traffic* and *Ahmad Late* (as nodes) and two directed edges is shown in Fig. 4.1. An edge from one node to another implies a direct dependency between them, with a child and parent kind of relationship. To quantify the strength of relationships among the random variables, conditional probability functions are associated with each node, such that $P = \{p(X_1|\Pi_1), \ldots, p(X_n|\Pi_n)\}$ where Π_i is the parent set of X_i in X. If there is a link from X_i to X_j , then X_i is a parent of X_j and thus it belongs to Π_j. For discrete random variables the conditional probability functions are represented as tables, called Conditional Probability Tables (CPTs). For a typical node A , with parents $B_1, B_2, \ldots, B_n$, there is associated a CPT as given by $P(A|B_1, B_2, \ldots, B_n)$.

The main principle behind BBNs is Bayes rule:

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)} \tag{4.1}$$

where $P(H)$ is the prior belief about a hypothesis H , $P(e|H)$ is the likelihood that evidence e results given H, and $P(H|e)$ is the posterior belief in light of evidence

e. This implies that belief concerning a given hypothesis is updated upon observing new evidence.

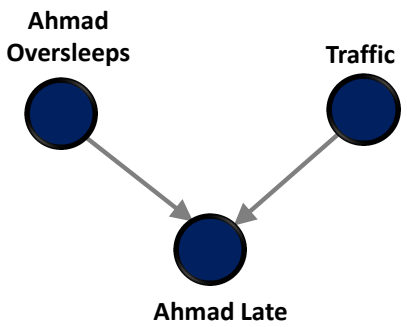


Figure 4.1: An illustrative Bayesian Belief Network.

4.3.1 Inference Features

BBNs support three types of learning: structural, parameter and sequential. The structure of the BBN can be constructed manually by a subject matter expert or through structure learning algorithms—PC and NPC algorithms [46,54]. Parameter learning uses past data as the basis for learning the parameters through algorithms. One such algorithm, Expectation Maximization (EM), is particularly useful for parametric learning [27]. In order for the model to reflect behavior in the problem domain, the parameters of the model need to be updated based on observations. This process is termed sequential learning [41]. Evidence about a particular node is used to update the beliefs (posterior probabilities) of other nodes of the BBN. The BBN framework supports predictive and diagnostic reasoning and uses efficient algorithms

for this purpose [30]. In this chapter, predictive reasoning will be the main approach implemented in the BBN framework.

4.3.2 Application to our framework

The performance of a student in a given class may be used as a measure to predict competence or skills in later classes [12]. In other words, the history of a student's academic skills tells us something about future performance. For instance, an 'A' high school student is generally expected to do better in college than a 'C' student, other factors being the same. Correspondingly, it makes sense that a college student who earns an 'A' in *Calculus II*, for instance, should be expected to earn a higher grade in *Calculus III* than those who earn a 'D'. In Fig. 4.2 , the application of BBN in the context of course network aims at predicting the grades of the courses for a given student based on the evidence of previous grades, age, gender, educational level of parents, emotional factors, etc.

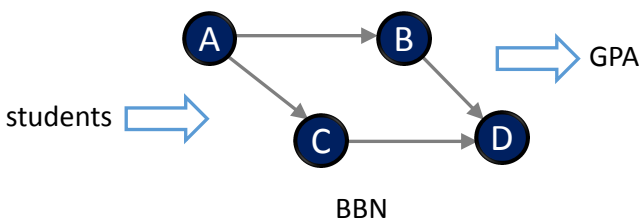


Figure 4.2: BBN in the context of a course network.

4.4 BBN Edges

Ultimately, degree attainment requires the satisfaction of all requirements associated with a degree program. The set of requirements associated with a particular degree program, along with the relationships between the individual requirements (e.g., course prerequisites) can be represented as directed acyclic graph, with a directed edge from node A to node B in the graph denoting that degree requirement A must be satisfied prior to the satisfaction of degree requirement B. Typically, a degree requirement is satisfied by passing a particular course, and the precedence relationships in the graph correspond to course prerequisites. A student satisfies all degree requirements, and therefore receives the associated degree, once they have traversed this graph, visiting every node according to the precedence relationships in the graph.

In our proposed framework, however, the edges of the BBN for the curriculum graph are not only restricted to prerequisite relationships. Basically a directed edge from node A to node B in the BBN of the curriculum graph denotes that the student performance in degree requirement A has a *direct influence (DI)* on predicting that of degree requirement B. Thus the presence of a *direct influence* edge between two requirements in the BBN does not imply the presence of a prerequisite relationship PR, however we hypothesize that the opposite is true. Accordingly PR edges are a subset of DI edges that is $PR \subseteq DI$. In other words, we consider in this work that the presence of prerequisite relationships among the courses in the BBN indicate a *direct influence* on predicting the level of performance in the direction of the edge.

4.5 BBN Nodes

Basically the variables influencing the predication of the performance of a student in a given course are not only restricted to the performance of previous courses. Many factors, other than performance of previous courses, have direct impact on predicting future performance. As mentioned in Chapter 1, studies have shown that age and gender [16,17], academic background [10], educational level of parents [56], emotional and social factors [8], and even the complexity measure of teacher's lecture notes [28] have direct influence on student progress. Typically, the BBN of a curriculum graph would be something similar to that illustrated in Fig. 4.3.

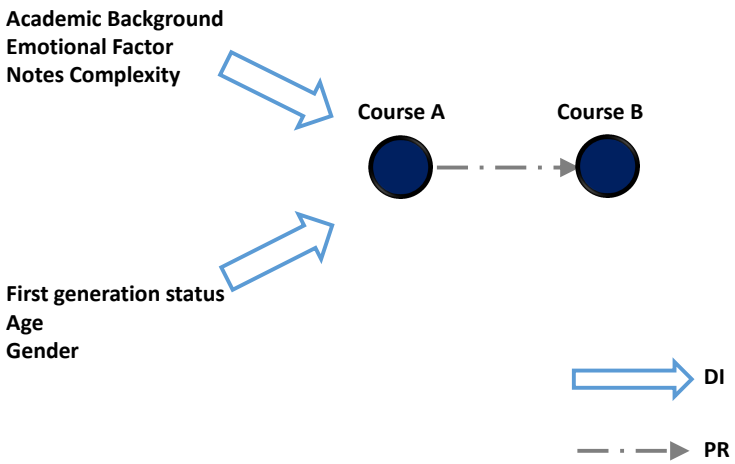


Figure 4.3: BBN model of the curriculum graph.

4.6 Implementation Aspects

For the purpose of a proof of concept, in this chapter we present a basic network topology. In particular, the only variable (i.e., node) that will be illustrated in the BBN is the performance of a student in a course (i.e., no variables related to age,

gender, educational level of parents, emotional factors, etc.), which will be a discrete variable. The states of the course variable are the letter grades associated with the course, that is $A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-$ and F. Also it is assumed that the only edge type present in the BBN is the prerequisite relationship PR. Hence, in this work, we model the BBN of a curriculum X consisting of n degree requirements as a directed graph $GF_X = (V, E)$, where each vertex $v_1,\ldots,v_n \in V$ represents a course in X, and there is a directed edge $(v_i, v_j) \in E$ from course v_i to v_j if v_i must be satisfied prior to the satisfaction of v_j . The final structure of the BBN for a curriculum graph will be something similar to that shown in Fig. 4.4.

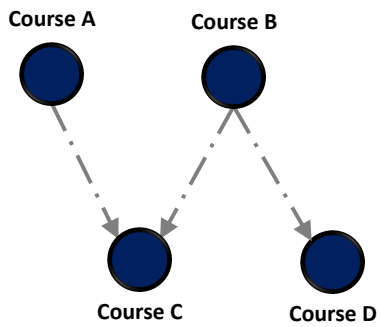


Figure 4.4: BBN model of the curriculum graph implemented in our framework. Note that the course variable is the only node presented in this BBN model and PR edges are the only links relating these type of nodes.

4.6.1 Decision-Making Policy

To meet the objective of predicting the grades of the courses to be taken by a student attending a given degree program, we need to design a policy to assign a grade for the courses to be taken in the future once we determine their respective marginal

probabilities based on the evidence of the grades of previously taken courses. Denote by $L = \{A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F\}$ the set of grade letters assigned to a course and $G = \{4.3, 4, 3.7, 3.4, 3, 2.7, 2.3, 2, 1.7, 1.4, 1, 0.7, 0\}$ the set of grades mapping L. For a course i, upon retrieving an evidence e, a decision is made using two methods:

1. Maximum a Posteriori Probability (MAP) estimate:

$$g = \underset{g \in G}{\operatorname{argmax}}\ p(g|e), \tag{4.2}$$

   where $p(g|e)$ is the marginal probability of course i states based on evidence e which is the set of grades of previous courses.

2. Expected Grade (EG) estimate:

$$\hat{g} = \mathbb{E}(G) = \sum_{g} g p(g|e). \tag{4.3}$$

Note that one of the predicted letter grades for a given course might be F. In this case no data is available to fill in the CPTs due to the fact that students cannot move to another class if they fail its respective pre-requisite(s). For instance, we cannot fill a CPT row querying about the probability of a student earning a C on $CalculusIII$ conditioned on getting F on $CalculusII$. Simply, the student must get D or higher on $CalculusII$ to go for $CalculusIII$. In other words, the student has to repeat the course and pass it. To overcome this problem, we use a Markov chain model. The transition probabilities are graphically represented by the transition diagram shown in Fig. 4.5 with a 13 state Markov chain model representing the letter grades. Thus in case the BBN model predicts a F grade for a given course, it will use the Markov chain model to choose the letter grade (other than F) with the maximum transitional probability, that is

$$g = \underset{l \in L}{\operatorname{argmax}}\ p(l) \tag{4.4}$$

where $p(l)$ is the transition probability.

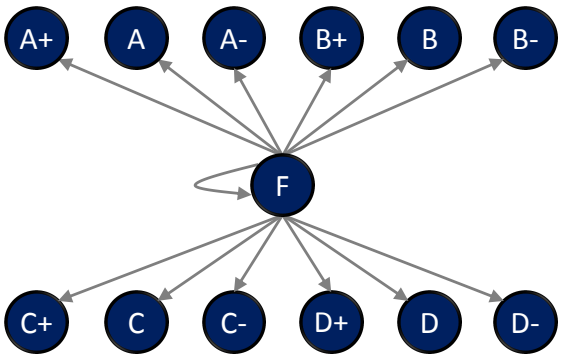


Figure 4.5: A 13-state Markov chain model.

4.7 Simulation Results

In an attempt to empirically validate our proposed BBN framework, we analyzed actual university data from the UNM[1]. For this we used the data of 115,746 students to generate the CPTs for all the courses in the BBN. Then we chose 400 students, who had already earned their degrees, randomly from different departments (eg., mechanical engineering department, chemical engineering department, electrical engineering department and nuclear engineering department) to test the framework. The performance of the framework is measured using mean squared error (MSE):

$$err^t = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2, \tag{4.5}$$

[1]All the UNM data used in this work are found at s3.amazonaws.com/employing-bayesian-belief-networks-for-course-networks/bbn-data.zip

where err^t is the MSE measured based on the evidence of the grades of the courses taken between semester 1 and semester t; n is the number of students;$\hat{Y}$ is a vector of n GPA predictions; $Y$is the vector of the true GPA (the actual GPA values of the students).

4.7.1 Data Pre-processing

It is well known that building relationships between courses based on pre-requisite links is not trivial. For example a course i, may be a co- or pre-requisite to another course j and/or vise-versa. In order to deal with such relationships, some assumptions are made:

1. If course i is a co- or pre-requisite to course j, we assume that i is a pre-requisite to j. In other words, we assume the worst case scenario where course i and j cannot be taken in the same academic term.

2. If course i is a co- or pre-requisite to j and vice-versa or in other words if courses i and j are co-requisites, we consider the worst case scenario in which one of the courses is considered to be the pre-requisite of the other. In this case we eliminate cycles from our graph.

4.7.2 Numerical Results

As mentioned previously, 400 students were chosen randomly from four different departments as the test set. The courses taken by these students are spread over 18 semesters (i.e., six years). The MSE is measured at each semester where the grades of the courses taken by the students up until that semester are entered as evidence to the BBN framework. The MSE is measured using two different methods illustrated in Eq. (5.3) and Eq. (5.4) to calculate the predicted GPA vector $\hat{Y}$. To demonstrate

the practicality of our approach, we compared our framework to another one where no edges are present. In other words, we generated another graph where we assumed that the performance of a student in one class does not have any influence on the performance of other classes (i.e. no edges are present). As in the BBN framework, we also measured the MSE using Eq. (5.3) and Eq. (5.4) to calculate the predicted GPA vector $\hat{Y}$. Note that no evidence is present anymore in Eq. (5.3) and Eq. (5.4) regarding the second framework. Fig. 4.6 illustrates the performance of both frameworks. This figure shows that the MSE values are decreasing gradually throughout the 18 semesters upon receiving new evidence e. The red curves show the MSE values for the BBN framework whereas the blue ones show those of the second framework. The dashed curves present the MSE values using the MAP estimate method illustrated by Eq. (5.3), whereas the solid ones presents those using the EG estimate method illustrated by Eq. (5.4). From the figure it is seen that the MAP estimate method outperforms that of the EG estimate in both frameworks. Besides, the curves show that the BBN framework outperforms the other framework in both methods (i.e., MAP and EG). These results clearly illustrate the influence of a student's present performance on predicting her future performance. Using the BBN framework, upon receiving the grades of the first semester (i.e., evidence e), for instance, the MSE value (using MAP estimate) is measured to be 0.16. However, using the second framework, the MSE value is measured to be 0.55. On the other hand, comparing the MSE values for both frameworks, using the EG estimate method, upon receiving the grades of the first semester, shows a gap as well. For semester one, the MSE value, using the BBN framework, is 0.37 whereas that, using the second framework, it is 0.616. On a scale of 4.3 (i.e., the maximum GPA value that can be achieved), it is obvious that the MSE value, using the second framework, is significantly high. This result illustrates the significance of the BBN framework in providing a better probability distribution, compared to the second framework, of the letter grades for a given course upon receiving an evidence e (i.e., marginal probability). Basically,

the results show that the BBN framework gives students a more accurate prediction about the probability distribution of the letter grades for their future courses.

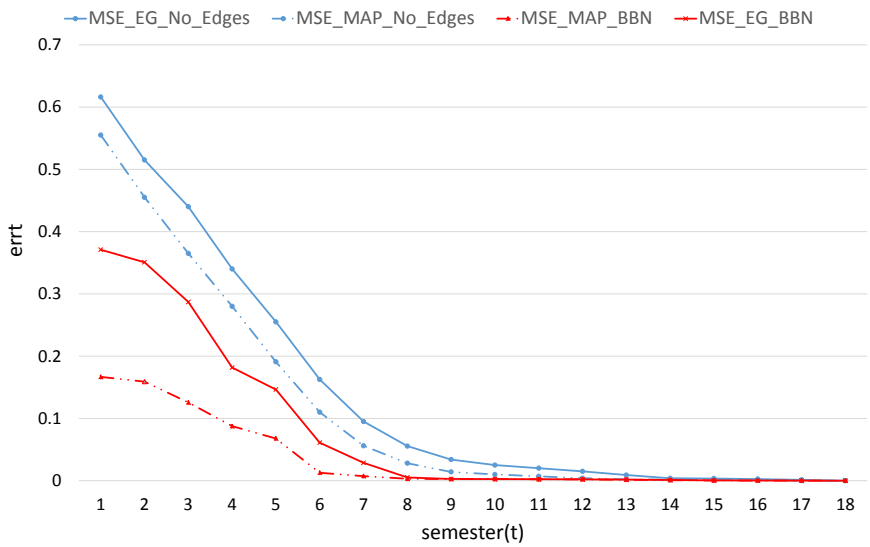


Figure 4.6: MSE values of the two frameworks for 18 semesters with 3 semesters per year. The red curves show the MSE values for the BBN framework whereas the blue ones show those of the second framework (i.e. no edges). Besides, the dashed curves presents the MSE values using the MAP estimate method illustrated by Eq. (5.3) whereas the solid ones presents those using the EG estimate method illustrated by Eq. (5.4).

Chapter 5

Employing Markov Networks on Curriculum Graphs

5.1 Introduction

In Chapter 4, we implemented a model to predict student performance by abstracting the courses of a curriculum along with their respective prerequisites into a BBN. A major limitation of this model, however, is considering that the performance in a given course can be only structurally influenced by its prerequisite courses. This is only partially true. The performance in courses in a given semester may be a good predictive indicator for the performance in courses in subsequent semesters even in the absence of prerequisite relationships. This component was missing in the BBN model. Thus, in this chapter, we take into account the student progress in order to predict future performance by using a MN to represent the curriculum graphs of the degree programs. Based upon the performance of a student in a given semester, the MN model predicts the future performance of the student in subsequent semesters. The model developed in this chapter was applied to the same degree programs and

the same students used in Chapter 4. This model outperforms the BBN model by predicting the GPA distribution of the students with minimal error.

5.2 Markov Networks

Markov networks are probabilistic models that are represented by undirected graphs and can, in contrast to directed graphical models, contain arbitrary cycles. The probability distribution factors over the maximal cliques ξ of the graph—these are the subsets of fully connected nodes. Each maximal clique $c \in \xi$ is associated with a potential function φ_c that assigns a positive value to the subset of random variables $\mathbf{x}_{(c)}$ represented by the clique [50]. The potential functions φ_c do not necessarily have a probabilistic interpretation, and are not directly related to marginal distributions of subsets of nodes. The joint distribution of a MN can be written as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \xi} \varphi_c(\mathbf{x}_{(c)}) \tag{5.1}$$

where

$$Z = \sum_{\mathbf{x}} \prod_{c \in \xi} \varphi_c(\mathbf{x}_{(c)}) \tag{5.2}$$

is a normalization constant (or partition function) that guarantees $p(\mathbf{x})$ integrates to 1. MN can be defined in terms of the conditional independence properties of each random variable. Each node v is conditionally independent of all other nodes, given its direct neighbors.

The values of some variables (nodes) in the graphical model are usually observed in a concrete application; inference means computing information about variables

$\mathbf{x}$, given observed or evident variables $\mathbf{y}$. Quantities of interest may be marginal distributions of one or some of the unobserved variables; for our purposes it will be an "optimal" configuration of all unobserved variables $\mathbf{x}$, given observed variable $\mathbf{y}$, which is governed by the posterior distribution $p(\mathbf{x}|\mathbf{y})$. Exact inference in graphical models is generally very hard, which is the reason why approximative inference is usually employed in practice. There are many different classes of approximative inference algorithms: variational, sampling-based, (local) optimization, graph-cuts, etc see [47] for an overview. Note that in this work we implemented a pairwise MN where the maximal cliques only connect pairs of nodes.

5.3 MN Edges

The MN associated with the set of requirements of a particular degree program, along with the relationships between the individual requirements can be represented as undirected cyclic graph, with an edge between node A and node B in the graph denoting the presence of a strong correlation in the performance of the student in both requirements.

5.4 MN Nodes

As previously mentioned, the variables influencing the prediction of the performance of a student in a given course are not only restricted to the performance of previous courses. Many factors, other than performance of previous courses, have direct impact on predicting future performance [8, 10, 16, 17, 28, 56]. Hence, edges, presenting strong correlation between nodes, are not only restricted to course nodes. A correlation can be found between different types of nodes. For example, there might exist a correlation between the performance of a student in one course and her respective

gender.

5.5 Implementation Aspects

For the purpose of a proof of concept, in this chapter we present a basic network topology. In particular, the only variable (i.e., node) that will be illustrated in the MN is the performance of a student in a course (i.e., no variables related to age, gender, educational level of parents, emotional factors, etc.), which will be a discrete variable. Similar to the BBN model, the states of the course variable are the letter grades associated with the course, that is $A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-$ and F. Hence we model the MN of a curriculum X consisting of n degree requirements as an undirected graph $GF_X = (V, E)$, where each vertex $v_1,\ldots,v_n \in V$ represents a course in X, and there is an undirected edge $(v_i, v_j) \in E$ between course v_i and course v_j if there exists a strong correlation coefficient r with high significance level (i.e., low *p–value*) [26]. In particular we consider that an edge exists between two courses if:

1. *R–squared* ≥ 0.11. Studies using linear regression models have reported *R–squared* values between 0.11 and 0.4 — values are not uncommon for human behavior studies (this is especially true in the field of grade prediction [10, 17, 29]).

2. *p–value* < 0.05

The final structure of the MN for a curriculum graph is illustrated in Fig. 5.1.

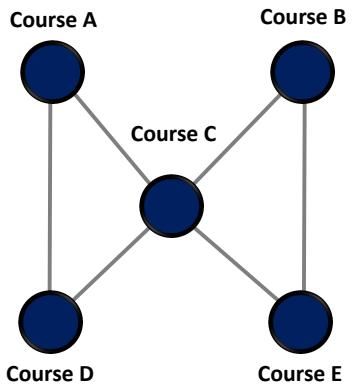


Figure 5.1: MN model of the curriculum graph implemented in our framework. Note that the course variable is the only node presented in this MN model.

5.5.1 Decision-Making Policy

Similar to the BBN model, we denote by: L={A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F} the set of grade letters assigned to a course and G={4.3, 4, 3.7, 3.4, 3, 2.7, 2.3, 2, 1.7, 1.4, 1, 0.7, 0} the set of grades mapping L. For a course i, upon retrieving an evidence e, a decision is made using two methods:

1. Maximum a Posteriori Probability (MAP) estimate:

$$g = \underset{g \in G}{\operatorname{argmax}}\ p(g|e), \tag{5.3}$$

   where $p(g|e)$ is the marginal probability of course i states based on evidence e which is the set of grades of previous courses.

2. Expected Grade (EG) estimate:

$$\hat{g} = \mathbb{E}(G) = \sum_{g \in G} g p(g|e). \tag{5.4}$$

5.6 Simulation Results

To empirically validate our proposed MN framework and compare its performance to that of the BBN, we used the same data of 115,746 students and generated the potential functions φ_c for the MN. In particular a potential function is chosen to be the joint distribution of any two connected nodes (i.e., courses). Then we chose the same 400 students we used in the BBN as our test collection. Again, the performance of the framework was measured using the mean squared error (MSE):

$$err^t = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2, \tag{5.5}$$

where err^t is the MSE measured based on the evidence of the grades of the courses taken between semester 1 and semester t; n is the number of students;$\hat{Y}$ is a vector of n GPA predictions; $Y$is the vector of the true GPA (the actual GPA values of the students).

5.6.1 Numerical Results

As mentioned previously, 400 students were chosen randomly from four different departments as our test collection. The courses taken by these students are spread over 18 semesters (i.e., 6 years with 3 semesters per year). The MSE is measured at each semester where the grades of the courses taken by the students up until that semester are entered as evidence to the MN framework. The MSE is measured using two different methods illustrated in Eq. (5.3) and Eq. (5.4) to calculate the predicted GPA vector $\hat{Y}$. Similar to the BBN model, to demonstrate the practicality of our approach, we compared our framework to another one where no edges are present. In other words, we generated another graph where we assumed that there is no correlation between courses. As in the MN framework, we also measured the

MSE using Eq. (5.3) and Eq. (5.4) to calculate the predicted GPA vector $\hat{Y}$. Note that no evidence exists anymore in Eq. (5.3) and Eq. (5.4), regarding the second framework, after removing the edges. Fig. 5.2 illustrates the performance of both frameworks. The blue curve shows the MSE values for the MN framework using the MAP estimate method illustrated by Eq. (5.3) whereas the red curve shows those using the EG estimate method illustrated by Eq. (5.4) (These two curves are clearly presented in Fig. 5.3). On the other hand, the purple curve shows the MSE values for the second framework (i.e., No_Edges) using the MAP estimate method whereas the green one shows those for the second framework using the EG estimate method. This figure shows that the MSE values are decreasing gradually throughout the 18 semesters upon receiving new evidence e. Apparently, from the figure, it is the case that the MAP estimate method outperforms that of the EG estimate in both frameworks. Furthermore, the curves show that the MN framework outperforms the other framework in both methods (i.e., MAP and EG). These results clearly illustrate the influence of a student's present performance on predicting future performance. Using the MN framework, upon receiving the grades of the first semester (i.e., evidence e), for instance, the MSE value (using MAP estimate) is measured to be 0.0449. However, using the second framework, the MSE value is 0.555. On the other hand, for semester one, the MSE value (EG estimate), using the MN framework, is 0.0657, whereas using the second framework it is 0.6162. On a scale of 4.3 (i.e., the maximum GPA value that can be achieved), it is obvious that the MSE value, using the second framework, is significantly higher. This result illustrates the significance of the MN framework in providing a better probability distribution, compared to the second framework, of the letter grades for a given course upon receiving an evidence e (i.e., marginal probability). Basically, the results show that the MN framework gives students a more accurate prediction about the probability distribution of the letter grades for their future courses. More importantly, the MN model outperforms the BBN. It is clearly shown that a marginal error of 0.0449 may be achieved using the

MN model whereas a marginal error of 0.16 can be achieved using the BBN model upon receiving the grades of the first semester. This result is predictable. The MN model takes into consideration implicit relationships among courses to predict the future performance that are not considered using the BBN model. It was shown in Chapter 4 that the BBN representation of a curriculum predicts the student performance based only on prerequisite relationships. Thus, more information is present in the MN model which is reflected by more prediction accuracy.

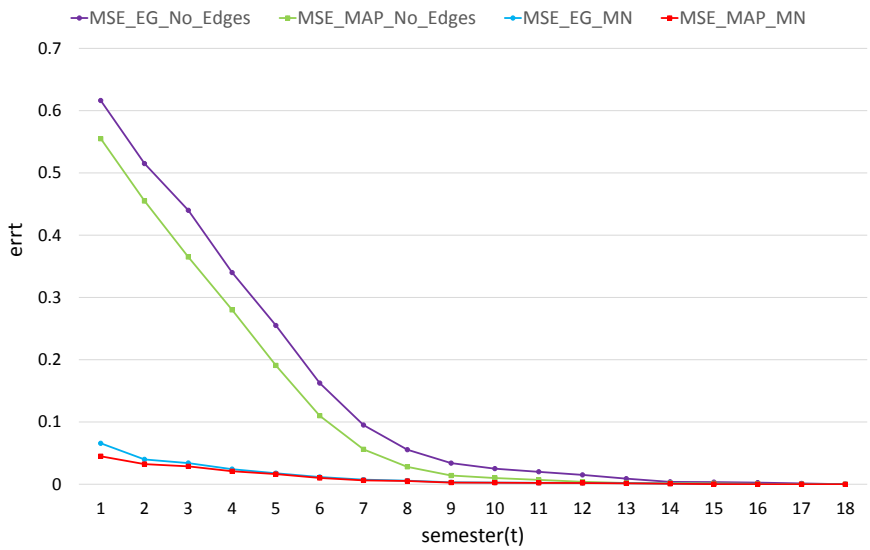


Figure 5.2: MSE values of the two frameworks for 18 semesters with three semesters per year. The purple curve shows the MSE values for the second framework (i.e., No_Edges) using the EG estimate method whereas the green one shows those using the MAP estimate method . The blue curve shows the MSE values for the MN framework using the EG estimate method whereas the red one shows those using the MAP estimate method.

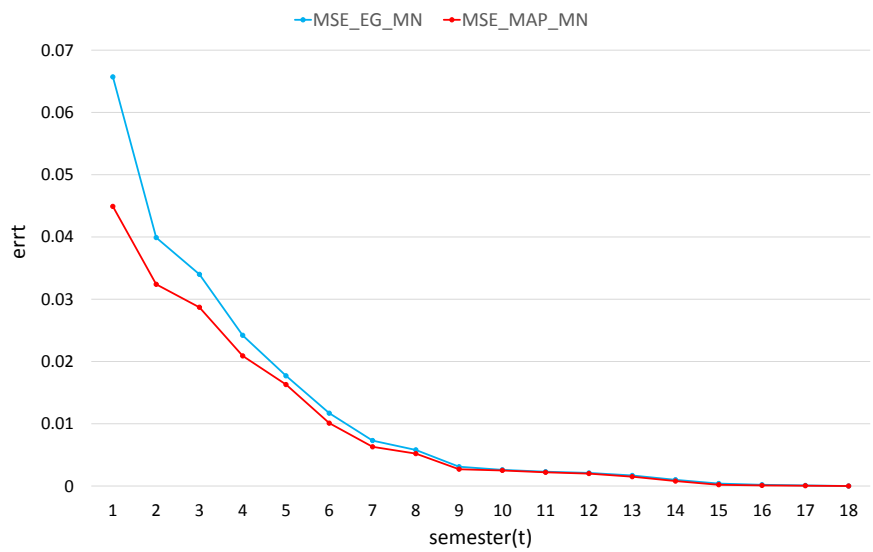


Figure 5.3: MSE values of the MN framework. The blue curve shows the MSE values using the EG estimate method whereas the red one presents those using the MAP estimate method.

Chapter 6

The Impact of Course Enrollment Sequences on Student Success

6.1 Introduction

In the previous chapters we defined crucial courses in a curriculum based on their respective blocking and delay factors. We discussed that it would be more efficient to move these types of courses to the earliest possible terms. We called this layout the "efficient" curriculum. It is efficient in a sense that it might give students better chances to graduate on time even if they fail some of these courses. Now that we have an "efficient" curriculum, it would be essential to follow the progress of students. Thus, in Chapters 4 and 5, we provided two predictive models that achieve this goal. Representing curricula as BBNs and MNs, we were able to predict the performance of students with high accuracy. This step is important as it offers students earlier intervention when needed. However one important thing is still missing in this pipeline process. What if students do not follow the order of the courses' sequences of the "efficient" curriculum? Would this impact the progress

of students? These questions are important to answer because most of the times students do not follow the order of the courses' sequences of curricula and thus it would be essential to study and analyze its impact on their progress. Thus, in this chapter we address student progress at the most basic level, by investigating the structural properties of individual curricula, taking into account the degree to which course enrollment sequences in a curriculum may impact student success.

6.2 Proposed Framework

In the previous chapters it was shown that improving the overall graduation rate of a university is facilitated by a smooth traverse of students through the degree requirements of its academic programs. Continuous progress motivates students to persist and continue firmly in their programs of study in spite of the difficulties they may encounter. A critical motivating factor involves sustaining relatively high grades while making steady progress. In this chapter we exclude pre-institutional factors, and work to uncover the best sequence of course enrollments that lead to high grades and graduation. Our approach involves analyzing the course enrollment sequences of students who graduated with high GPAs, versus the course sequences of those who did not. The notion being, we would like current students to imitate the behavior of the successful students who preceded them. The framework we developed is as follows:

Step 1. Split the data representing student information and course enrollment history into n datasets $\{D_1, \ldots, D_n\}$ corresponding to their respective labels $\{L_1, \ldots, L_n\}$. One label can be students who graduated with "high GPAs" while another can be students who graduated with "low GPAs".

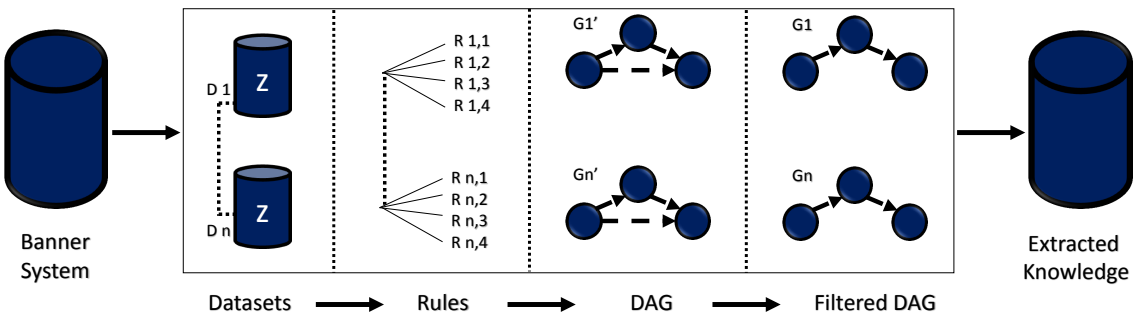


Figure 6.1: The course enrollment sequence analysis framework.

Step 2. Generate course enrollment sequential patterns $\{R_1, \ldots, R_n\}$ corresponding to the respective datasets $\{D_1, \ldots, D_n\}$ using Sequential Pattern Mining (SPM).

Step 3. Generate Directed Acyclic Graphs (DAGs) $\{G'_1, \ldots, G'_n\}$ representing the patterns $\{R_1, \ldots, R_n\}$ generated in step 2.

Step 4. Generate DAGs $\{G_1, \ldots, G_n\}$ by applying a transitive reduction algorithm to filter out the transitive edges within graphs generated in step 3.

Step 5. Apply different graph theory techniques and complex network metrics to analyze, study and compare the graphs generated in step 4.

Figure 6.1 shows the steps of the proposed framework used to extract knowledge from student course enrollment histories.

6.3 Implemented Techniques

This section presents a detailed explanation of the different methods used to design our proposed framework. It also illustrates the different assumptions made while designing the framework.

6.3.1 Course Enrollment Sequential Patterns

SPM is a new method in the data mining field used to extract knowledge [6]. The input data of this method is made up of a set of transactions. Each transaction contains a set of items which could be associated with a time flag. The output data is a sequential pattern that is also made up of a set of items. The sequential pattern shown in the output data is typically generated based on a user-defined minimum support value. This value reflects the percentage of the input transactions that contain this particular sequence. A pattern might look something like $A \Rightarrow B$, where A and B are two item-sets explained as, if A occurred, then B is most likely to occur, given the minimum support value. This method has been used in different fields such as stock market analysis [14, 61], weather observation [21], e-learning [20] and drought management [18].

In this chapter we apply SPM in order to extract sequential patterns from student course enrollment histories so that we can study and analyze these patterns with the goal of improving student performance and hence success. An example course enrollment sequential pattern is as follows:

$\{PHYC161L\} \rightarrow \{ECE371\}$;
$\{PHYC161\} \rightarrow \{ECE371\}$;
$\{PHYC161\} \rightarrow \{ECE314\}$;

$\{MATH316\} \rightarrow \{ECE314\}$;
$\{MATH316\} \rightarrow \{ECE213\} \rightarrow \{ECE314\}$;
$\{PHYC161L, PHYC161\} \rightarrow \{PHYC262, ECE213\} \rightarrow \{ECE371\}$;

The last pattern in this example is interpreted as: students who take $PHYC$ $161L$ and $PHYC$ 161 together are likely to take $PHYC$ 262 and ECE 213 in a following semester, followed by ECE 371 in a latter semester. The fifth pattern indicates that students who take $MATH$ 316 in one semester are more likely to take ECE 213 in the next semester, followed byECE 314.

6.3.2 Course Enrollment SPs as a DAG

A course enrollment DAG is a directed graph with no directed cycles. That is, it is formed by a collection of courses and directed edges, each edge connecting one course to another, such that there is no way to start at some course c and follow a sequence of edges that eventually loops back to c. The directed edges show the sequence, or the flow, of course enrollments that students are likely to follow in a particular academic program or institution. The DAG of the sequential patterns presented in the previous section is shown in Fig. 6.2.

6.3.3 Transitive Reduction of the DAG

A transitive reduction of a directed graph is the deletion of a number of edges in a way that preserves the reachability measure of the given graph. It is important to consider the transitivity relationship links between the vertices of the course enrollment DAG. These types of links must be deleted. For example if course A is preceding both courses B and C, while C itself is preceding B, then there is no need to show that

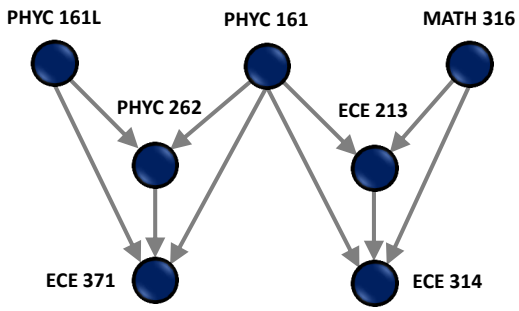


Figure 6.2: DAG of course enrollment sequential patterns.

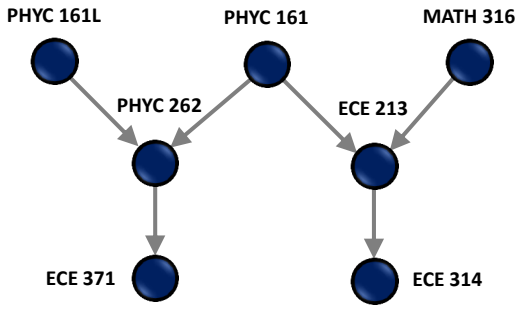


Figure 6.3: Filtered DAG using a transitive reduction algorithm.

A is preceding B. Otherwise, A assumes extra information that is not deserved. We use a transitive reduction algorithm to filter out the transitive edges within the course enrollment DAG generated in step three [7]. The transitive reduction step on the DAG shown in Fig. 6.2 is illustrated in Fig. 6.3.

6.3.4 Graph Metrics

In this chapter we use graph theoretic tools and measures in order to study and analyze the structure of course enrollment DAGs. We use these to create comparison

measures among course enrollment DAGs in order to extract some knowledge that might help students improve their academic performance. The graph measures that will be used in this work are the following:

Cosine Similarity. Measures the similarity of two vertices in a graph by counting the number of neighbor vertices they share [43]. We modified the cosine similarity algorithm a bit in order to better fit our model. In particular, the similarity metric we used measures the similarity of a vertex v in two different DAGs G_1 and G_2 by counting the number of vertices with similar sequence positions that v shares. The modified algorithm is defined as following:

- Let X be the set of courses preceding course v in G_1.
- Let Y be the set of courses preceding course v in G_2.
- Let Z be the set of courses of both X and Y:

$$Z = X \cup Y$$

- For course $v \in \{G_1, G_2\}$, create a 3-by-$|Z|$ adjacency matrix M^v where

$$M^v_{1j} = \begin{cases} 1 & if \;\; j \in X \\ 0 & Otherwise \end{cases}$$

and

$$M^v_{2j} = \begin{cases} 1 & if \;\; j \in Y \\ 0 & Otherwise \end{cases}$$

and

$$M^v_{3j} = \begin{cases} \frac{1}{|S^v_{1j} - S^v_{2j}| + 1} & if \;\; j \in \{X \cap Y\} \\ 0 & Otherwise \end{cases}$$

where S^v_{1j} and S^v_{2j} are the positions of course j with respect to v in the preceding sequences of G_1 and G_2 respectively.

Accordingly, the cosine similarity of course v in G_1 and G_2 is

$$similarity_v(G_1, G_2) = \begin{cases} \frac{\sum_j M^v_{1j} * M^v_{2j} * M^v_{3j}}{|M^v_1||M^v_2|} & if \ \ |Z| > 0 \\ 1 & Otherwise \end{cases}$$

As an example, consider the two DAGs shown in Fig. 6.5(a) and Fig. 6.5(b). The matrix M used to compute the cosine similarity of vertix v is shown in Table 6.1. The courses preceding v in this example are A, B, C and D. The sequence position of these courses with respect to v are shown in Fig. 6.5(a) and Fig. 6.5(b).

$$similarity_v(G_1, G_2) = \frac{1 + 1/2 + 1/2 + 0}{\sqrt{4}.\sqrt{3}} = \frac{2}{2\sqrt{3}} \simeq 0.6$$

Breadth First Search. Traverses a graph starting from a source vertex and then explores the neighboring vertices before going to the next level neighbors [43]. This method was to determine the enrollment term of course v in a DAG G, denoted as $t_v(G)$. This measure is important in order to compare the enrollment term of v in different DAGs. The algorithm used to compute $t_v(G)$ is detailed in Algorithm 1. Fig. 6.4 provides a visualization of Algorithm 1.

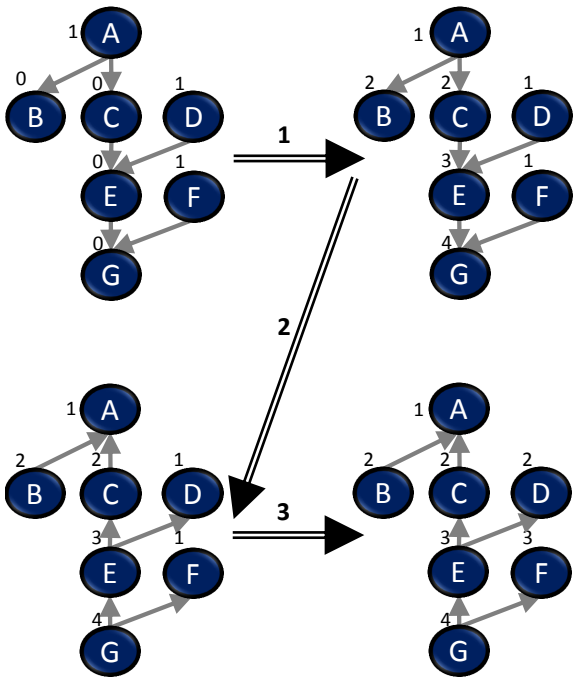
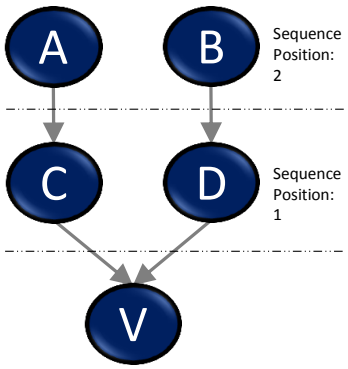


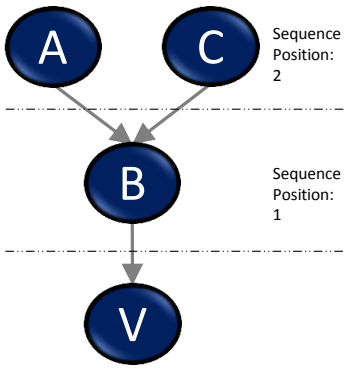
Figure 6.4: This figure shows the process used in the TAA in order to compute the term enrollment values for courses A, B, C, D, E, F and G.

6.4 Case Study: Electrical Engineering Students at UNM

In order to empirically validate our course enrollment framework, we analyzed actual university data provided by UNM. In particular, we studied and analyzed the course enrollment history for students enrolled in the Electrical Engineering (EE) program at UNM. The primary goal was to determine if the sequence of course enrollments of students who graduated with high GPAs is different from those who graduated with relatively low GPAs. The idea will be to use this information in advising sessions in order to guide new students in a manner that might improve their academic performance, i.e., by following the general enrollment patterns of the successful students who preceded them.



(a) Sequence position of courses A, B, C and D with respect to V in the DAG G_1.



(b) Sequence position of courses A, B and C with respect to V in the DAG G_2.

Figure 6.5: This figure shows two DAGs G_1 and G_2 with their respective course enrollment sequences. In particular it shows the sequence position of courses A, B, C and D with respect to V.

M	A	B	C	D
M_1^v	1	1	1	1
M_2^v	1	1	1	0
M_3^v	1	1/2	1/2	0

Table 6.1: A matrix M used to compute the cosine similarity of vertix v in the DAGs shown in Fig. 6.5(a) and Fig. 6.5(b).

6.4.1 Basic Statistics

As mentioned in the introduction, there are many factors, in addition to course enrollment sequences, that influence the final GPA of a university student. For instance, high school GPA is highly correlated with student success, and females tend to slightly outperform males in college [9, 59]. Thus, we performed some basic statistics in order to compute the mean high school GPA and the gender distribution for the datasets D_1 (EE students who graduated with "high GPA") and D_2 (EE students who graduated with "low GPA"). The results are presented in Table 6.2

and Table 6.3.

Based on these results, we noted that D_1 and D_2 are almost statistically equivalent. This indicates that there are factors other than high school GPA and gender that influence the student performance in universities. In this work we explore the extent to which course enrollment sequences influence student success.

	mean	*standard deviation*
D_1	3.6	0.42
D_2	3.4	0.48

Table 6.2: The mean and standard deviation values of the high school GPA for the datasets D_1 and D_2.

	$Male(\%)$	$Female(\%)$
D_1	83	17
D_2	86	14

Table 6.3: The gender distribution for the datasets D_1 and D_2.

6.4.2 Data Processing

We extracted from UNM's student information system the course enrollment histories of all EE students who were awarded a degree and were admitted as First-Time Full-Time (FTFT) students. This data was then divided into two datasets:

Dataset D1. A sequence database that contains the course enrollment histories of the students who graduated with "high GPA" values (i.e. ≥ 3.5).

Dataset D2. A sequence database that contains the course enrollment histories of the students who graduated with relatively "low GPA" values (i.e. < 3.0).

Using the generalized sequential pattern mining with item intervals algorithm [23], we then generated the sequential patterns $R1$ and $R2$ representing $D1$ and $D2$ respectively. We derived all the sequential patterns respecting a minimal support value of:

For D1$\rightarrow$ $support = 60\%$;

For D2$\rightarrow$ $support = 60\%$;

Next, we generated graphs $G1'$ and $G2'$ by converting $R1$ and $R2$ to DAGs, respectively. A transitive reduction algorithm is then applied on $G1'$ and $G2'$ to delete the transitive links and hence generate $G1$ and $G2$, respectively. The DAGs $G1$ and $G2$ representing the SPs $R1$ and $R2$ are shown in Fig. 6.6(a) and Fig. 6.6(b) respectively.

6.4.3 Basic Analysis

Based on the graphs shown in Fig. 6.6(a) and Fig. 6.6(b), it is evident that the sequence of course enrollments for "high GPA" students is quite different from that of "low GPA" students. For example, in Fig. 6.6(a), enrollment in the courses PHYC 160, PHYC 161 and PHYC 161L is not common at UNM. It can be shown that only 39% of "low GPA" students enroll in these courses at UNM. The rest of the students in this group either take these courses at a different institution (and transfer them

$Courses$	$Similarity(G_1, G_2)$	$t(G_1)$	$t(G_2)$
MATH 162	1	1	1
MATH 163	1	2	2
PHYC 160	0	2	0
PHYC 161	0	3	0
PHYC 161L	0	3	0
MATH 264	0.82	3	3
ECE 203	0.58	4	3
MATH 316	0.61	4	4
PHYC 262	0.37	4	8
MATH 314	0.67	5	5
ECE 213	0.67	5	5
ECE 360	0.63	7	7
ECE 314	0.23	6	8
ECE 420	0.32	9	13

Table 6.4: The cosine similarity and term enrollment values for courses shown in Fig. 6.6(a) and Fig. 6.6(b).

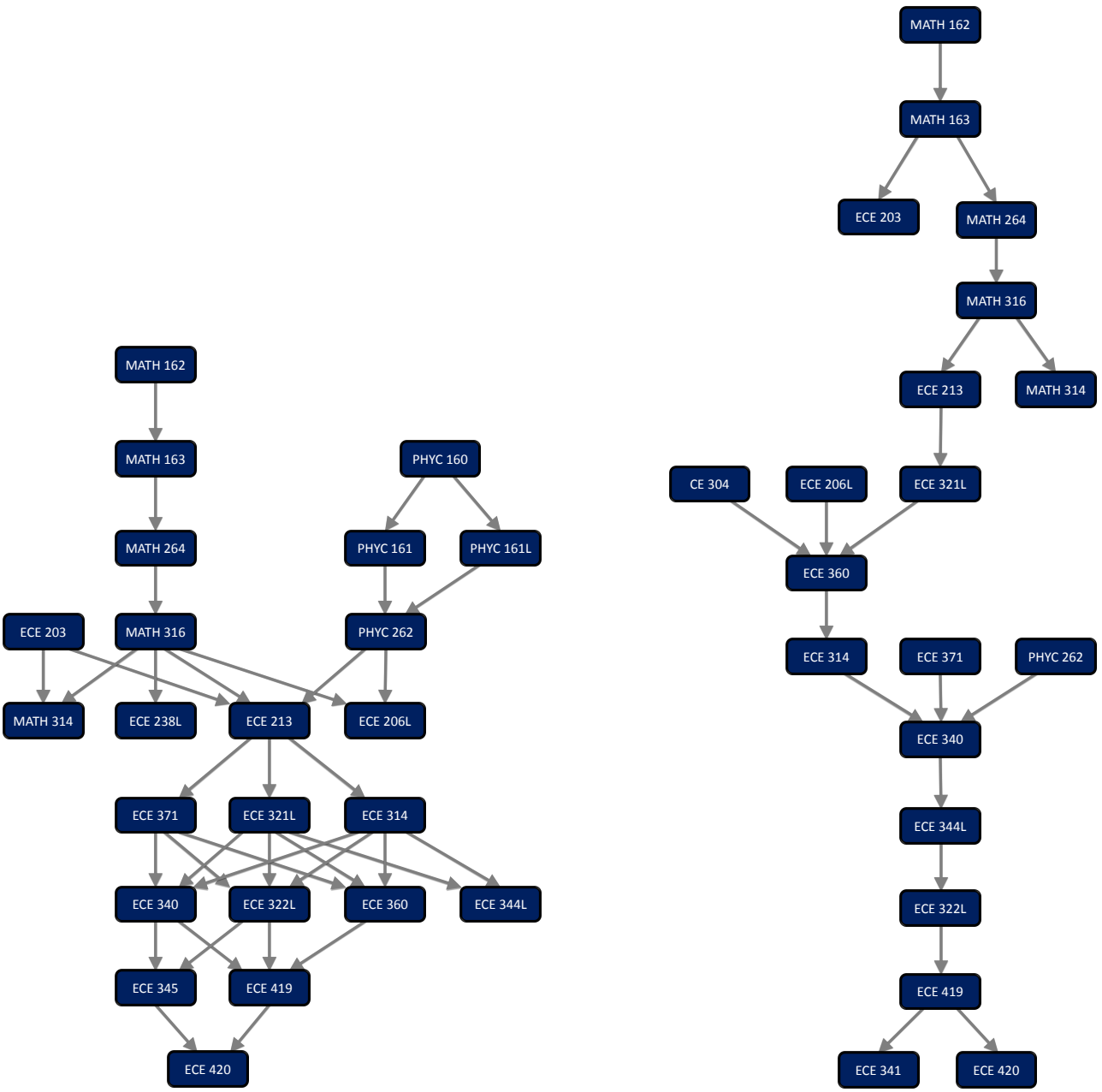
to UNM) or satisfy these requirements by replacement exams. However, this is not the case with "high GPA" students. Statistics showed that nearly 65% of these students enrolled in these courses at UNM. This result might suggest that it would be better for EE students to enroll in these courses at UNM exclusively, rather than other options, especially since these courses precede many other core courses in the program. Another example that shows the functionality of the proposed framework is the enrollment sequence of ECE 360. In Fig. 6.6(a), ECE 360 is taken after ECE 371, ECE 314 and PHYC 262; however, it is the other way around in Fig. 6.6(b). The difference in the enrollment sequence of ECE 360 in these two figures is shown by its cosine similarity value shown in Table 6.4. Another interesting observation found in this case study is the fact that the longest path in Fig. 6.6(b) is greater than that of Fig. 6.6(a). This indicates that students with "low GPA" value, on average, tend to earn their degrees later than those of "high GPA" students, that is, their time-to-degree is longer. This fact has a direct influence on the university's

graduation rate.

Algorithm 1 Term Assignment Algorithm (TAA).

TAA(Graph $G = (V, E)$, array S), S is the set of vertices without incoming edges, V is the set of vertices in G, E is the set of edges in G

```
1   for each v ∈ V do
2     d[v] ← 0, unmark all vertices
3   for each s ∈ S do
4     d[s] ← 1, mark the source
5   for each s ∈ S do
6     Enqueue(Q, s)
7       while Empty(Q) = false do
8       v ← Dequeue(Q)
9       for each u ∈ adjacent[v] do
10          if d[u] < d[v] + 1 then, is vertex u unmarked?
11            d[u] ← d[v] + 1, mark vertex u
12            Enqueue(Q, u)
13  Reverse the direction of the edges of G
14  for each s ∈ S do, S is the set of vertices without
    incoming edges
15    Enqueue(Q, s)
16      while Empty(Q) = false do
17      v ← Dequeue(Q)
18      for each u ∈ adjacent[v] do
19        if d[u] < d[v] − 1 then, is vertex u unmarked?
20          d[u] ← d[v] − 1, mark vertex u
21          Enqueue(Q, u)
22  Reverse the direction of the edges of G
```



(a) DAG G_1 (b) DAG G_2

Figure 6.6: The DAGs $G1$ and $G2$ representing the SPs $R1$ and $R2$ generated using the course enrollment histories of all undergraduate EE students who earned a degree at UNM.

Chapter 7

Conclusion

Higher education attainment proves to be one of the most important factors that characterize the rise of any society or country. Not only does it influence the social status of individuals but also the economic factors of a country in general. Unfortunately, many countries encounter major difficulties in this domain. This is driven by numerous factors and perhaps the nature of the educational process offered by many academic institutions constitutes the major bottleneck. The services offered by institutions, the competence of the instructors, the advisement arrangements provided to students, and other conditions play important roles in determining the efficacy of higher education. Thus attempts to provide assistant or support in this direction would be highly valuable.

Studies show that some services offered by some universities are effective in providing tools that would help students proceed smoothly in their academic lives. Most of these tools provide help based on the assumption that students are the source of obstruction in higher education. For example, many studies relate low graduation rates in universities to students with low ACT scores and low high school GPAs. Other studies relate low graduation rates to race, ethnicity, and gender factors. However

none of these studies, for example, investigate the issue of low graduation rates from the perspective of curricula structure. In this dissertation we explore the impact of curricula structure on student progress and graduation rates. We argued that there is an inverse correlation between the complexity of a curriculum and the graduation rate of students attempting that curriculum. We validated this claim by analyzing actual university common curricular pattern. In particular, using complex network analysis, graph theory, and machine learning techniques, we proposed a framework that quantify the complexity of a curriculum. First we introduce a new measure to compute the cruciality of the courses within a curriculum and accordingly compute the complexity of a curriculum as the sum of the crucialities of all courses in the curriculum. The framework is extended further to create an "efficient" curriculum for a particular department where efficiency is characterized by lessening the risk of delayed graduation. To achieve this goal, we implemented a new optimization model, CBCB, that uses the regulations of the well-know problem of the BACP. CBCB assigns courses with relatively higher crucial values to the earliest terms while maintaining a balanced workloads of the terms. This layout outperforms other models (i.e., BACP and RBCB) by not only assigning relevant courses to the earliest terms but also minimizing the distance between them. Note that CBCB is modeled as a multi-objective optimization problem with linear objective functions which is another advantage over the RBCB model implemented using quadratic non-linear functions. As a future work, we will extend CBCB by adding a new criterion characterized by course difficulty. With this extension, we will give the student the ability to determine the level of difficulty he or she wants in every term and accordingly design a layout for the curriculum that best fits the required constraints. Now that we have an "efficient" curriculum, it would be essential next to follow the progress of the students attempting a particular curriculum. Perhaps the best way to track student progress would be to predict their academic performance in advance and accordingly give them suitable advice. To achieve this goal, we implement two predictive models

using BBNs and MNs. The results show that these networks may easily model a curriculum graph and can be used to predict the future progress of a student. It is shown that we can predict the final GPA of a student with a marginal error of 0.0449 upon receiving the grades of the first semester. This initial work will be extended in the future to model multiple variables (e.g., student initial condition, age, gender, educational level of parents, emotional factors, instructor difficulty, etc.) in the BBN and the MN models in addition to the "course" variable. We anticipate that this additional information will improve the performance of our framework. Finally we address student progress at the most basic level, by investigating the structural properties of individual curricula, taking into account the degree to which course enrollment sequences in a curriculum may impact student success. Using data mining methods we presented a framework that models student course enrollment sequences as directed acyclic graphs for further study and analysis. We introduced some measures to quantify and compare courses among different directed acyclic graphs. Based on real data, our results show the influence of course enrollment sequences on the final GPA value. These results also show how a student's time-to-degree is affected by course sequences. Students who graduate faster tend, on average, to follow a different course enrollment sequences than those who get delayed. This application is therefore very useful in tracking the progress of students and to intervene (via advisement, academic support, etc.) in order to improve graduation rates.

References

[1] Education knowledge and skills for the jobs of the future. The White House. Available at https://www.whitehouse.gov/issues/education/higher-education.

[2] Does diversity make a difference? three research studies on diversity in college classrooms. Technical report, American Council on Education (ACE) and American Association of University Professors (AAUP), Washington, DC, 2000.

[3] Performance funding for higher education. National Conference of State Legislatures, Feb. 2012. Available at http://www.ncsl.org/research/education/performance-funding.aspx.

[4] Weekly address: Congress should back plan to hire teachers. The White House, Office of the Press Secretary, Aug. 2012. Available at https://www.whitehouse.gov/the-press-office/2012/08/18/weekly-address-congress-should-back-plan-hire-teachers.

[5] Predictive analytics for student success: Developing data-driven predictive models of student success. Technical report, University of Maryland University College, Adelphi, Maryland, January 2015.

[6] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, pages 3–14, Washington, DC, USA, 1995. IEEE Computer Society.

[7] A. V. Aho, M. R. Garey, and J. D. Ullman. The transitive reduction of a directed graph. *Society for Industrial and Applied Mathematics*, 1(2), June 1972.

[8] J. Bennedsen and M. E. Caspersen. Optimists have more fun, but do they learn better? : On the influence of emotional and social factors on learning cs and math. 18(1):1–16.

[9] J. M. Braxton. Student success in college: Creating conditions that matter. *Journal of College Student Development.*, 48(5):614–616, 2007.

[10] D. F. Butcher and W. A. Muth. Predicting performance in an introductory computer science course. *Commun. ACM*, 28(3):263–268, Mar. 1985.

[11] C. Castro and S. Manzano. Variable and value ordering when solving balanced academic curriculum problems. *CoRR*, cs.PL/0110007, 2001.

[12] A. T. Chamillard. Using student performance predictions in a computer science curriculum. In *In ITICSE '06: Proceedings of the 11th Annual Conference on Innovation and Technology in Computer Science Education*, pages 260–264, 2006.

[13] M. Chiarandini, L. Di Gaspero, S. Gualandi, and A. Schaerf. The balanced academic curriculum problem revisited. *Journal of Heuristics*, pages 1–30, 2012.

[14] G. Das, K. ip Lin, H. Mannila, G. Renganathan, and P. Smyth. Rule discovery from time series. pages 16–22. AAAI Press, 1998.

[15] L. DeAngelo, R. Franke, S. Hurtado, J. Pryor, and S. Tran. Completing college: Assessing graduation rates at four year colleges. Technical report, Higher Education Research Institute, UCLA, Los Angeles, CA, Nov. 2011.

[16] R. F. Deckro and H. W. Woundenberg. Identifying factors that influence performance of non-computing majors in the business computer information systems course. 21(4):431–446.

[17] R. F. Deckro and H. W. Woundenberg. Mba admission criteria and academic success. 8(4):765–769.

[18] J. Deogun and L. Jiang. Prediction mining an approach to mining association rules for prediction. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, volume 3642 of *Lecture Notes in Computer Science*, pages 98–108. Springer Berlin Heidelberg, 2005.

[19] L. Di Gaspero and A. Schaerf. Hybrid local search techniques for the generalized balanced academic curriculum problem. In M. Blesa, C. Blum, C. Cotta, A. Fernndez, J. Gallardo, A. Roli, and M. Sampels, editors, *Hybrid Metaheuristics*, volume 5296 of *Lecture Notes in Computer Science*, pages 146–157. Springer Berlin Heidelberg, 2008.

[20] U. Faghihi, P. Fouriner-viger, R. Nkambou, and P. Poirier. The combination of a causal and emotional learning mechanism for an improved cognitive tutoring

agent. In *Trends in Applied Intelligent Systems*, volume 6097 of *Lecture Notes in Computer Science*, pages 438–449. Springer Berlin Heidelberg, 2010.

[21] H. Hamilton and K. Karimi. The timers ii algorithm for the discovery of causality. In *Advances in Knowledge Discovery and Data Mining*, volume 3518 of *Lecture Notes in Computer Science*, pages 744–750. Springer Berlin Heidelberg, 2005.

[22] D. Heckerman. A tutorial on learning with bayesian networks. Technical report, Learning in Graphical Models, 1996.

[23] Y. Hirate and H. Yamana. Generalized sequential pattern mining with item intervals. *Journal of Computers*, 1(3), 2006.

[24] B. Hnich, Z. Kiziltan, and T. Walsh. Modelling a balanced academic curriculum problem. In *Proceedings of CP-AI-OR-2002*, 2002.

[25] D. Hochbaum. Complexity and algorithms for nonlinear optimization problems. *Annals of Operations Research*, 153(1):257–296, 2007.

[26] J. P. Hoffmann. *Linear Regression Analysis: Applications and Assumptions*. Second edition edition, 2010.

[27] F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer Co., second edition edition.

[28] K. J. Keen and L. Etzkorn. Predicting students' grades in computer science courses based on complexity measures of teacher's lecture notes. *J. Comput. Sci. Coll.*, 24(5):44–48, May 2009.

[29] J. Konvalina. Identifying factors influencing computer science aptitude and achievement. *AEDS Journal*, 16(2):714–731, 1983.

[30] K. B. Korb and A. E. Nicholson. *Bayesian Artificial Intelligence*. CRC Press Co., second edition edition.

[31] G. L. Kramer. *Student academic services : an integrated approach*. Jossey-Bass, San Francisco, 2003.

[32] D. P. Kroese, T. Brereton, T. Taimre, and Z. I. Botev. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6), 2014.

[33] G. D. Kuh, J. Kinzie, J. A. Buckley, B. K. Bridges, and J. C. Hayek. What matters to student success: A review of the literature. Technical report, National Postsecondary Education Cooperative, U.S. Department of Education, Commissioned Report for the National Symposium on Postsecondary Student Success: Spearheading a Dialog on Student Success, 2006. Available at `http://nces.ed.gov/npec/pdf/kuh_team_report.pdf`.

[34] G. D. Kuh, J. Kinzie, J. H. Schuh, and E. J. Whitt. *Student Success in College: Creating Conditions That Matter*. Jossey-Bass, San Francisco, 2005b.

[35] G. D. Kuh, J. Kinzie, J. H. Schuh, and E. J. Whitt. *Student Success in College: Creating Conditions That Matter*. Jossey-Bass, San Francisco, CA, 2010.

[36] G. D. Kuh and P. D. Umbach. College and character: Insights from the national survey of student engagement. *New Directions for Institutional Research*, 2004(122):37–54, 2004.

[37] T. Lambert, C. Castro, E. Monfroy, and F. Saubion. Solving the balanced academic curriculum problem with an hybridization of genetic algorithm and constraint propagation. *Artificial Intelligence and Soft Computing–ICAISC 2006*, pages 410–419, 2006.

[38] R. Marler and J. Arora. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, 26(6):369–395, 2004.

[39] J. Monette, P. Schaus, S. Zampelli, Y. Deville, and P. Dupont. A cp approach to the balanced academic curriculum problem. In *Symcon'07, The Seventh International Workshop on Symmetry and Constraint Satisfaction Problems*, 2007.

[40] C. Moore and N. Shulock. Student progress toward degree completion: Lessons from the research literature. Technical report, The Institute for Higher Education Leadership & Policy, California State University, Sacramento, 2009.

[41] K. P. Murphy. Dynamic bayesian networks: Representation, inference and learning, 2002.

[42] Y. Z. nal and zgr Uysal. A new mixed integer programming model for curriculum balancing: Application to a turkish university. *European Journal of Operational Research*, 238(1):339 – 347, 2014.

[43] M. Newman. *Networks: An Introduction*. Oxford University Press, Oxford, NY, 2010.

[44] R. Paradise and B. Rogoff. Side by side: Learning by observing and pitching in. *Ethos*, 37(1):102–138, 2009.

[45] E. T. Pascarella and P. T. Terenzini. *How College Affects Students: A Third Decade of Research*. Jossey-Bass, San Francisco, 2005.

[46] C. N. G. Peter Spirtes and R. Scheines. *Causation, Prediction, and Search*. MIT Press, second edition edition.

[47] S. Roth. *High-order Markov Random Fields for Low-level Vision*. Ph.d. dissertation, Brown University,Department of Computer Science, 2007.

[48] C. L. Ryan and K. Bauman. Educational attainment in the united states: 2015. Technical report, United States Census Bureau, Suitland, Maryland, March 2016.

[49] A. Schaerf. A survey of automated timetabling. *Artificial intelligence review*, 13(2):87–127, 1999.

[50] U. Schmidt. Learning and evaluating markov random fields for natural images. Masters thesis, Technische Universitt Darmstadt,Department of Computer Science, February 2010.

[51] A. Slim, J. Kozlick, G. L. Heileman, and C. T. Abdallah. The complexity of university curricula according to course cruciality. In *Proceedings of the 8th International Conference on Complex, Intelligent, and Software Intensive Systems*, Birmingham City University, Birmingham, UK, 2014. IEEE.

[52] A. Slim, J. Kozlick, G. L. Heileman, J. Wigdahl, and C. T. Abdallah. Network analysis of university courses. In *Proceedings of the 6th Annual Workshop on Simplifying Complex Networks for Practitioners*, Seoul, Korea, 2014. ACM.

[53] W. Stadler. Fundamentals of multicriteria optimization. In W. Stadler, editor, *Multicriteria Optimization in Engineering and in the Sciences*, volume 37 of *Mathematical Concepts and Methods in Science and Engineering*, pages 1–25. Springer US, 1988.

[54] H. Steck. Constraint-based structural learning in bayesian networks using finite data sets.

[55] C. Strange and J. H. Banning. *Educating by design: creating campus learning environments that work*. Jossey-Bass, San Francisco, 2001.

[56] S.-M. R. Ting and T. L. Robinson. First-year academic success: A prediction combining cognitive and psychosocial variables for caucasian and african american students.

[57] V. Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 1(45):89–125, 1975.

[58] V. Tinto. *Leaving College: Rethinking the Causes and Cures of Student Attrition*. University of Chicago Press, Chicago, IL, 1987.

[59] A. Venezia, P. M. Callan, J. E. Finney, M. W. Kirst, and M. D. Usdan. The governance divide: A report on a four-state study on improving college readiness and success. Technical report, The National Center for Public Policy and Higher Education, San Jose, CA, Sep. 2005.

[60] J. Wigdahl, G. L. Heileman, A. Slim, and C. T. Abdallah. Curricular efficiency: What role does it play in student success? In *Proceedings of the the 121st ASEE Annual Conference and Exposition*, Indianapolis, Indiana, USA, 2014. IEEE.

[61] D.-L. Yang, Y.-L. Hsieh, and J. Wu. Using data mining to study upstream and downstream causal relationship in stock market. In *JCIS*. Atlantis Press, 2006. Available at `http://www.bibsonomy.org/bibtex/2cc44647df3f74088ef48fe3b08d599bc/dblp`.

[62] L. Zhang. Does state funding affect graduation rates at public four-year colleges and universities? *Educational Policy*, 23(5):714–731, Sep. 2009.