Restructuring Curricular Patterns Using Bayesian Networks

Ahmad Slim
Lebanese American University
1102, Beirut, Lebanon
ahmad.slim@lau.edu.lb

Gregory L. Heileman
The University of Arizona
Tucson, AZ 85721, U.S.
heileman@arizona.edu

Chaouki T. Abdallah
Georgia Tech
Atlanta, GA 30332, U.S.
ctabdallah@gatech.edu

Ameer Slim
The University of New Mexico
Albuquerque, NM 87131, U.S.
ahs1993@unm.edu

Najem Sirhan
The University of New Mexico
Albuquerque, NM 87131, U.S.
najem83@unm.edu

ABSTRACT

Recent studies proved the existence of a relationship between the complexity of university curricula and graduation rates. As a result, extensive efforts have been done in an attempt to restructure curricula in order to improve graduation rates. In this paper, we propose a new model for evaluating and quantifying the impact of restructuring curricula on graduation rates using a Bayesian network framework. We validate our model by analyzing a common curricular pattern found in most of the engineering programs. We demonstrate its usefulness using actual data for students at the University of New Mexico. We also extend this model to include a helpful tool that can be used to predict student performance. The advantage of our work is characterized by its data-driven nature which makes it more reliable than other proposed models.

Keywords

Curricular analytics, Bayesian networks, education, curriculum complexity, student success, graduation rate

1. INTRODUCTION

Recently a significant amount of work has been done on curricular analytics to show its impact on student success [16, 15, 12, 13, 2, 14]. The work done mainly spots the light on the importance of curricula structure on student performance characterized primarily by graduation and retention rates. These studies argue that the complexity of prerequisite dependencies between the requirements of a curriculum can increase the risk of students stopping out and eventually dropping the school. A lot of work has been done recently identifying factors that help students retain their school and hence graduate at faster rates [6]. This includes new learning pedagogy styles, dorms, flipped classrooms, learning centers, etc. However, such factors solely might fail to significantly contribute to student success if other institutional factors are overlooked; essentially curricula structure [11]. As mentioned earlier, it is already proved that the structure of a curriculum has a direct impact in student success [2, 3]. In this regard, Klingbeil and Bourne considered a case study and analyzed the structure of a common curricular pattern found in most of the engineering programs (Figure 1) [5]. They noticed that, in the sophomore year, most of the engineering programs require Differential Equations as a prerequisite to a domain specific course (in electrical engineering programs, Circuits I is domain specific; in mechanical engineering, Mechanics (statics and dynamics) is domain specific; etc.). They also noticed that all the learning outcomes in Differential Equations, except for solving linear differential equations, are not necessarily required to pass the domain specific course. Thus they suggested to create a new course in the freshman year to teach how to solve linear differential equations along with the Precalculus materials. They called this new course Engineering 101. As a result of this observation, they pointed out that Differential Equations is not required anymore as a prerequisite for the domain central course; only Engineering 101 does. This resulted in a revised curricular pattern shown in Figure 2. Klingbeil and Bourne claim that this new curricular pattern will help students graduate in a timely fashion. This is driven by the fact that the students are not required anymore to follow the long chain of prerequisites before they are allowed to take a domain specific course-as it is the case in the original curricular pattern. This new pattern is now pursued by a number of universities [5]. And to validate the legitimacy of such changes to curricular patterns, a number of researchers came up with different mathematical models that prove the significance of these changes on student success outcomes. Slim et al. came up with a metric that quantifies the complexity of any curricular pattern [16]. Their metric showed that the revised engineering pattern shown in Figure 2 is less complex than that of the original one shown in Figure 1[2]. Thus, according to their metric, students are expected to finish their degrees at a faster pace. Furthermore, Hickeman showed that the revised engineering pattern can significantly improve graduation rates [3]. He proved that by implementing a Monte Carlo simulation through which virtual students are allowed to flow through a curricular pattern. For more details on different models see [2]. Although these models put a mathematical foundation to prove the advantage of restructuring curricula, they still have a major limitation. None of these models include actual student data in their implementations. In other words, these models only show the advantage of restructuring curricular patterns in its abstract state without using any data-driven approaches. This makes the models less reliable in proving the need to revise any curricular pattern. In this paper we propose a data-driven model that uses actual student data to achieve this. Particularly we use a Bayesian Network (BN) model to statistically prove the validity of any effort to restructure

a curricular pattern. This is mainly characterized by adding and removing courses/prerequisites within a curricular pattern which can be neatly captured using a BN. The main motivation behind our proposed model is the ability to use the notion of hidden/latent variables in building a BN. These hidden nodes represent new added courses in the revised curricular pattern. Thus they can be used to check the validity of the changes made to the original curricular pattern and accordingly decide whether to apply these changes or not. It is important to note that our model can be generalized to find the optimal structure of a revised curricular pattern using different methods of structural learning for a BN [1]. However we will not cover this part in our paper. We will leave it for a future work. The remainder of this paper is structured as follows. In Section II we present the details of our proposed framework and provide a case study. In Section III we present a number of applications for our proposed model and show some simulation results. Finally, Section IV presents some concluding remarks.

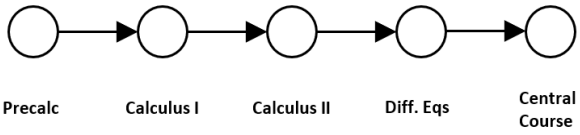

Figure 1: The original curricular pattern found in most engineering programs. The nodes represent the courses and the directed edges represent the prerequisite dependencies.

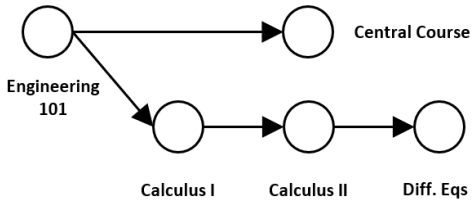

Figure 2: The revised curricular pattern.

2. BAYESIAN NETWORKS AND HIDDEN VARIABLES

A BN is a directed acyclic graph (DAG) representing correlations between a number of random variables [4]. The graph-like structure reflects a confined representation of the joint probability distribution underlying these variables. The presence of an edge between two nodes indicates the existence of a relationship between two variables and the direction of the edge indicates the direction of the causal relationship. That is a directed edge from node A to node B indicates that A causes B. In BNs, these types of relationships are quantified by conditional probability tables (CPTs). The main feature of a BN is its compact representation of the conditional dependencies of the random variables. However, in some applications the BN gets complicated and thus in this case adding hidden nodes would be essential for two main reasons [8]:

1. **Knowledge discovery**: reveals interesting relationships among the variables of the data
2. **Lower complexity**: attains lower structure complexity of the network

For example, consider the case where we observe a bunch of variables representing different patient's symptoms. The joint probability distribution for these symptoms might be highly connected. In this case, the BN representation of these symptoms would be highly

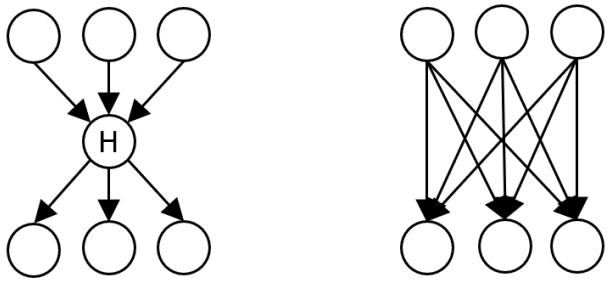

Figure 3: Two DAGs (one with a hidden node and the other without) representing the relationship between symptoms, causes and mediating factors. Symptoms, such as chest pain, are represented by the leaves. Causes, such as smoking and diet, are represented by the roots. Mediating factors, such as heart disease, are represented by the hidden nodes. This figure shows how hidden nodes can reveal a better understanding of the relationship between variables by attaining a lower structure complexity of the network.

complicated. However if we introduce a "cause" node representing the underlying disease for these symptoms then we can get a noticeably simpler network. We call this cause node a hidden node. Figure 3 (inspired from [7]) illustrates this example in more details. In a similar scenario in an educational context, a curricular pattern can be modeled as a BN. A node maybe a course, and the states of the node would be the possible letter grades (i.e., $A, B+, B, C+$, etc.). A directed edge from course A to course B indicates that the performance in A influences that in B. In this context adding a hidden node to the BN is equivalent to adding a new course to a curricular pattern. This hidden node might represent the underlying prerequisite course that needs to be taken prior taking other courses. This notion constitutes the bulk of our proposed work and the subsequent sections elaborate more about this idea. Following this notion, restructuring the BN of any curricular pattern would include these steps:

- Removing an existing course(s) (i.e., removing a node)
- Adding a new course(s) (i.e., adding a hidden node)
- Removing an existing prerequisite(s) (i.e., removing a directed edge)
- Adding a new prerequisite(s) (i.e., adding a directed edge)

We denote the restructured BN by R, characterizing the revised curricular pattern. Once R is constructed, we fit the CPTs using actual student data, denoted by D, and then compute the likelihood of R, $p(D/R)$. Similarly, we denote by O the BN of the original curricular pattern. Once it is constructed, we compute its likelihood, $p(D/O)$. If $p(D/R)$ is greater than $p(D/O)$, then the revised version of the curricular pattern fits the student data better than that of the original one. In this case, the proposed revised curricular pattern can be a good candidate to replace the original one. To elaborate more about this, we present a case study in the following section.

2.1 Engineering Curricular Pattern: A Case Study

In this section, we consider, as a case study, the engineering curricular patterns shown in Figure 1 and Figure 2. Recall that the graph in Figure 1 represents the original pattern whereas that in Figure 2 represents the revised one. For these two graphs, we construct two BNs denoted by O and R respectively. These two BNs are shown in

Course Number	Course Name
MATH 150	Precalculus
MATH 162	Calculus I
MATH 163	Calculus II
MATH 264	Calculus III
MATH 316	Differential Equations
PHYC 160	General Physics I
PHYC 161	General Physics II
ECE 203	Circuits I
ENG 101	New Course Proposed

Table 1: Engineering courses taught at freshman and sophomore level at UNM.

Figure 4 and Figure 5. The variables used to construct O and R are shown in Table 1. These variables represent actual courses taught at the University of New Mexico (UNM). The states of each of these variables are the letter grades: A, B, C and D/F where D and F are assumed to represent one state. The CPTs for both O and R are computed using a dataset, denoted by D, for $1,000$ UNM student. Each row in this dataset contains the letter grades achieved in the courses shown in Table 1. Some grades for some students were not available. Thus, the dataset included missing values. In addition, ENG 101 in Figure 5 is considered a hidden variable because it is supposed to represent the new proposed course and thus we do not have the letter grade values. Therefore, to compensate for hidden and missing values in our dataset, we used the expectation-maximization (EM) method to compute the CPTs for O and R [9]. Using EM, we computed the ratio of the likelihood, $\frac{p(D/R)}{p(D/O)}$, to be 2.89. This means that R fits the student data better than O. This result suggests that the proposed revised curricular pattern is a good candidate to replace the original one. Not only does it have a less complex structure but also the revised curricular pattern has the potential to improve student performance when compared to the original pattern. We concluded this using actual student data which is an advantage over other proposed models in literature [2]. In the following sections we present more applications of BNs in the context of course network.

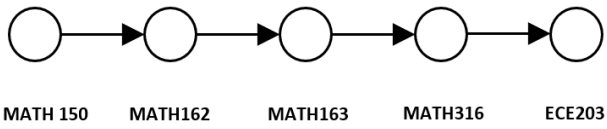

Figure 4: The original curricular pattern found in the electrical engineering curriculum.

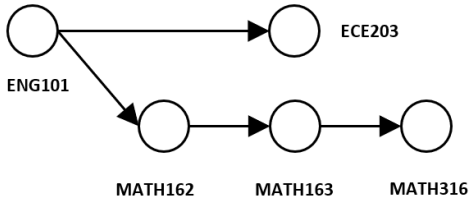

Figure 5: The revised version of the curricular pattern in the electrical engineering curriculum.

3. INFLUENCE OF STUDENT CHARACTERISTICS ON ACADEMIC PERFORMANCE

As mentioned earlier, many institutions are dedicating lot of efforts on student success. Colleges and universities are applying ever more sophisticated analytical tools to track their student progress in an attempt to improve their performance characterized mainly by graduation and retention rates. Intuitively, early indicators of student performance in this context is crucial to provide suitable interventions when needed. Thus predicting the performance of students in future courses is essential to achieve it. In this regard, historical information about previous academic achievement of a student could be used to project future performance. For example, a student who receives a 'B' in Calculus I is expected to receive a better grade in Calculus II compared to those who receive a 'D'. A BN in the context of course network can capture the correlation in performance between such courses. Further, it can be used to predict the letter grades of a student in subsequent classes based on the grades of previous classes. The accuracy of prediction can further be improved by adding other factors related to student characteristics. Factors such as age, gender, high school GPA, socioeconomic status, etc. proved to influence student performance [11]. For this reason, it would be rational to add such factors as additional variables to the BN of a course network. The advantage of using a BN model to capture all these variables together is two-folded: it can be used as a knowledge discovery model that can neatly display the correlation between the variables and also it can be used as an inference tool to predict student performance. In this section, we construct a BN for eight engineering courses along with five other variables representing student characteristics. The eight courses are: MATH 150, MATH 162, MATH 163, MATH 264, MATH 316, PHYC 160, PHYC 161 and ECE 203 (Table 1). These courses are considered to be the most crucial classes at the freshman level. As for student characteristics, we considered Gender, ACT score, and high school GPA. As mentioned earlier the student characteristics are proved to influence student performance. Thus it would be interesting to discover and visualize how all these variables are related to each other. In the following section we show the constructed BN along with some applications. However, it is important to mention here that a similar work has been done in [10]. The authors of this work used a domain expert to construct the BN for these variables. This means that the process of constructing the BN is not automated and doesn't guarantee a good fit to the student data. In this paper, however, we automated the process of learning the structure of the BN using a score-based learning algorithm [1]. Particularly, we used the hill climbing (hc) greedy search that explores the space of the DAGs by single-arc addition, removal and reversals.

3.1 A BN for Engineering Courses

To construct and validate our framework, we collected a dataset of 3,000 undergraduate student in the college of engineering at UNM. The dataset represented different demographical and academical variables of the students. The states for each of these variables are show in Table 2. It is important to note that MATH 162 is the prerequisite for MATH 163, MATH 163 is the prerequisite for MATH 316 and MATH 264, and MATH 316 is the prerequisite for ECE 203. The constructed BN is presented in the graph shown in Figure 6. It is tempting to interpret this graph in terms of causality. In particular, it seems that ACT score, high school GPA and gender, in contrast to ethnicity, causally influence the performance of students in these engineering courses. Also, this graph shows that the performance in PHYC 160 influences that in MATH 264 and ECE 203. This is an interesting observation because, according to the department policy, PHYC 160 is not a required prerequisite for neither of these courses. Though it has an impact on both these courses. Another interesting observation is the absence of any correlation in performance between MATH 316 and ECE 203 even though MATH 316 is a prerequisite to ECE 203. This observation confirms the fact that

Variable	States
Gender	Female, Male
Ethnicity	7 different ethnicities
ACT score	Integer between 10 and 36
High school GPA	Real value between 1 and 4
MATH 150,162,163,264,316	A,B,C, D/F
PHYC 160,161	A,B,C, D/F
ECE 203	A,B,C, D/F

Table 2: The courses and the student characteristics used to build the BN.

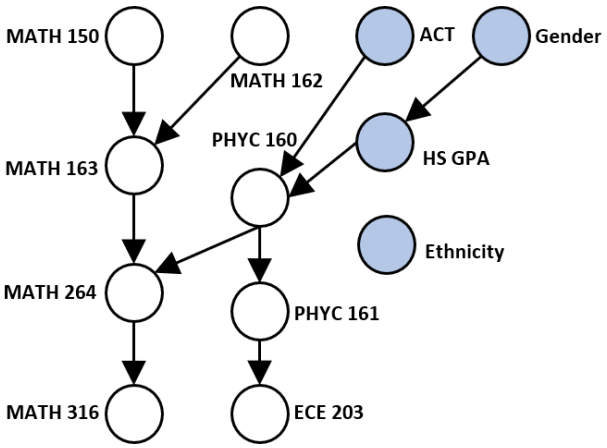

Figure 6: The constructed BN using UNM student data.

only a small portion of the learning outcomes in MATH 316, namely the ability to solve linear differential equations, are actually used in ECE 203 and the absence of a link between these two courses proves it. This is another evidence that supports the claim that it is needed to restructure the original curricular pattern shown in Figure 1.

4. CONCLUSION

In this paper we presented a framework that models a curriculum as a Bayesian Network (BN). We showed that this model, in the context of a course network, can be used to quantify any effort to restructure curricular patterns. In particular, we used the notion of hidden variables to achieve this. We validated our proposed model using a common curricular pattern found in most of the engineering programs. For that, we used actual data for students at the University of New Mexico. The results showed that the likelihood of the revised version of the engineering curricular pattern is higher than that of the original one. This suggests that the revised version can help students perform better in their courses as well as graduate at a faster pace. The advantage of our model over other proposed models in literature is its data-driven nature which makes it more reliable. Furthermore, we extended our model to use it as a knowledge discovery and inference tool. Particularly we added variables related to student characteristics (e.g. gender, ACT score, high school GPA, etc.) and showed how they can influence student performance. We also showed how to exploit the constructed BN to predict the grades of a student in following semesters based on grades of previous semesters.

5. REFERENCES

[1] D. Heckerman. Learning in graphical models. chapter A Tutorial on Learning with Bayesian Networks, pages 301–354. MIT Press, Cambridge, MA, USA, 1999.

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

[3] M. Hickman. Development of a Curriculum Analysis and Simulation Library with Applications in Curricular Analytics. Master's thesis, The University of New Mexico, 2017.

[4] M. Horny. Bayesian networks. Technical report, Boston University School of Public Health, 2014.

[5] N. W. Klingbeil and A. Bourne. The wright state model for engineering mathematics education: Longitudinal impact on initially underprepared students. In *2015 ASEE Annual Conference & Exposition*, Seattle, Washington, June 2015. ASEE Conferences.

[6] L. K. Lau. Institutional factors affecting student retention. 2003.

[7] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[8] T. L. Perez and L. Kaelbling. Techniques in artificial intelligence (sma 5504). Massachusetts Institute of Technology, MIT OpenCourseWare, Fall 2002.

[9] D. Prescher. A tutorial on the expectation maximization algorithm including maximum likelihood estimation and em training of probabilistic context free grammars. *ArXiv*, 2004.

[10] A. Sharabiani, F. Karim, A. Sharabiani, M. Atanasov, and H. Darabi. An enhanced bayesian network model for prediction of students' academic performance in engineering programs. *2014 IEEE Global Engineering Education Conference (EDUCON)*, pages 832–837, 2014.

[11] A. Slim. *Curricular Analytics in Higher Education*. PhD thesis, The University of New Mexico, 2016.

[12] A. Slim, G. L. Heileman, W. Al-Doroubi, and C. T. Abdallah. The impact of course enrollment sequences on student success. In *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, pages 59–65, March 2016.

[13] A. Slim, G. L. Heileman, M. Hickman, and C. T. Abdallah. A geometric distributed probabilistic model to predict graduation rates. In *2017 IEEE Cloud Big Data Computing (CBDCom)*, pages 1–8, Aug 2017.

[14] A. Slim, D. Hush, T. Ojha, , C. T. Abdallah, G. L. Heileman, and G. El-Howayek. An automated framework to recommend a suitable academic program, course and instructor. In *Proceedings of the Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, San Francisco, CA, USA, 2019. IEEE.

[15] 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.

[16] 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.