ASEE's 123rd Annual • Conference & Exposition • New Orleans, LA • June 26-29, 2016

Paper ID #16257

Identifying Course Trajectories of High Achieving Engineering Students through Data Analytics

Omaima Almatrafi, George Mason University
Dr. Aditya Johri, George Mason University

Aditya Johri is Associate Professor in the Information Sciences & Technology Department. Dr. Johri studies the use of information and communication technologies (ICT) for learning and knowledge sharing, with a focus on cognition in informal environments. He also examine the role of ICT in supporting distributed work among globally dispersed workers and in furthering social development in emerging economies. He received the U.S. National Science Foundation's Early Career Award in 2009. He is co-editor of the Cambridge Handbook of Engineering Education Research (CHEER) published by Cambridge University Press, New York, NY. Dr. Johri earned his Ph.D. in Learning Sciences and Technology Design at Stanford University and a B.Eng. in Mechanical Engineering at Delhi College of Engineering.

Huzefa Rangwala, George Mason University
Dr. Jaime Lester, George Mason University

Identifying Course Trajectories of High Achieving Engineering Students through Data Analytics

Abstract

In this paper we present findings from a study that compares course trajectories of students who performed well academically and graduated in four years and with those of low achieving student. The goal of this research is to identify factors related to course-taking choices and degree planning that can affect students' academic performance. The data for the study was collected from three majors within an engineering school at a large public university: civil, environmental, and infrastructure engineering (CEIE), computer science (CS), and information technology (INFT). The data includes more than 13,500 records of 360 students. Analysis shows that low performers postponed some courses until the latter end of their program, which delayed consequence courses and their graduation. We also found that low performers enrolled in multiple courses together at the same semester that their counterparts do not usually take concurrently. The methods used in this paper, frequent pattern mining and visualization, help uncover student pathways and trajectories with direct impact for advising prospective and current students. The findings can also be used to improve engineering programs' curriculum.

1. Introduction

As higher education institutions continually make investments to improve the quality of their academic programs, it has become increasingly important to develop a better understanding of factors that shape students' success. Thus, researchers examine the impact of demographic, socio-economic, and environmental factors, on student success[1]. Within STEM education, and especially within engineering education, additional factors such as identity and motivation contribute to student success[2]. Other factors such as conceptual understanding, misconceptions, and representational expertise are also being investigated[3].

As different forms of data about student performance are made available, it is possible to examine the factors that were previously not considered. Many higher education administrators are using data-driven decision making to identify strategies that can improve retention and quality of education[4]. While prior work has demonstrated a connection between these factors and student success, many relationships have gone unexamined primarily due to the lack of sufficient and useful data. It is well-known that students' grade point average (GPA) is the strongest predictor for students' retention; and specific courses in the first year may influence students to dropout, migrate or positively shape their career path[5], however, course taking pattern is still an area that needs more research. This work highlights that student success – their performance and degree completion time – can depend on the courses students take and the sequence in which they take these courses. Although students often enter an academic program with similar academic qualifications, once they are in the program they perform differently. Many factors can lead to differences in performance including selecting courses that limit better performance in future semesters due to a lack of proper foundation. Therefore, a better understanding of course patterns for high GPA students can shed light on more productive trajectories and pathways. A comparison with students who do not perform well can shed further light on this issue.

The objective of this study is to identify the academic path i.e., sequence of courses that high- and low- achieving students follow and find out how they differ from each other. An effective

way to address this problem is through data analytics. In this paper, a well-known data mining technique, apriori-based frequent pattern mining, has been utilized to find the frequent courses in each semester. Additionally, we used network graphs to visualize the relationship between courses that are co-enrolled in a semester and courses that are enrolled across two consecutive semesters. The visualization is a powerful way to view the most significant and interesting patterns and confirms the results from the algorithm. The results of the study is for departments who design and implement policies for programs' curricula that can improve students' academic performance and graduation rates. In addition, the result can be used for advising current and prospective students about the academic paths that can play a role in improving their learning outcomes and performance. Studying engineering programs trajectories of course-taking paths of high-performers and comparing it with low performers can reveal some insights about course taking choices that is helpful for advisors and students.

This paper is organized as follows: Section 2 presents relevant work on educational data mining and frequent pattern mining. Section 3 provides a brief description of the algorithm that has been utilized in the study followed by the research study and analysis in section 4. Section 6 concludes the paper with some insights and future work.

2. Literature Review

2.1 Related Work in Educational Domain

Educational data mining is a raising research area that is concerned with developing techniques to analyze student associate data collected from a myriad of sources to answer educational related questions[6]. There are many publications on analyzing students' data to predict their performance; the majority use classification methods to study the background qualifications such as language, mother's education, and family income as primary factors to predict the academic performance[1]. They found that some of these factors are highly correlated with the students' performance. Others predict course/term grade and students at risk, from last year performance[7, 8].

In 2014, Peña-Ayala surveyed the recent work in educational data mining and found that more than half of the approaches in the current educational data mining research, for the period between 2010 and the first quarter of 2013, focus on student modeling, which leave other functionalities such as assessment, student behavior modeling, student support and feedback, and curriculum underdeveloped. About 8.5% of the current work in the survey focused on curriculum, domain knowledge, sequencing and teacher support, and 6.6% used association rule and pattern mining as an approach for their analysis[9].

Pattern mining and association rules have been utilized in educational context to achieve different goals. Su et al. used sequential pattern mining as part of Learning Portfolio Mining approach that extracts learning features, constructs a personalized activity tree and sequential rules for learners[10]. They test the approach on 45 high school students in Taiwan. The result shows that the approach was useful; and students who received personalized guidance achieved better performance. Another publication applied pattern-mining technique and association rules mining for students' profiling[11]. Romero et al. also implement apriori algorithms to discover infrequent association rules that reveal the relation between learners' behavior when using a LMS and their final grades[12]. Damaševičius applied association rule mining to assess student

academic results in a course, rank course topics based on their importance to the final course grade, and extract recommendations for course content improvement[13].

None of the research to our knowledge, studied course taking patterns in order to improve students' performance, which is the gap we are investigating in this paper. As mentioned before specific courses in the first year, and students' performance influence students to dropout[5]; thus, the set of courses and the sequence of taking them can play a role in students' academic success.

2.2 Related Work in Frequent Pattern mining

Frequent pattern mining is a thriving field in data mining. It was first proposed by Agrawal et al. in 1993 when they sought to find an association rule for market basket analysis[14]. Since then, many researchers have followed-up and extended the algorithm and its application. One of the most common frameworks for frequent pattern mining is the support based framework[15]. The concept of a frequent pattern is when a subset of the dataset appears frequently no less than a predefined threshold. There are different methodologies that have been developed to find the frequent patterns from a set of transactions such as Apriori, and FP-growth. In the next subsection, Apriori algorithm is explained in more details since it is the one used in this study.

The biggest challenge for frequent pattern mining is the enormous search space, which increase exponentially with the number of items appearing in the dataset[16]. Although Apriori algorithm was developed to be efficient and scalable- it reduces the size of candidates substantially- it will suffer from other problem if the minimum support is low or in case of long patterns or large number of frequent patterns[17]. For the purpose of this study, Apriori works well because the dataset size is manageable and the number of frequent patterns is not large per semester.

3. Apriori Algorithm

Apriori algorithm is a common method for pattern mining. Mathematically, assume we have a database of T transactions, and each transaction has i items, a set of items I called a pattern if it occurs at least p number of times where p is the minimum support. In an educational context, a transaction will be a set of courses a student takes per semester, and courses are the items. So a set of courses is considered frequent if there is more than the pre-specified number (minimum support) of students having the same courses in the same semester. The algorithm pseudo code shown below shows the technical steps to identify the frequent itemsets[18]:

```
1: p = min_support                          (pre-specified by the user)
2: L1 = { i | i ∈ I ∧ count(i) ≥ p}         (frequent items of size 1)
3: for (k=2; Lk-1 != ∅ ; k++)
4:         Ck = generate candidates         (LkXLk, then eleminate any (k-1)-size itemset < p)
5:         for each t in T:
6:                 Ct = subset(Ck, t)       (identify all candidates belong to t)
7:                 for each c in Ct:
8:                         increment count(c)
9:                 end for
10:        end for
11:        Lk = { c | c ∈ Ct ∧ count(c) ≥ p}   (extract frequent itemsets of size k)
12: end for
13: return result = ∪ Lk
```

4. Research Study and Analysis

4.1 Datasets

The data used in this paper is a structured data collected from a large public university. It includes students, courses and degrees data for students starting from Summer 2009 through Spring 2014. In this study, we utilized a subset of the data focusing on engineering majors, in particular, students matriculated at one of the three engineering majors: civil, environmental, and infrastructure engineering (CEIE), computer science (CS) or information technology (INFT). We accumulated students from all the years (2009-2014) and excluded students who switched majors. The total number of records is 13,950 compiles 630 students, and 637 courses. We did not include transfer students in this study because of the variation of time spent in the program after transferring, and the individual differences in their preparation for the majors. Hence, we eliminate any disturbance in the trajectories of the course taking patterns.

To do the analysis, we split each group of students into high achieving group (H), who have 3.0 GPA or higher in the semester under study, and low achieving group (L), who have GPA less than 3.0. Both groups were comparable in size in all majors. Table 1 summarizes the data used in the study:

Table 1: data summary

Major	Group	Number of students	Number of courses
CS	High performing (GPA>=3.0)	208	330
	Low performing (GPA<3.0)	199	295
INFT	High performing (GPA>=3.0)	73	224
	Low performing (GPA<3.0)	53	183
CEIE	High performing (GPA>=3.0)	115	208
	Low performing (GPA<3.0)	131	202

4.2 Study Description

For each group, we extracted the courses that students are enrolled in, at each semester and identify the relationship between the courses (Phase 1). After that, we extracted the relationship between the courses taking at semester s_t and s_{t+1}, where t is the semester term (Phase 2). The data that has been extracted was represented in matrices, which in turn used to draw a network of frequent courses.

Next, apriori algorithm has been applied to the list of transactions after they have been divided to H and L groups of transactions. A transaction is the list of courses a given student enrolled-in in a given semester. An example of a transaction is t_1 = {CS101, ENGL100, HIST100, MATH105}, where CS101, etc, are items, whereas {CS101, MATH105}, for instance, is a pattern if it appears in the dataset more than the minimum support. The minimum support value, user specified, is the proportion of the transaction that contains the itemset. The min_support chosen for this study is 0.25, which means if fourth of the transactions contains the itemset, it considered frequent. The min_support is not in term of count but in percentage to account for the unequal number of transactions in different groups and majors.

4.3 Evaluation Metrics

Frequent pattern mining task is the first and essential step in the association rule mining. There are some methods to measure the interestingness in the association rules, but in this study we have not applied the association rule mining; we only revealed the frequent patterns. Thus, making sense of the mined patterns and the interpretation left to the domain experts. To make the interpretation easier to read, we use visualization tool to plot the courses sequential pattern for each group. This way we facilitate the task of domain experts to understand and compare the trajectories of courses for both groups and uncover the insights. The graphs are shown in the next section.

4.4 Experimental Results

After applying apriori algorithm, we get all the frequent patterns for each term in the following form: [['MATH213'], ['CS262'], ['CS310'], ['CS262', 'CS310']]

To visualize the relationship between the courses, we obtain the nodes (courses) and edges (links between the courses) properties from the matrices. We have two types of edges: undirected edges, which are between nodes of Phase 1, and directed edges between nodes in two consecutive semesters to represent the transition from term t to t+1.

5. Discussion and Findings

The following figures show the frequent courses pattern for both groups of students. Some visualization techniques have been utilized to ease the reading of the graph, in particular: (i) the size of the node represents the percentage of students taking that class, (ii) the nodes are labeled by the name of the class, (iii) the thickness of the edge represents the percentage of students taking both of the connected classes, (iv) the courses taken at the same semester are aligned vertically and coded by the same color, and sorted horizontally by semester- far left is the first semester and it goes on up to the eighth semester in most cases. It is important to mention that nodes in the graph are frequent (meet or exceed the minimum support), however we loose that for the edges to include all links between courses if the percentage of students taking both connected courses is 20% or more.

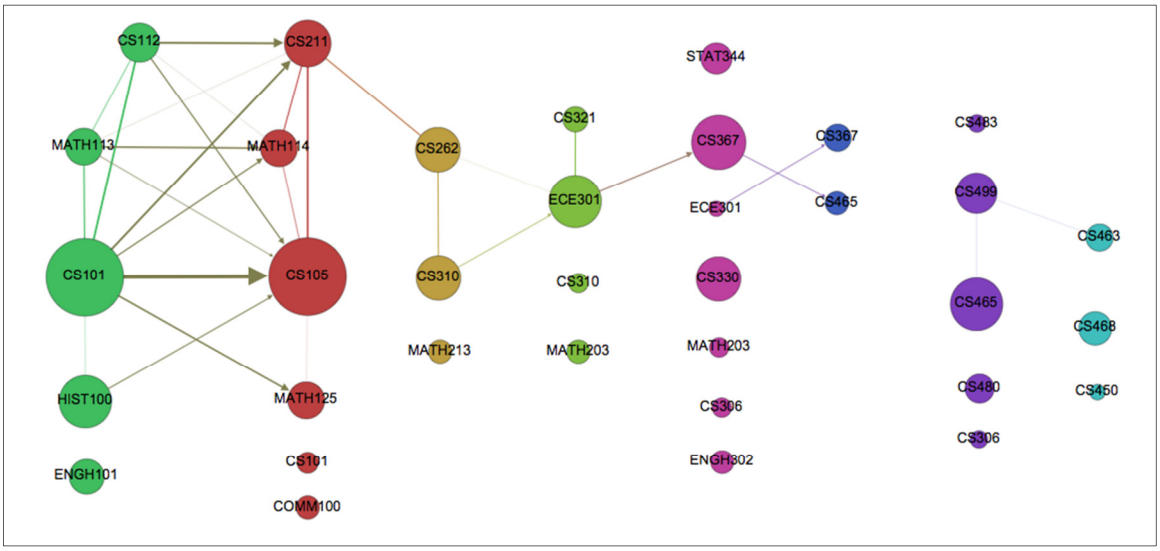

Figure 1: Trajectory of frequent courses for high achieving students (CS)

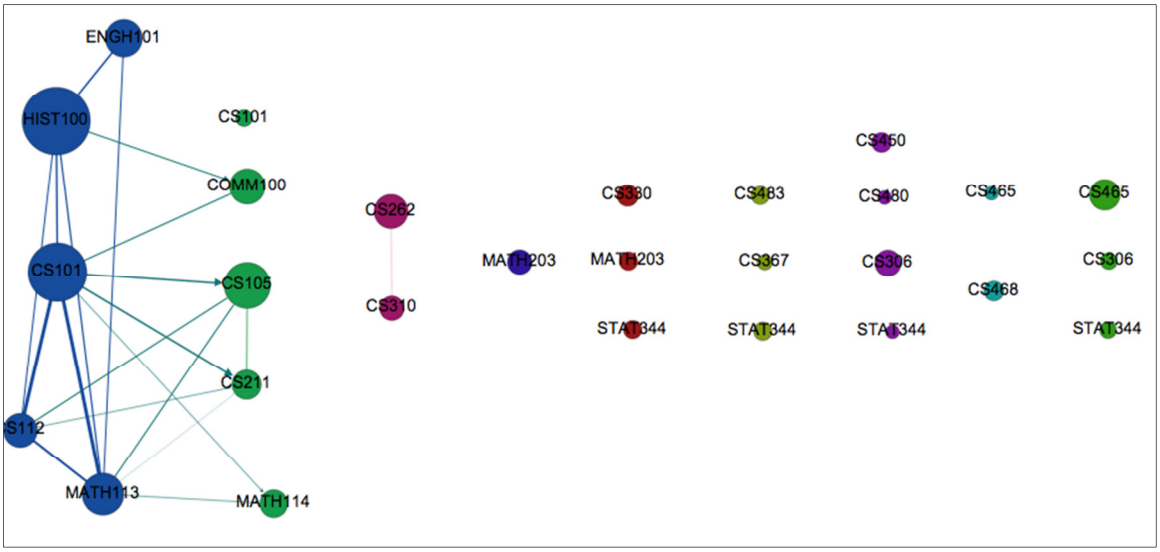

Figure 2: Trajectory of frequent courses for low achieving students (CS)

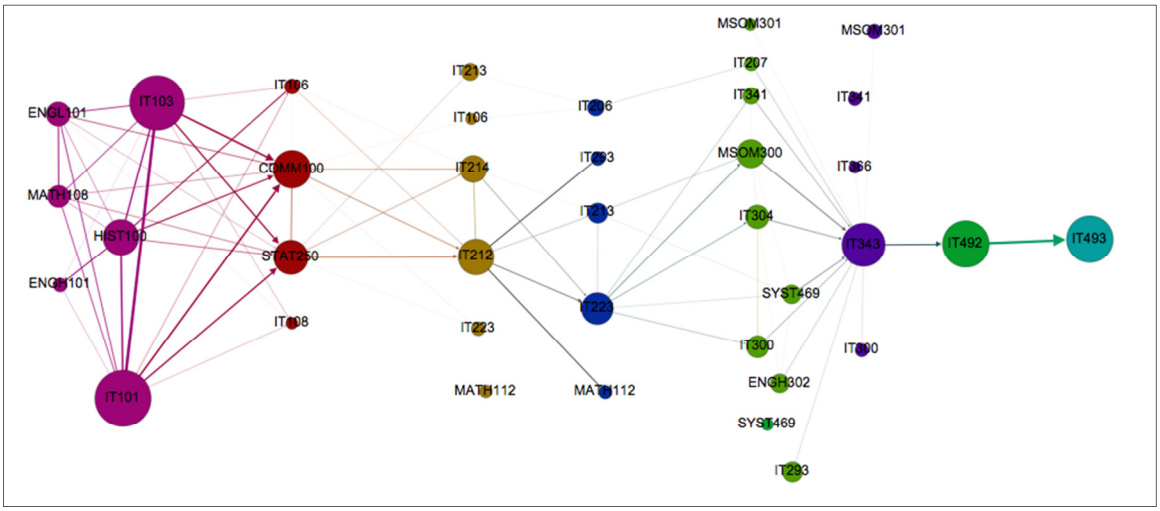

Figure 3: Trajectory of frequent courses for high achieving students (INFT)

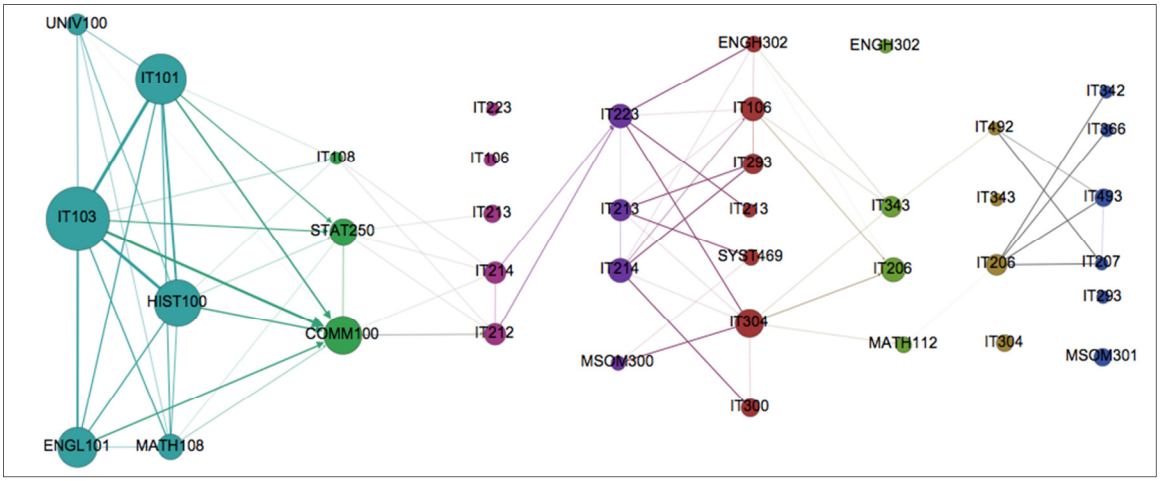

Figure 4: Trajectory of frequent courses for low achieving students (INFT)

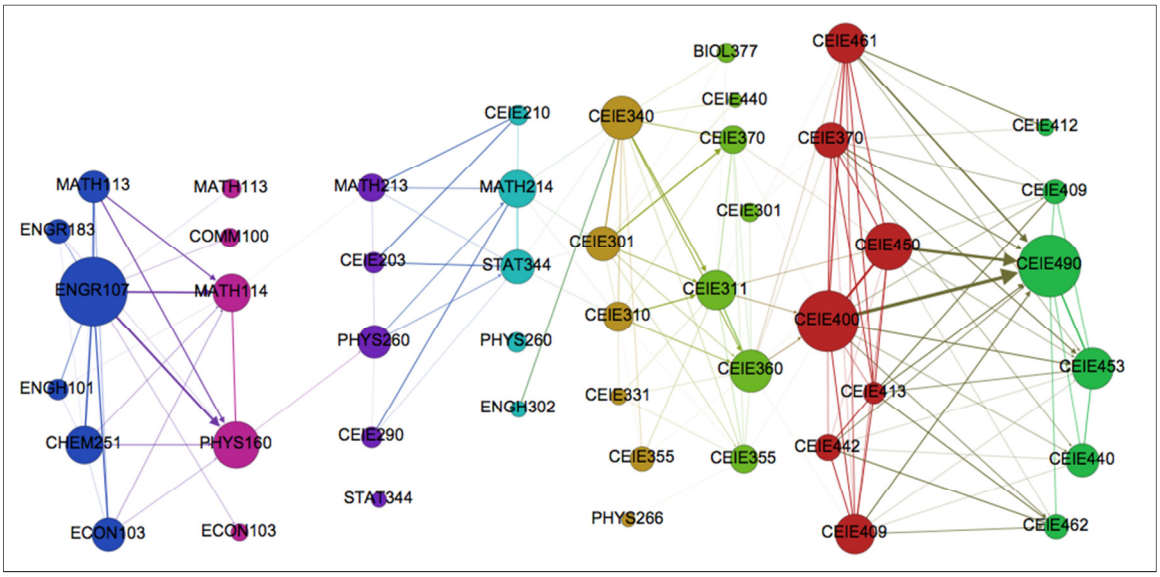

Figure 5: Trajectory of frequent courses for high achieving students (CEIE)

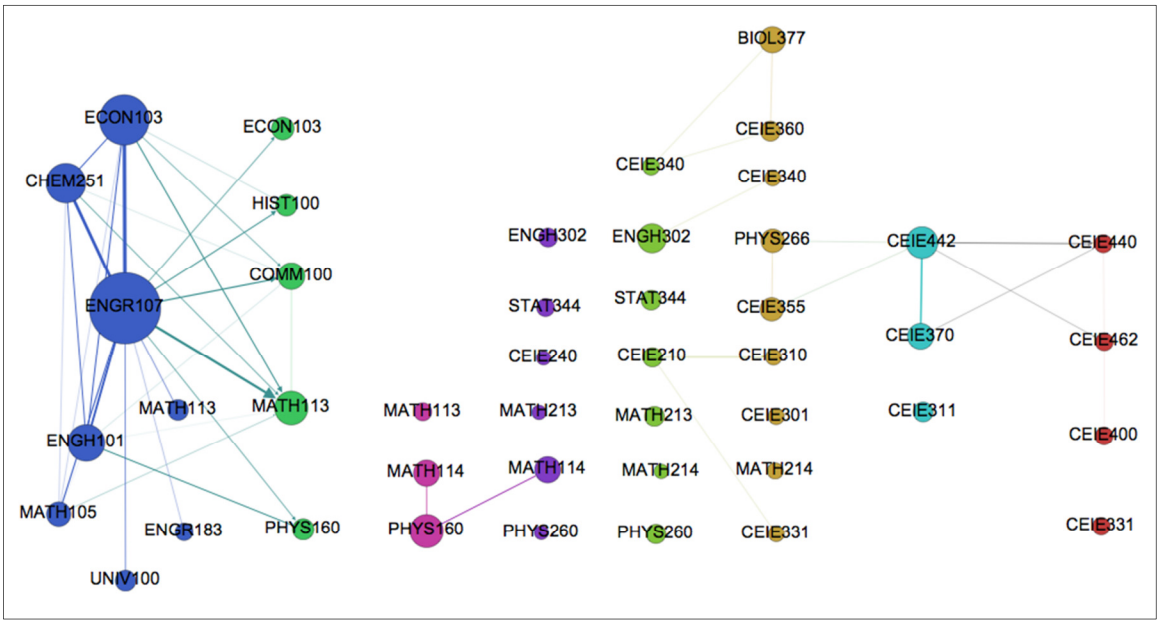

Figure 6: Trajectory of frequent courses for low achieving students (CEIE)

From the graphs, we can summarize the findings as follows: there are some patterns of courses that are frequent in low achieving students that are not frequent in their counterparts. For instance, it is frequent for L group students' to take CS101, CS112, and MATH113 together in the first semester, or HIST100, and MATH113; both patterns are not frequent in the high achieving group. It may suggest that these set of courses together require heavy workload that affect their performance.

Some courses are frequent in more than one semester, which is valid but some courses are frequent in more than one semester and different than the other group. For example, from the CS

graphs, STAT334 is a frequent course for H group in the 5th semester, whereas in the L group, it is frequent in the 4th, 5th, 6th, 8th, and 9th semesters. This observation raises a question weather this course is a bottleneck for the low performing students. In some cases, we can infer that from the graph if we found a link between the same course nodes in two consecutive semesters.

Similar results were found in IT and CEIE. In addition, we found that L group students postponed some courses later in their studies. For example, in IT graphs, IT206 is frequent in the 4th semester for the H group; in contrast, it is frequent in the 6th and 7th semesters for their counterparts L. As a result some consequence courses delayed, which result in late graduation.

In CEIE graphs, there is a distinct feature other than the similarity in the above findings. The graph is more connected in the H group, particularly in the last two years. It may suggest that students in CEIE (H group) have formed their cohort and are enrolling in classes together. On the other hand, we watch enormous repetitions of the courses taken in different semesters for the L group, which may propose a split in the group; some students may fail a prerequisite course that hindered them from advancing with the rest of the group.

6. Conclusion

In this paper, we present the importance of tracing students' course taking patterns in pursuance of understanding the appropriate sequential courses that could improve students' performance and education quality. We project the trajectories of courses for three engineering majors where students were split into two groups high- and low- achieving groups based on their cumulative GPA. Some major insights from the trajectory of the frequent courses include: low performers postponed some courses toward the end of the program, and take a collection of courses together that their counterparts do not usually take. The findings have many implications on different levels; it can be used by programs' policy makers to design new policies that improve the programs' curricula, for example, by sitting some courses as prerequisites to other courses. Moreover, it can be used by academic advisors, current and prospective students to increase their awareness of the paths and course-taking choices that may improve the students' performance and help them graduate on time. One possible future work is to identify the bottleneck courses and investigate the paths that lead to failing or passing them.

Acknowledgements

This work was supported in part by NSF Grant# 1447489. We would like to thank our informants for participating in the field studies reported here. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Pandey, U. K. and Pal, S. (2011), "A Data Mining View on Class Room Teaching Language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, 277-282, ISSN:1694-0814

[2] Perez, T., Cromley, J. G., & Kaplan, A. (2014). The role of identity development, values, and costs in college STEM retention. *Journal of Educational Psychology*, *106*(1), 315.

[3] Moore, T. J., Miller, R. L., Lesh, R. A., Stohlmann, M. S., & Kim, Y. R. (2013). Modeling in engineering: The role of representational fluency in students' conceptual understanding. *Journal of Engineering Education*, *102*(1), 141-178.

[4] Picciano, A. G. (2012). The Evolution of Big Data and Learning Analytics in American Higher Education. *Journal of Asynchronous Learning Networks*, *16*(3), 9-20.

[5] Reason, R. D. (2009). Student variables that predict retention: Recent research and new developments. *Journal of Student Affairs Research and Practice*, *46*(3), 850-869.
[6] http://www.educationaldatamining.org/JEDM/index.php/JEDM
[7] Pandey, U. K., & Pal, S. (2011). Data Mining: A prediction of performer or underperformer using classification. *arXiv preprint arXiv:1104.4163*.
[8] Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*.
[9] Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, *41*(4), 1432-1462.
[10] Su, J. M., Tseng, S. S., Wang, W., Weng, J. F., Yang, J. T. D., & Tsai, W. N. (2006). Learning portfolio analysis and mining for SCORM compliant environment. *Educational Technology & Society*, *9*(1), 262-275.
[11] Parack, S., Zahid, Z., & Merchant, F. (2012, January). Application of data mining in educational databases for predicting academic trends and patterns. In *Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on* (pp. 1-4). IEEE.
[12] Romero, C., Romero, J. R., Luna, J. M., & Ventura, S. (2010, June). Mining Rare Association Rules from e-Learning Data. In *EDM* (pp. 171-180).
[13] Damaševičius, R. (2009). Analysis of academic results for informatics course improvement using association rule mining. In *Information Systems Development* (pp. 357-363). Springer US.
[14] Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, *15*(1), 55-86.
[15] Aggarwal, C., & Han, J. (2014). *Frequent Pattern Mining*. NewYork: Springer.
[16] Goethals, B. (2003). Survey on frequent pattern mining. *Univ. of Helsinki*.
[17] Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, *8*(1), 53-87.
[18] Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Addison-Wesley.