Duke Undergraduate Academic Trajectories: Network Analysis and Course Recommender System

Vanessa Tang, Sangseok Lee

Abstract

Currently, universities rely on academic advisors and students' individual research to guide major and course selection. However, with thousands of possibilities, it is extremely time-consuming and difficult to search through all the options. Using historical enrollment and graduation data, we propose several tools to facilitate major and course selection for undergraduate students at Duke University. First, a series of dashboards aid students in major and course selection as well as course timeline planning. Second, network and alluvial plots show curriculum and course flows from semester to semester to help identify common academic trajectories. We then use network measures to delve deeper into the structure and patterns between majors over time. Third, a course recommender system provides personalized course recommendations to students. Through the use of these tools, both students and advisors can better plan majors, courses, and academic pathways.

Introduction

The abundance of courses available and the variety of academic requirements are often overwhelming for students. First, with hundreds of different majors to choose from, students often find it difficult to decide which major to pursue, especially when they have ill-defined or several different interests. Second, even within a major, students have hundreds and even thousands of options of classes to take, varying in difficulty, specialty, timing, professor, and workload. Third, students must fulfill graduation requirements but can do so in a variety of ways. This leaves students with an almost infinite number of options when choosing a major and classes. A student's final decisions for classes depend on the risk vs. reward trade-offs perceived by each student with his or her individual needs and objectives (Jiang, 2019). Therefore, it is important to provide students with sufficient information about potential academic pathways to ensure that their class selections help guide them towards their goals. Identifying patterns and common courses from students within the same major can help guide students in course selection. Rather than manually searching through the massive amount of courses, students can narrow their search to courses that align with their major.

To do so, we provide several deliverables to facilitate the course selection process for students, given historical data on student enrollment and graduation for undergraduate students at Duke University. These include an integrated planner for major and course selection, visualizations of academic trajectories, as well as a course recommender system. The three main goals of this project are the following:

1. To track academic pathway from the perspective of both students and academic advisors
2. To understand academic pathway structure by using network analysis methods
3. To develop a course recommendation system.

Background

Currently, most universities, including Duke, rely on academic advisors to help students select classes. Students also spend a significant amount of time individually researching courses, majors, and professors to determine which classes to take. This process requires thorough research and thought about potential academic pathways. Despite help from advisors, individual research, and even peer feedback, students still have difficulty determining which classes to take because there are so many options. Furthermore, many first year students enter university not knowing what major they want to pursue. The current system relies on students deciding their major based on personal interests, potential career paths, or even pursuing the same major as their peers. Overall, there is little guidance for students in terms of major and course selection. This issue underscores the importance of a tool to help facilitate this class and major selection process for students.

One proposed tool for facilitating course selection is a dashboard to support learning in the academic setting (Verbert, 2013). Most dashboards focus on graphical representations of historic data on students and courses to give professors a better understanding of course activities and teaching practices (Few, 2006). We built on this idea by creating a series of dashboards using historical student data to help track majors and courses throughout the eight semesters at Duke. The planning dashboards centralize information for both students and advisors and can help facilitate more informed and flexible decision making.

Academic advisors and our client also aimed to better understand students' course-taking patterns. Pathway visualizations can help show course progressions and flows for each major over time (McFarland, 2006). Visualizations of academic flows for courses within a major can help identify the most popular pathways and course progressions across each semester, and thus help guide course selection.

To better understand the structure within a major, we approached the problem as a network and used network measures such as centrality and modularity to identify patterns and differences among majors and across semesters. This is an uncommon technique in terms of recommendation systems, but highly valuable in understanding structure and patterns seen from visualizing academic pathways. Centrality and modularity can also help identify core courses within a major and term.

In addition to the above visualizations, a course recommendation engine could also provide a more personalized approach to aiding students in the course selection process (Lee and Cho, 2011). Recommendation engines are commonly seen today in online platforms such as Netflix, Spotify, and Amazon, but are very uncommon in the academic setting. Unlike feedback from advisors or peers, course recommendation systems can further help manage the curriculum and keep track of students' academic progress in a more personalized manner (Lee and Cho, 2011). Thus, in comparison to dashboards or academic trajectory visualizations, recommender systems learn from past students' course-taking patterns, and thus, provide a much more personalized approach to guiding course selection.

Overall, the development of a tool to facilitate course and major selection could be extremely beneficial for students. It could enable students to select a major that aligns more with their courses and interest, enable selection of courses that will fulfill requirements, as well as help students save time when looking for classes (Verbert, 2013). Past work includes implementation of dashboards and recommendation engines (Verbert, 2013; Chen, 2004), and our project builds off these ideas to create a more tailored approach to undergraduate courses and major planning at Duke University.

To accomplish the first goal of tracking academic trajectories, we developed a series of dashboards to facilitate major and course selection as well as course planning, and constructed visualizations of academic trajectories based on network analysis. Then, to better understand the structure and patterns seen through these visualizations, we used network measures such as centrality and modularity to quantify differences between majors and across semesters. To accomplish the last goal of developing a course recommendation engine, we built and tested several recommendation algorithms to identify one model that best predicted the classes a student would take in a certain term of interest. This schema is detailed in figure 1.

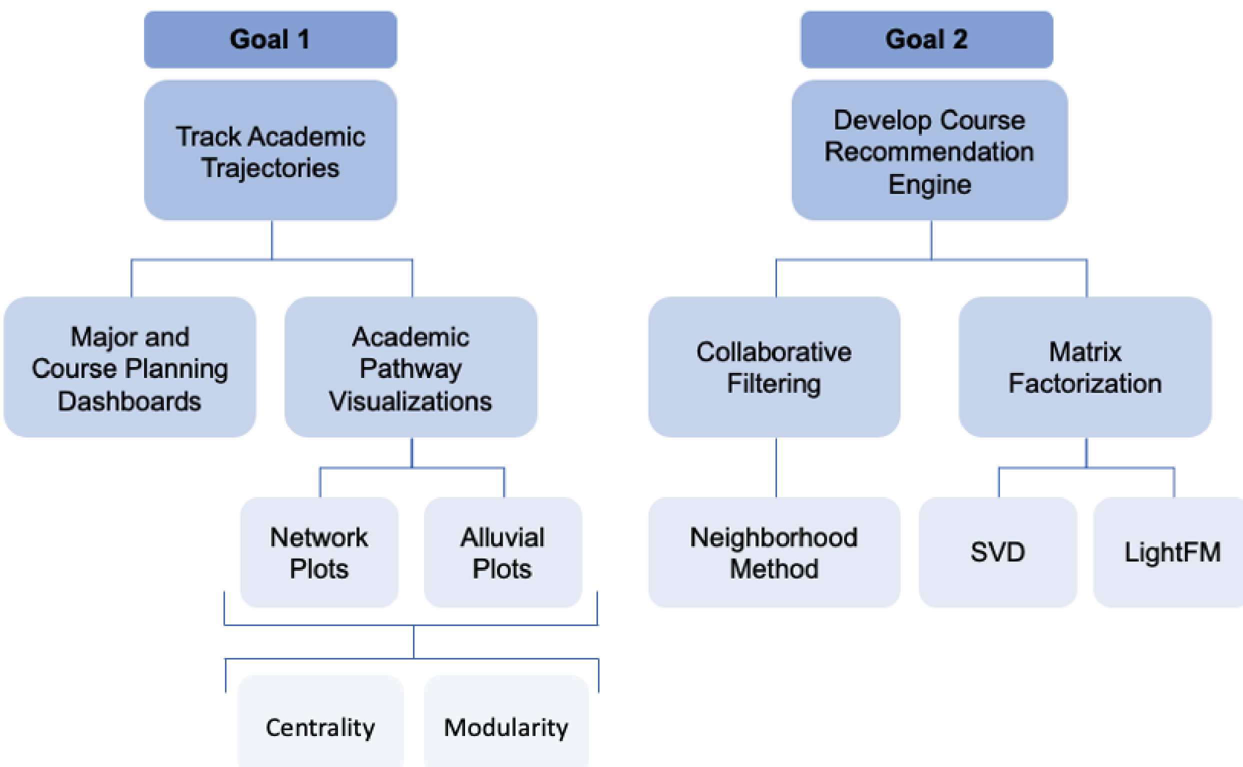


Figure 1. Overall schema of project, including primary goals and steps taken to work towards each goal.

Data

From the Duke Registrar, we have two datasets: one with degree data and a second with enrollment information, each spanning 2013 to 2020. The degree dataset contains graduation degree information, including graduated students' de-identified IDs, graduation year, and major, minor, certificate, or secondary degree descriptions. The second dataset contains all course enrollment information, including de-identified IDs corresponding to those in the degrees dataset in addition to course descriptions, grades received, and academic year and term for every course each student took at Duke. The two datasets were combined by merging on student ID (`Calculation ID`) then cleaned and reorganized to create a dataset with one row for every course taken per student and separate columns for major, minor, certificate, and secondary degree information merged from the degrees dataset. The detailed data dictionary can be found in the appendix. The final dataset has 325,107 rows with 13,513 students and 4,224 unique classes. There are 141 different primary majors.

At Duke, every student graduates with a major, but students have the option of combining three of these options: major (MAJ), minor (MIN), secondary (SEC), and certificate (CER). As there are many different combinations of majors, minors, secondary, and certificates and all students must graduate with a major, we are focusing mainly on course planning for a student's primary major in this project. Future work could help students identify secondary majors as well as minors, secondaries, or certificates.

Major and Course Planning Dashboards

Introduction

The first goal of this project was to track academic pathways to facilitate the course selection process for Duke undergraduate students. To do so, we constructed a series of Tableau dashboards to aid in major and course selection as well as course planning. Tableau dashboards were chosen because they enable easy public access through Tableau Public, allow for extensive filtering for majors and/or classes, and are generally very user-friendly. The dashboards can be accessed through the following link:

Duke University: Major and Course Planning

Methods

When creating the dashboards, we focused on three primary goals, each of which contributed to one page on our dashboard site. The three goals are to help students 1) select a major, 2) search for classes, and 3) plan courses over time. We chose to focus on both major and course selection because major requirements are one of the main reasons why students choose classes. Aiding with major selection is important because many students, especially in their First Year, are undeclared majors, and students can change majors at any time at Duke. Thus, choosing courses based on major is one of the key guiding factors for these dashboards. These visualizations are based entirely on historical data of course enrollments and required no additional processing aside from the preprocessing and data cleaning previously detailed.

The first goal is to help undergraduate students select a major. This is particularly helpful for First Year students who may not know what major to pursue or for students with an undeclared major. In order to help students select a major, we chose to use course-taking history to show the most popular majors given the courses a student has already taken. This is helpful because it can identify a major pathway that the student is already working towards. For example, if a student has already taken Economic Principles and Laboratory Calculus, they may already be on track toward completing an Economics major. This led to the creation of the *Major Selection* page. On this page, students enter in the classes they have already taken, which serve as a filter to show the most common major in terms of number of students given the classes entered.

The second goal is to help students find courses based on their major. This is helpful because major requirements are one of the main reasons why students choose to take certain classes. Working towards this goal, we chose to show students the most common classes to take for their selected major and semester. This can help students pick classes based on what previous students with the same major have also taken. To help students choose classes based on major and semester, we created the *Course Selection* page. On this page, students search for their major and then select the term for which they are searching for classes. These serve as filters for the dataset, and show the popular courses for the selected major and term.

The third major goal of these dashboards is to help students plan their courses over time. This can help with course load planning to ensure that students aren't overloading semesters with too many difficult or time-consuming classes. Furthermore, it can help show students when it is common to take a certain class to ensure they have time to complete course prerequisites and are not taking an upper-level class too early. In order to help with course planning, we created the *Course Planning* page, which shows the number of students who have taken the courses entered over time. Similar to the major selection page, this page also shows the most common major given the courses entered. This helps students ensure that the courses they plan on taking in the future align with their planned major. Though this could be done by entering the same courses on the *Major Selection* page, this page is targeted to planning in the future and also doesn't require the students to re-enter classes on the first page. On this page, students enter the classes they are interested in taking, which then filters for students who have taken the classes entered and shows the most popular majors within the subset of students. It also plots the number of students who took the classes entered for each semester.

Note that the *Major Selection* and *Course Planning* pages require students to enter in classes. Small classes or classes that have not been frequently offered may alter results because there are few students who take the selected class. Thus, using a small class as a filter may result in a very small subset of students remaining. This is one important limitation that is innate to the data and our problem. Future work could exclude small classes, though this must be done with caution to limit data loss.

For each of these pages, bar plots were chosen based on data visualization principles. As all plots rely on comparison of counts to show popularity, bar plots easily show differences in bar length, making it easy to quickly identify the most popular bar. Furthermore, the absolute numbers are less important in these plots, while the relative lengths of bars are more important. Thus, when applicable, we sorted bars from highest to lowest count to ensure that the most common choices for courses or majors were shown at the top. The only plot where bars were not sorted was when the counts are shown over time. In this case, the x-axis is shown in chronological order from first year fall term to fourth year spring term because this is the order in which students take classes at Duke. Overall, we selected bar plots for all plots based on visualization principles to make each plot quickly and easily understandable.

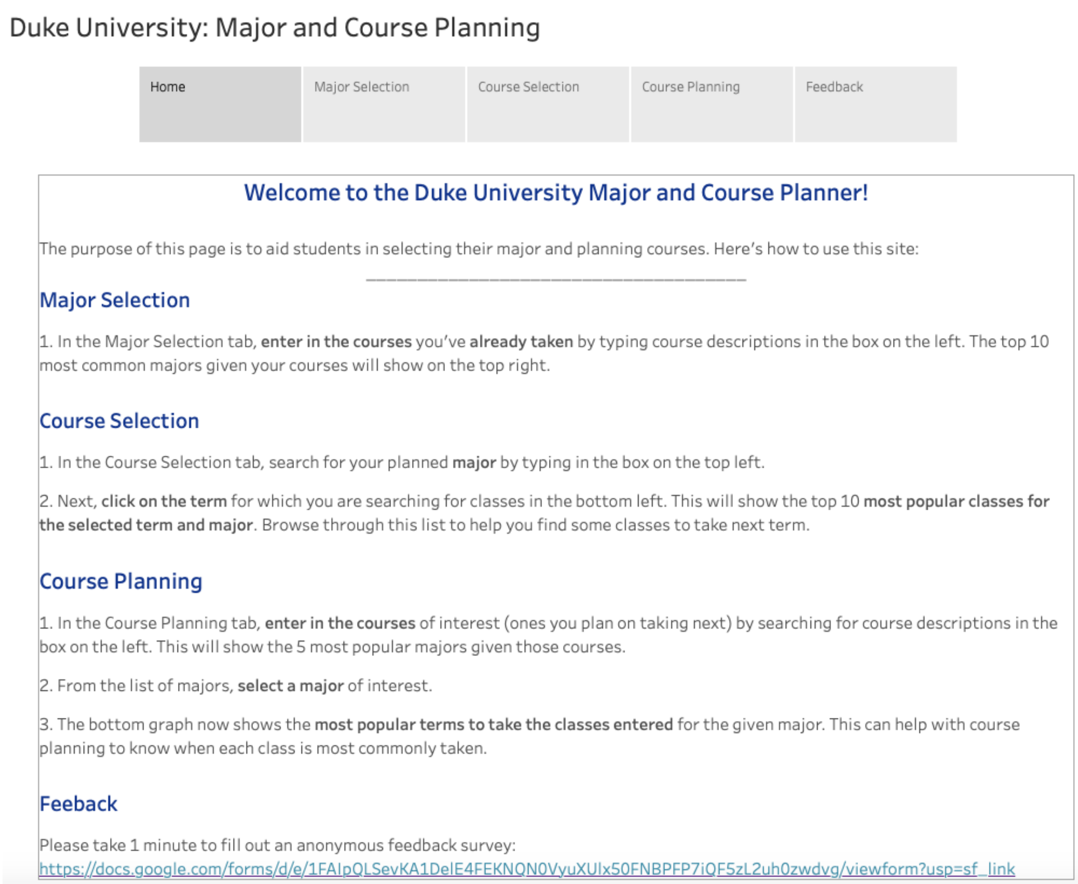
Lastly, we implemented a feedback form to enable users to rate each page and provide comments and suggestions for improvement. This is crucial to the development of these dashboards, because they are a prototype product and require feedback from users to determine if these dashboards are helpful. This feedback can be used to improve and adjust the dashboards based on actual users' comments and ratings.

Pages were ordered based on feedback from our client, where pages focusing on courses already taken precede the page focused on courses to take in the future. The pages follow a logical flow, where *Major Selection* is first because it is the guiding factor for course selection and planning in the next two pages. *Course Selection* is second, because students should already have a planned major, and it helps students fitler through courses to take in the future. *Course Planning* is third, because it focuses on planning for future terms and is based on the *Course Selection* page before. The *Feedback* page is last so that students and advisors can fill out the survey after using the dashboards.

Results

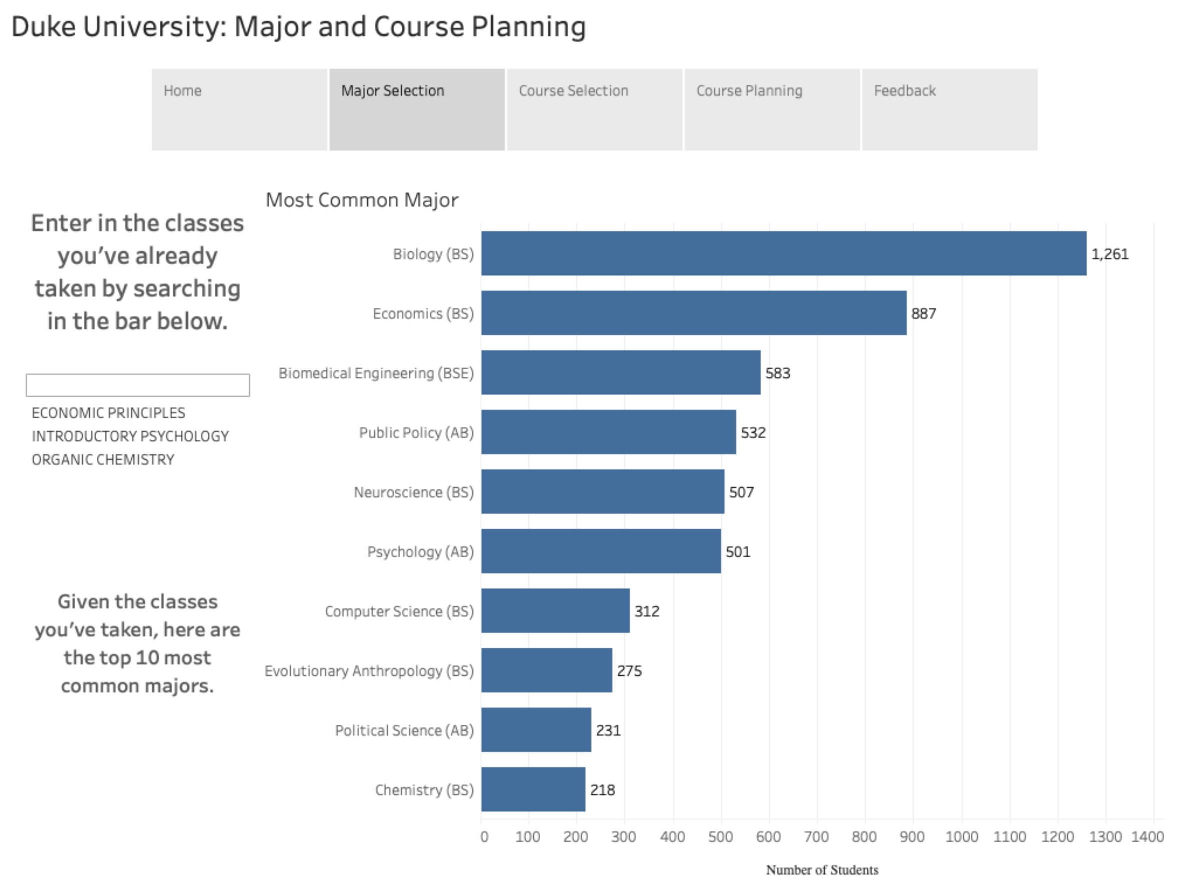
The website consists of the following five main pages:

1. **Home**
   The Home page provides brief instructions on how to use the dashboards as well as a link to the feedback form.



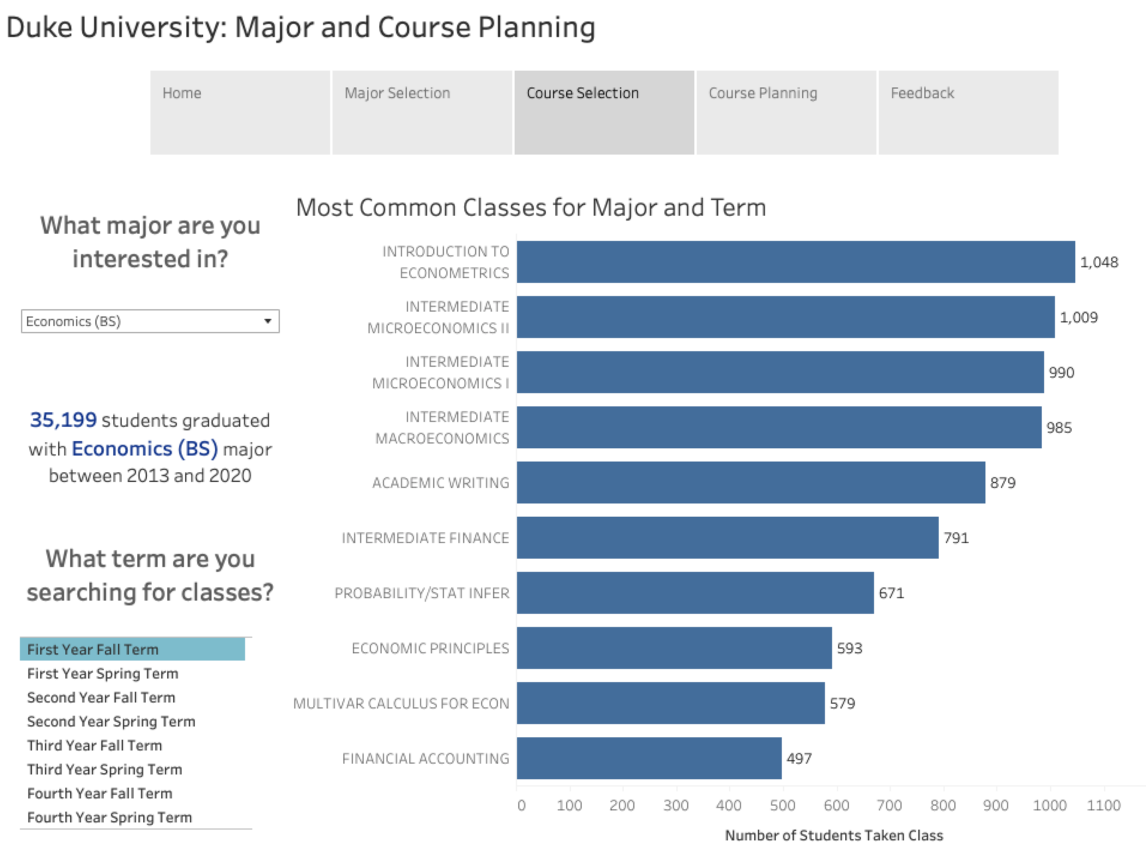
2. **Major Selection**
   The Major Selection page enables students to search for classes they have *already taken*, then shows the top 10 most popular majors given the classes entered. The purpose of this

page is to help narrow down the search for majors based on classes a student has already taken, as they may already be on track to complete a certain major. This can be especially helpful for first year students who may need more guidance in major selection.



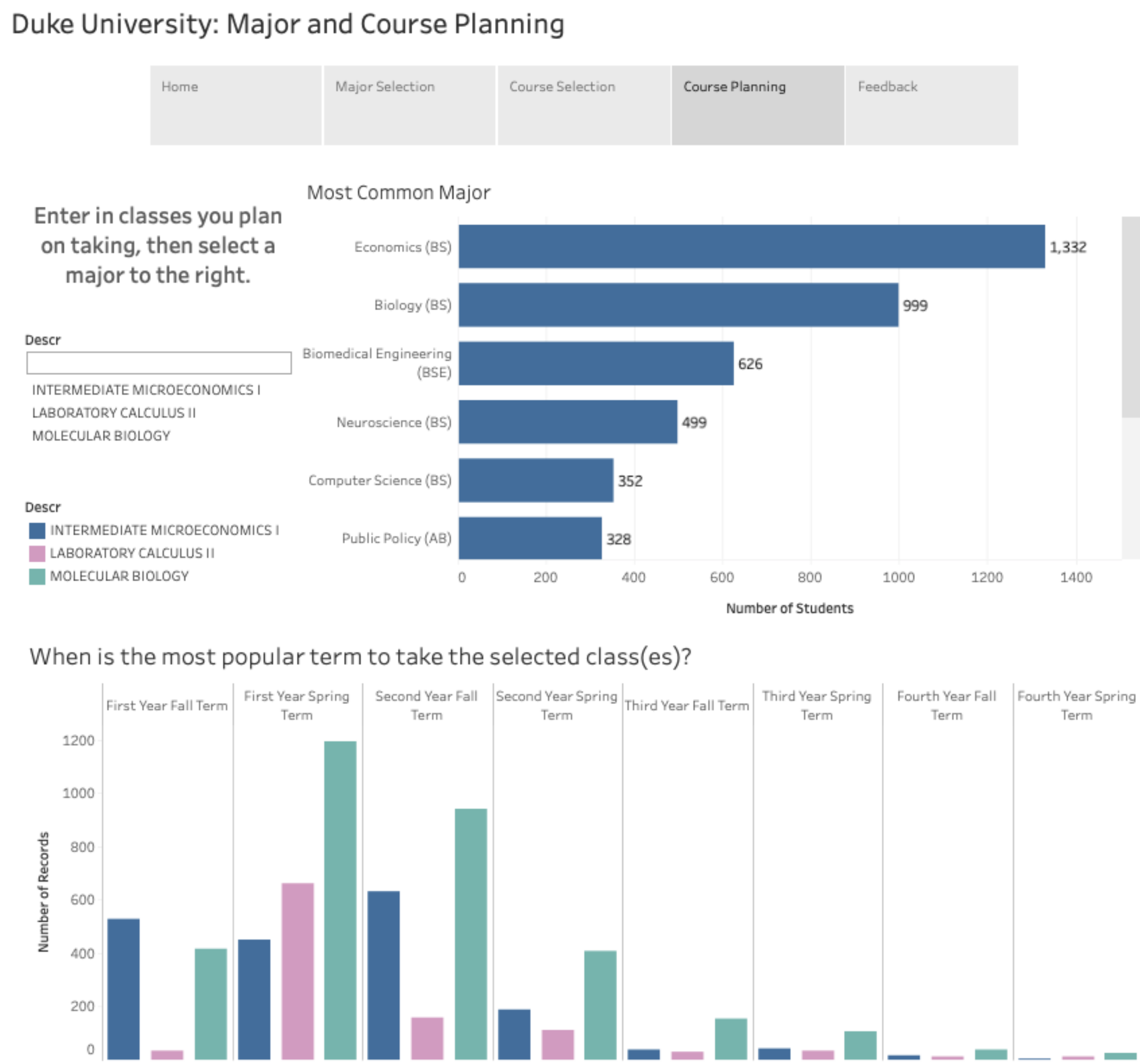
3. Course Selection

The Course Selection page enables students to search for a major of interest, then select the term for which they are searching for classes. It then shows the top 10 most popular classes to take for the selected major and term as well as the total number of students who graduated with that major between 2013 and 2020. The purpose of this page is to show students the common classes to take for their major and term. This can essentially identify requirements or even other non-major classes that many students within a major decide to take.

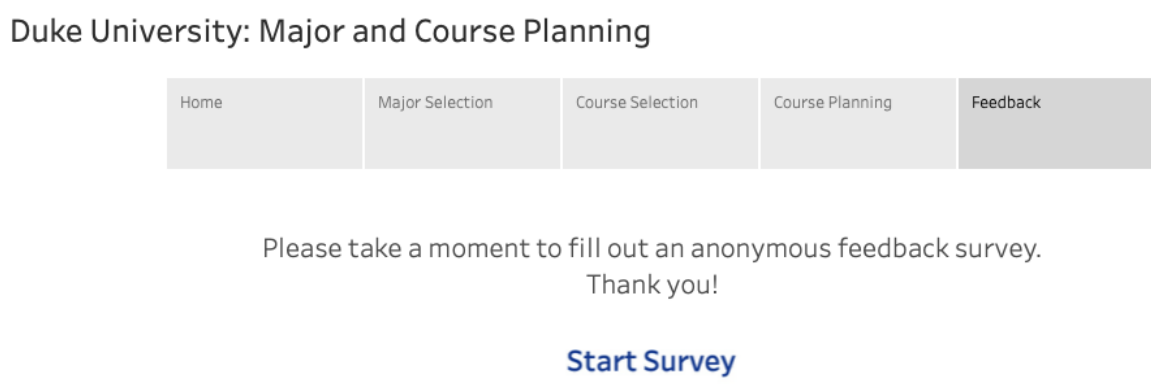


4. **Course Planning**

The Course Planning page enables students to search for classes they plan on taking *in the future*. Based on these classes, it shows the top 10 most common majors as well as a timeline of when the classes entered have historically been taken. By showing popular majors given the courses entered, students can determine if these classes align with their major. The timeline feature can also help students balance their course load. For example, if it is common to take Molecular Biology in both the First Year Spring term and Second Year Fall term but the student's First Year Spring term already has a heavy load, then he/she can choose to wait until the Second Year to take that class knowing that it is a common trend.



5. **Feedback Survey**
   The Feedback page provides a link to a Google form where students and advisors can rate each page and provide comments. The purpose of this page is to enable us to receive feedback and continue making improvements on these dashboards.



Overall, this website consists of several dashboards that aim to facilitate major and course selection as well as course planning for students. These dashboards aim to be both easily understandable and helpful for the entire undergraduate student population as well as the academic advisors.

These dashboards help accomplish the first goal of this project of tracking academic pathways while creating a final end product usable by both students and advisors. We aim to continue improving these dashboards based on feedback from students and advisors. All in all, these dashboards serve to aid undergraduate students in selecting both majors and courses as well as planning course load.

Feedback and Future Work

We invited academic advisors of the Duke Registrar for a panel interview and gathered feedback regarding the dashboards. Overall, the feedback was extremely positive with a future direction focusing on integrating more information within the tool. Advisors were impressed with the synthesis of a large amount of information in a relatively ingestible way. Based on their feedback, we identified the following primary areas for future work for these dashboards:

- **Addition of more course data:** Future work could synthesize more data to allow students to see more course information, including locations, times, requirements, prerequisites, descriptions, and professors. These attributes often guide student course selection and could be a helpful addition to the dashboards.

- **Addition of major requirement data:** Currently, major requirement data is not available, as these often change from semester to semester. However, future work could synthesize course information with major requirement information to better guide students towards completing the necessary classes for their major.

- **Searching by course department and number:** The current implementation enables course searching by the name of the course. Searching by department and number (ex: Econ 101) would also be helpful to students. However, course numbering and/or names often change, so this would require contiguity across all data to ensure reliable results.

- **Filtering through Areas of Knowledge:** Within the Trinity College of Arts and Sciences at Duke, there are five areas of knowledge: Arts, Literatures, and Performance; Civilizations; Natural Sciences; Quantitative Studies; and Social Sciences. All students in Trinity must complete 2 credits within each area to graduate. Thus, this requirement guides the course selection process for many students, as Trinity is the largest college within Duke University. Based on feedback from advisors, it would be helpful to enable filtering classes by a selected Area of Knowledge, then show the most popular classes and/or the classes with the highest average grade. This could help students choose common classes that satisfy the Areas of Knowledge requirement as well as identify classes that are likely to be easier, as indicated by the higher average grade. Future work could integrate data with Areas of Knowledge information to combine with the current dataset on grades and enrollment.

  Note that providing grade information is controversial as it may discourage students from taking certain classes, even if they are major or graduation requirements (Main, 2014). Likewise, it may affect class enrollment and discourage advisors from recommending students to take a certain class knowing that historical data shows low average grades (Main, 2014). Thus, we did not provide grade information, and it must be used cautiously due to its potential implications for guiding course selection.

- **Integration with DukeHub:** Ideally, this tool would be completely integrated with DukeHub, Duke's platform for course searching and registration. This would link the additional information of description, syllabus, timing, professor, etc. as well as enable students to add their courses to their cart then register.

- **Identifying patterns in combinations of classes:** Future work could identify classes that are commonly taken together to further help identify academic pathways. This could also identify combinations of courses where students are more likely to receive lower grades. Likewise, it could identify combinations of more difficult (lower average grade) courses taken with easier (higher average grade) courses to help students balance course loads and avoid taking too many difficult courses at the same time.

There is a wide variety of additions that can be made to this tool to further facilitate course and major selection for undergraduate students. While most of these suggestions are out of the scope of this project or not currently possible with the data provided, they can guide future work on this project to improve the tool for both advisors and students. Notably, almost all feedback from advisors were additions to the dashboard rather than revisions to the existing work. This emphasizes the effectiveness of our dashboards in providing useful information in an understandable way to help guide the course and major selection process for Duke undergraduate students.

Academic Trajectory Visualizations

Introduction

While the major and course planning dashboards provide a deliverable usable for students and advisors, we also were interested in analyzing course progressions and patterns. Thus, we implemented network visualization tools based on social network analysis. *Academic trajectories* can be defined as the progression of classes a student takes throughout their undergraduate career. As stated by McFarland, academic trajectories show the “patterned flow of students across courses” (McFarland, 2006). This can be helpful in identifying common patterns of courses to take for each major, while highlighting prerequisites or courses that should be taken before proceeding to higher level courses. Furthermore, tracking pathways through visualizations can emphasize a common course-taking pattern for certain majors, thus identifying majors that may have more structure than others.

Methods and Results

Network Plots

To represent participant flows across courses, students’ schedules are used from two consecutive semesters to create large affiliation matrices (Friedkin and Thomas, 1997; Sorenson, 1987). In each affiliation matrix, rows are students, columns are courses, and cell values of 1 and 0 indicate membership in a particular semester. All observations where a student failed or dropped out mid-semester were excluded. For example, let A_F be an affiliation matrix of all students’ memberships in the first year fall semester and let A_S be an affiliation matrix of all students’ memberships in the first year spring semester. Matrix algebra is used to calculate participant flows across courses (Wasserman and Faust, 1994). To construct a mobility matrix, A_F and A_S must have the same number and order of students (rows) so that the transposed affiliation network of A_F (A_F^T) can be multiplied by the affiliation network of the spring semester (A_S). Through matrix multiplication, a mobility matrix of origins by destinations is created, **M** = A_F^TxA_S. The cell values in this matrix indicate the number of students who transitioned from the Fall semester courses to the Spring semester courses. When presented in table form, the results are considered transition frequencies from origin states to destination states (Chase, 1991)

Transition frequencies show the volume of flows but do not test the likelihood that students will move from an origin state to a destination. Such a test can be constructed from the transition frequency tables by dividing the values in each row by the total number of students who are in the course (Chase, 1991; Sorenson, 1987). The new cell values represent maximum likelihood tests or transition probabilities that indicate the proportion of students from each course who move to the next position.

With the transition frequency table, we draw a network plot for the top 10 classes from each semester (Figure 2). Each node represents a class and each line indicates students’ transition. The yellow line indicates course transition probabilities greater than 0.2. This plot shows a pattern in course transitions from semester to semester and can help guide students and advisors in choosing courses for the following semester.

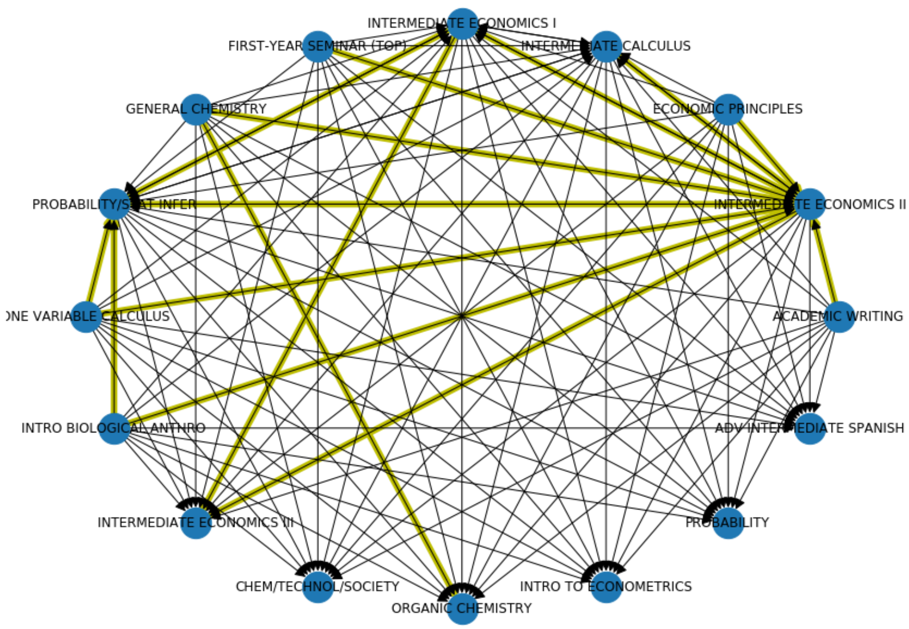


Figure 2. Network plot for economics majors from First Year Fall term to First Year Spring term. Each node is a class in the top 10 classes for each semester, directional lines indicate transitions from one class to the next, and yellow lines show transition probabilities greater than 0.2.

Alluvial Plots: Two Semesters

While network plots help identify relationships among courses and common course-taking patterns, they do not indicate the popularity or number of students who take certain course progressions. Furthermore, they do not show as clear of a “pathway” from semester to semester. To help combat some of the drawbacks of network plots, we implemented alluvial plots. In this case, alluvial plots show the flow of students from the popular classes in one semester on the left to the popular classes in the following semester on the right. The thickness of the bands indicate the number of students who progress from the class on the left to that on the right. Note that the actual implementation of the alluvial plots enables hovering with tooltips and shows the number of students from class to class. For the purpose of this paper, only an image of the alluvial plot is shown in Figure 3, and interactive plots can be found in our GitHub repository.

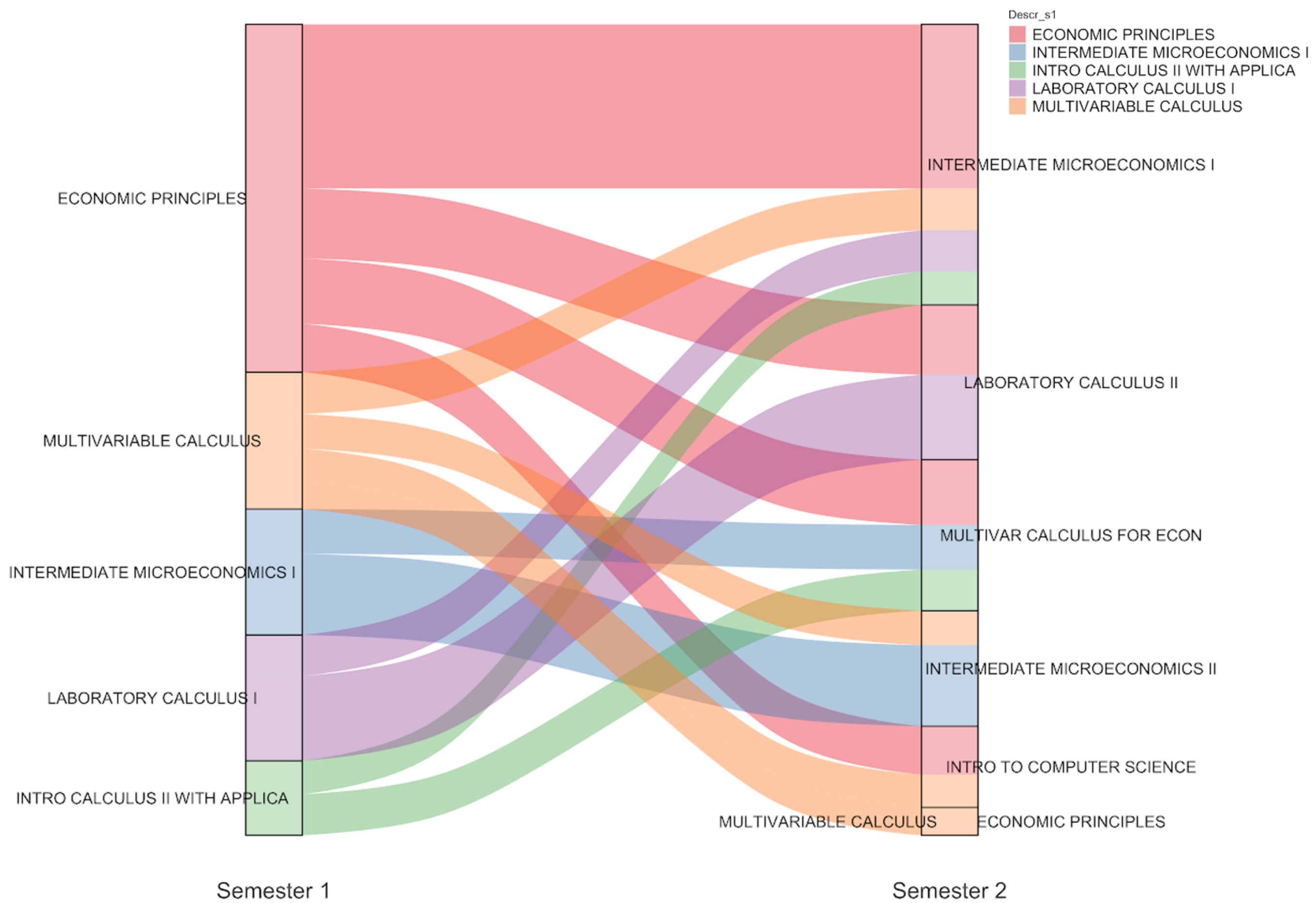


Figure 3. Alluvial plot for economics majors from first (First Year Fall term) to second (First Year Spring term) semester.

Figure 3 shows an alluvial plot for economics majors and identifies common course progressions from the first to the second semester. Based on the thickness of the pink bar, we can see that the most popular course for economics majors to take their first semester (First Year Fall term) is Economic Principles. Following the thickest bar, we can see that it is then most common to take Intermediate Microeconomics I during the second semester (First Year Spring term). On the other hand, some economics majors then take Intro to Computer Science their second term, but it is a much less popular pathway. Overall, these alluvial plots show common course-taking patterns from semester to semester to highlight major-based course progressions.

Alluvial Plots: Eight Semesters

In addition to using alluvial plots to track semester to semester changes, we also used alluvial plots to show pathways over all eight semesters. This is another main advantage of alluvial plots over network plots, as they are easier to interpret especially when tracking all eight semesters. The chronological ordering from left to right with the eight-semester alluvial plots is more intuitive and shows pathways in a clearer manner.

For each eight-semester alluvial plot, vertical bands indicate the semester from First Year Fall term to Fourth Year Spring term and are ordered chronologically based on the semester system. Similar to the two-semester alluvial plots, band thickness indicates the number of students in each progression, and the actual implementation of these alluvial plots also enables hovering with tooltips. However, there are two important differences between the two- and eight-semester alluvial plots. First, because of the wide variety of classes possible, all classes not in the top 20 most popular classes are grouped together as “other.” This makes the alluvial plots significantly easier to interpret, and because we aim to identify common pathways, we are less concerned with less-popular courses. Thus, grouping them together leads to simpler plots and minimal loss in information. Second, students are grouped based on following the *exact same* pathway over all eight semesters. To create a simpler visualization, a threshold value must be applied within each major where all pathways with fewer students than the threshold value are dropped before plotting. These threshold values are indicated in each plot title.

Notably, the threshold value for each major was manually tuned to create plots visually similar in complexity. Too high of a threshold may eliminate many students, leading to less complex plots, but also less data. Too low of a threshold results in overly complex alluvial plots with several thin bands, making the plot more difficult to interpret and pathways more difficult to identify. Higher threshold values indicate that more students within that major take a similar pathway over all eight semesters. Conversely, lower threshold values indicate that students take a wider variety of classes over time and cannot be grouped together as easily. For example, for economics (Figure 4), the threshold value is 8, while for biomedical engineering (BME) (Figure 5), the threshold value is 20. Therefore, there seems to be a clearer pathway for BME than for economics. This is also evidenced by the overall wider bands in BME in comparison to economics, where there are more thin bars with fewer students. For BME, there are a few courses throughout semesters 1 through 4 that a large portion of students take. In contrast, the alluvial plot for economics has many thin bands and fewer thick bands, indicating that economics majors likely take a wider variety of courses.

These discrepancies in threshold values and band patterning between economics and BME majors indicates differences in course-taking patterns and course progressions. This is helpful in differentiating between majors that have clearer, more structured pathways and majors where students take a wide variety of courses. In the context of our project, this may also be indicative of recommendation engine performance, as majors with more variability may be more difficult to predict, and recommending courses may be more difficult. Overall, comparing the eight-semester alluvial plots for economics and BME majors highlights course-taking patterns and structural differences between these majors.

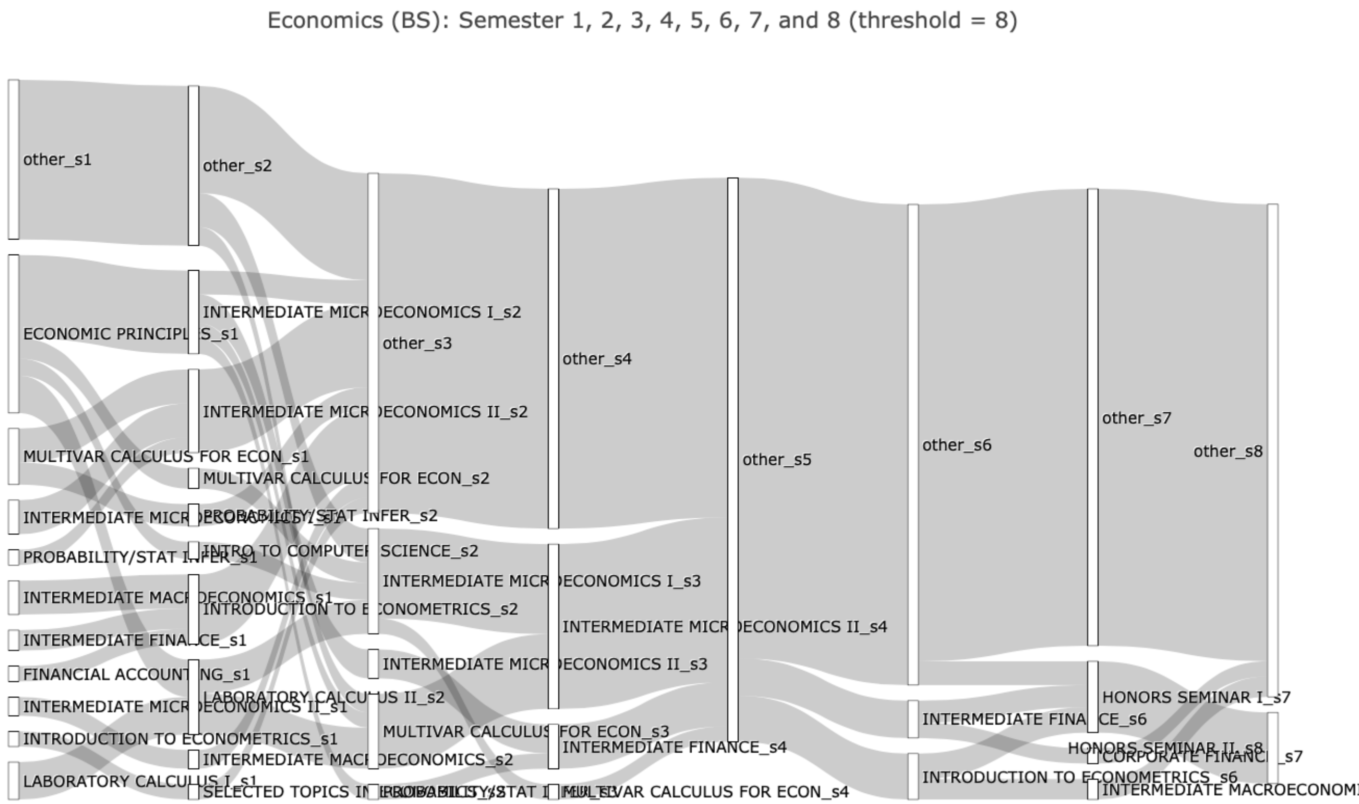


Figure 4. Eight semester alluvial plot for economics majors.

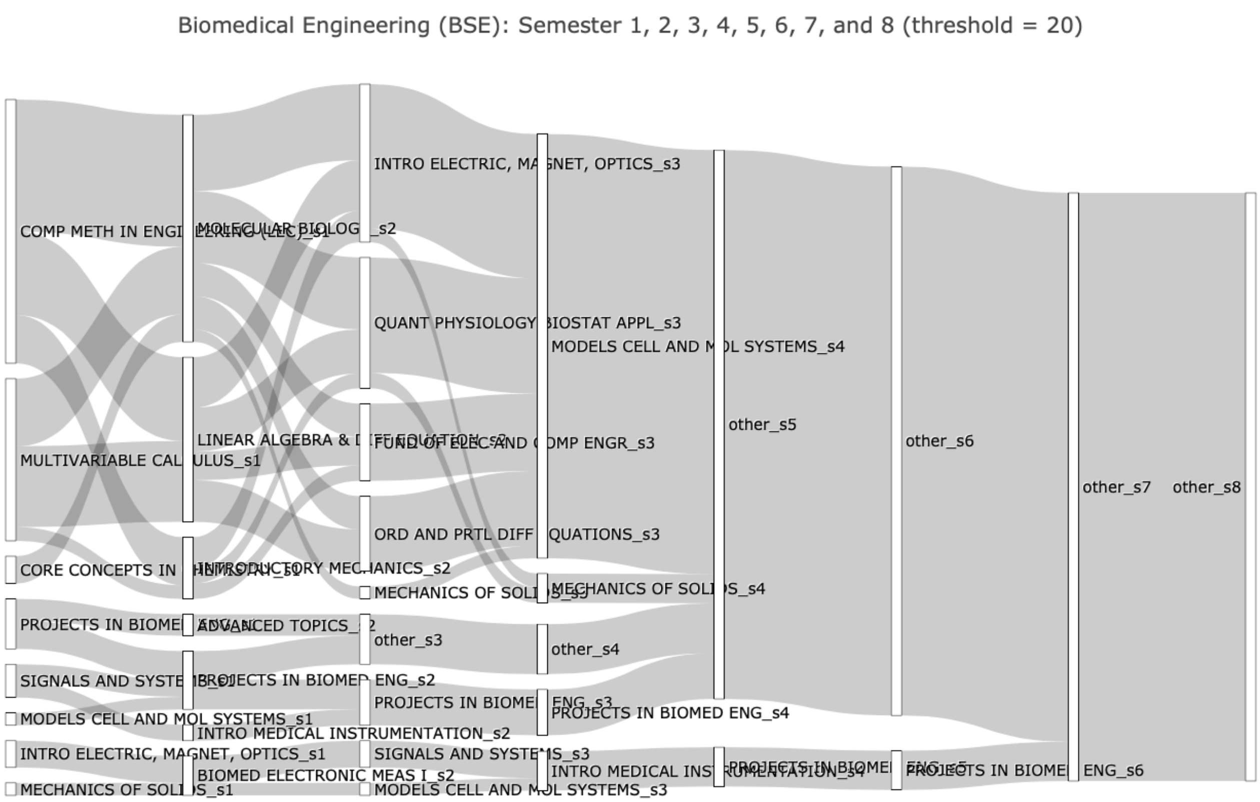


Figure 5. Eight semester alluvial plot for biomedical engineering majors.

Discussion

One of the main advantages of alluvial plots is that they are more conducive to highlighting pathways over time. The flow of bars from left to right is more intuitive in terms of time in comparison to the circular pattern of network plots. Furthermore, they show the scale of each course progression through the thickness of bars, providing a measure of popularity or commonness. This can help identify more common versus less common course progressions. For these reasons, it is helpful in identifying common academic pathways and is very beneficial for research purposes.

Despite some of these advantages, the primary disadvantage of alluvial plots is that they are a less common or familiar plotting style, making them less interpretable to the average person. Thus, we chose not to include them in our final deliverable for advisors and students. In terms of making information ingestible and usable for advisors and students, feedback from advisors allowed us to conclude that this course progression information would be helpful if presented in a different manner. Future work could include the addition of a dashboard that would present course progression information in a more ingestible manner for students and advisors.

Network Structure

Introduction and Background

Though the network and alluvial plots help identify patterns and flows, to understand academic pathway structure, we used centrality and modularity to help quantify differences between majors and across semesters. Through plotting, we found that each major followed different patterns and trends. For example, some biomedical engineering (BME) seemed to have a very clear path that a large portion of students followed throughout all eight semesters, evidenced by the higher threshold value (Figure 5). However, economics students took a much wider variety of classes, evidenced by significant branching in the alluvial plots and lower threshold value (Figure 4). Likewise, we observed differences among semesters, where certain semesters seemed to have core courses while others did not. Based on these observations, we aimed to quantify these differences between majors and across semesters by using network measures. Treating each major as its own network allows us to measure structuredness and connectivity among classes and identify reasons behind the course-taking patterns observed.

Again, we compared two majors, economics and biomedical engineering (BME), using both centrality and modularity across all eight semesters. We chose economics and BME, because at Duke, they are two very popular majors, thus providing more data than less popular majors. Furthermore, economics and BME are two very different majors, where economics falls more into arts and sciences while BME is engineering. Lastly, network and alluvial plots showed clear differences in these two majors in terms of pathways and course progressions (Figures 4 and 5). Because of these reasons, we decided to focus on comparing economics and BME majors in terms of centrality and modularity. We chose centrality and modularity because centrality helps to quantify the importance of individual courses within a major and modularity helps to quantify the strength of course clusters within a major.

Centrality

Within a network, centrality identifies when some nodes are more important or central than others (Crucitti, 2006). Centrality was first introduced in the context of social systems, where a relation was assumed between the location of an individual in the network and his/her influence and/or power in group processes (Wasserman, 1994; J. Scott, 2000). Various centrality measures have been proposed over the years to quantify the importance of an individual in a social network and have been applied to biological, technological, and geographical networks (J. Scott, 2000). Degree centrality counts how many neighbors a node has. In directed networks, there are two versions of this measure: in-degree and out-degree. In-degree is the number of incoming links or the number of predecessor nodes, and out-degree is the number of outgoing links or the number of successor nodes. The standard centrality measure is based on the idea that important nodes are those with the largest number of ties to other nodes in the network. This gives

$$C_D(j) = \sum_{j=1}^{n} A_{ij}$$

where A_{ij} is the (i,j)th entry of the adjacency matrix and n is the total number of rows in the matrix. Thus, the degree centrality (C_D) is calculated by adding up all possible values of the cells designated by the row i and column j combination in matrix A. A higher value of a node indicates that the node is more central compared to other nodes.

Modularity

In order to measure multiple connections among classes, we use a quality function known as *modularity* (Newman and Girvan, 2004), which is the most widely used to quantify the strength of community structures in the network literature (Fortunato, 2010). For a weighted and directed network and a cluster partition C = {$C^{(1)}$,$C^{(2)}$, … , $C^{(K)}$} of the classes into K clusters, the modularity of the partition can be calculated by generalizing the formula given in Leicht and Newman (2008). This gives

$$QC = \frac{1}{w}\sum_{i \in V}\sum_{j \in V}(w_{ij} - \frac{w_i^{out} w_j^{in}}{w})\, \delta(C_{i,} C_j))$$

where w_{ij} is the (i, j)th entry of the weighted adjacency matrix W, $w_i^{out} = \sum_j w_{ij}$ and $w_j^{in} = \sum_j w_{ij}$ are, respectively, the weighted in- and out-degrees of class i and j, and $w = \sum_i \sum_j w_{ij}$ is the total weight of the network. Finally, $\delta(C_{i,} C_j)$is the Kronecker delta function, which is equal to 1 if both *i* and *j* belong to the same cluster and zero otherwise. Intuitively, this equation might be understood as the sum over the proportion of within-cluster transitions minus what would be expected under a random mixing model, where students move randomly from class to class given the constraint that the weighted in- and out-degree of each class is preserved. A modularity of zero indicates that the classification scheme is no better than the random mixing model, while high values of Q indicate a tendency of transitions to be confined within the classes above what is expected by the random mixing model. In other words, networks with high modularity have dense connections between the nodes within clusters but sparse connections between nodes in different clusters.

We applied a walktrap algorithm to detect communities through a series of short random walks with the idea that the nodes encountered on any given random walk are more likely to be within a community than not. walks of length are suggested to use 4 or 5 steps so that we used 5 steps to detect communities.

Results

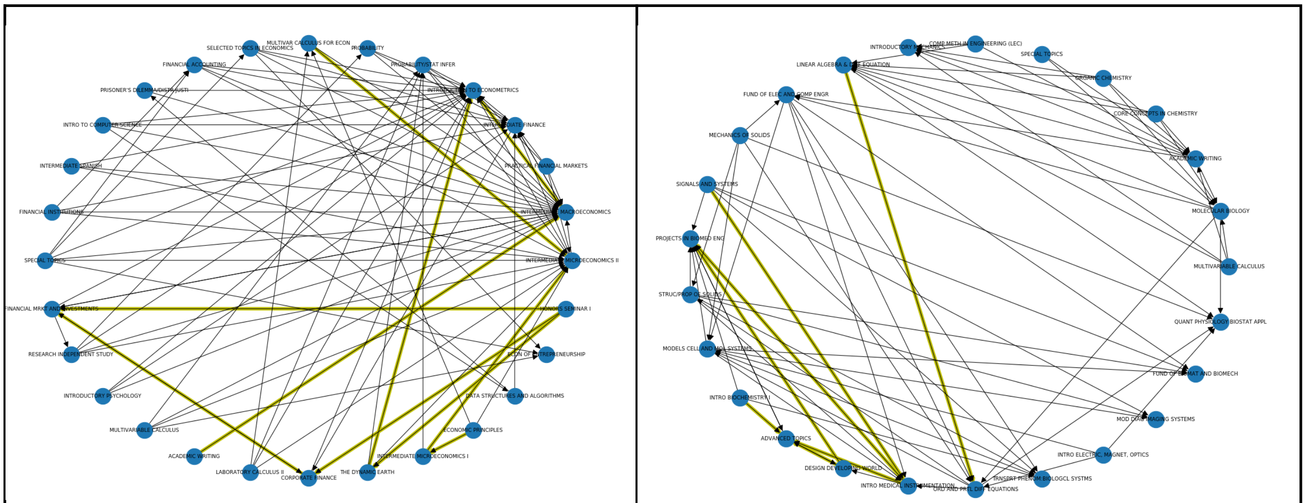
We calculated and compared centrality and modularity for economics and biomedical engineering (BME) across all eight semesters (Table 1). Economics has a higher average centrality than that of BME, but the centrality changes for each semester. Generally, centrality is higher in the third year. This may indicate that students potentially tend to focus on key major requirements or core courses during the third year rather than taking a larger variety of courses. Likewise, first and fourth year have relatively lower centrality. During the first year, students are often still determining their major, and during the fourth year, students have often completed requirements and taken other diverse classes for fun.

Table 1. Centrality of two majors (Economics and Biomedical Engineering) across all eight semesters.

	1st Yr Fall	1st Yr Spring	2nd Yr Fall	2nd Yr Spring	3rd Yr Fall	3rd Yr Spring	4th Yr Fall	Average
Economics	0.386	0.446	0.384	0.613	0.755	0.662	0.512	0.54
Biomedical Engineering (BME)	0.182	0.332	0.431	0.478	0.568	0.719	0.54	0.46

To further see the difference in centrality between these two majors, we use the network plots to graph the semesters for the highest and lowest centrality (Figure 6). The semester with the highest centrality is the 3rd year fall semester for economics. The network plot shows that students tend to take a wide variety of classes, but intermediate macro, intermediate micro II, and intro to econometrics tend to be very popular courses during this term, as there are several inbound arrows for these nodes (Figure 6a). The semester with the lowest centrality is the first year fall term for BME, and the network plot shows that there doesn't seem to be a central course that a large proportion of students took (Figure 6b). Some classes do tend to be popular, but not as noticeable as those in the network plot for 3rd year fall for economics.

(a) Highest Centrality	(b) Lowest Centrality



(a) 3rd Year Fall semester in Economics major

(b) 1st Year Fall semester in Biomedical Engineering major

Figure 6. Comparison of major and semester with the highest and lowest centrality. a) The major and term with the highest centrality is the third year fall term for economics majors. b) The major with the lowest centrality is the first year fall term for biomedical engineering majors.

Second, Table 2. shows the modularity of the two majors. Unlike the centrality, economics has slightly lower modularity than BME on average, where high modularity scores indicate dense connections between the nodes within clusters but sparse connections between nodes in different clusters. In other words, the higher the modularity, the more structured the major. Table 2 indicates that BME is more structured than economics, but again, modularity varies across semesters. Modularity is the highest in the first year fall semester in both majors and the lowest in the third year spring semester in both majors. For both majors, modularity gradually decreases over time. This indicates that there are dense connections between classes within clusters during the first two years and more random clusters during the last two years. This aligns with the commonly seen trend where students tend to focus on major courses initially, then take a wider variety of courses during their fourth year.

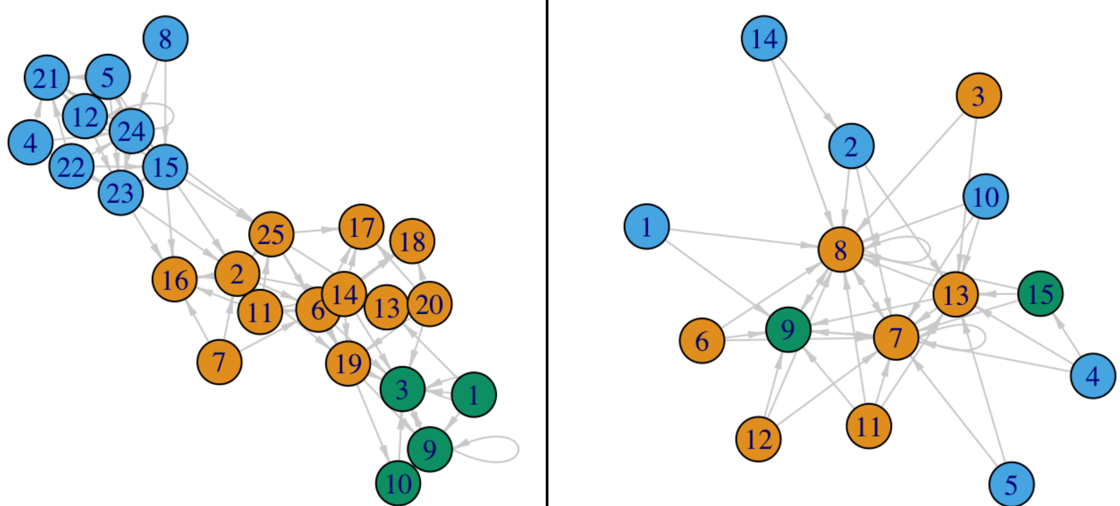
Table 2. Modularity of Economics and Biomedical Engineering across semesters.

	1st Yr Fall	1st Yr Spring	2nd Yr Fall	2nd Yr Spring	3rd Yr Fall	3rd Yr Spring	4th Yr Fall	Average
Economics	0.407	0.308	0.245	0.190	0.180	0.124	0.150	0.23
Biomedical Engineering (BME)	0.415	0.316	0.359	0.336	0.11	0.08	0.17	0.26

To further see the difference in modularity between these two majors, we compare two network plots divided into three clusters and again graph the semesters with the highest and lowest modularity (Figure 7). In each plot, one node represents a course labeled with a number. The semester with the highest modularity (Figure 7a) is the 1st year fall semester for BME, which shows clear clusters in Figure 7. The orange cluster in the 1st year fall semester for BME includes math courses such as 2 (ORD AND PRTL DIFF EQUATIONS), 16 (QUANT PHYSIOLOGY BIOSTAT APPL), and 17 (FUND OF BIOMAT AND BIOMECH). The blue cluster contains chemistry courses, such as 12 (ORGANIC CHEMISTRY) and 22 (CORE CONCEPTS IN CHEMISTRY). In contrast, the semester with the lowest modularity (Figure 7b) is the first year fall term for BME, which shows more random clusters. Courses 7 (SPECIAL TOPICS) and 8 (PROJECTS IN BIOMED ENG) are the most central in the 3rd year spring semester in BME. This imbalanced concentration may lead to interrupting the process of random walks when detecting communities in the network.

Figure 7. Comparison of highest and lowest Modularity. Appendix II includes the description of course numbers.

(a) Highest Modularity	(b) Lowest Modularity



1st Year Fall semester in Biomedical Engineering major	3rd Year Spring semester in Biomedical Engineering major

Recommendation Model

Introduction and Background

Undergraduate students often choose courses based on their own research, guidance from academic advisors, or advice from peers. However, the number of course combinations available and the variety of academic requirements are often overwhelming. Therefore, a recommendation system that uses historial course data to recommend courses could help guide students by providing more tailored course recommendations.

There are two major branches of recommendation systems: content-based filtering and collaborative filtering (CF) (Charu, 2016). Content-based filtering uses item features to recommend other items similar to what the user previously liked. However, this model will only perform well given sufficient item features in the dataset (Charu, 2016). To address some of the limitations of content-based filtering, CF uses similarities among both users and items to provide recommendations (Charu, 2016). This model recommends items to a user based on the preferences of another similar user (Charu, 2016).

In the last decade, CF has been an increasingly common method for developing recommendation systems. One major advantage of CF is that it can perform well without relying on item features. Clustering has also been one of the extensively used concepts in CF. B. M. Sarwar et al. (2002) argued that clustering improves the performance of recommendation, and P. Adamopoulos (2014) proposed a new probabilistic neighborhood-based approach as an improvement of the standard k-nearest neighbor algorithm. It is based on classical metrics of dispersion and diversity as well as on some newly proposed metrics. P. Knees et al. (2014) proposed a normalization technique called mutual proximity in the nearest neighbor selection phase to rescale the similarity space and symmetrize the nearest neighbor relation. They proved that incorporating normalized similarity values into the neighbor weighting step leads to increased rating prediction accuracy.

One of the major factors in CF that greatly influences the recommendation accuracy is the selected similarity measure. Almost all previous works are based on the well-known Pearson correlation measure. However, Pearson correlation does not take into account users' preferences. Data sparsity resulting from a large, binary, item-based matrix is another common problem that contributes to generating incorrect recommendations. In addition, using a large dataset requires more time for computing similarities among users in order to build an effective neighborhood for the active user. Moreover, in our case, class recommendations must consider the temporal nature driven by the semester system. These challenges make our problem unique, so we created a more custom cosine-similarity neighborhood method that aims to find the neighborhood where students share the same interests and take similar classes over time (Figure 4).

Neighborhood-based Methods

The cosine-similarity neighborhood method consists of two primary parts, the identification of the class-to-class relationship followed by the student-to-class relationship. Q. Liu et al. (2014) proved that combining user-based and item-based methods performed significantly better than single method (item-based or user-based). By using the proposed method, we are able to consider both the relationship between classes as well as student preferences. Specific processes are discussed in detail below.

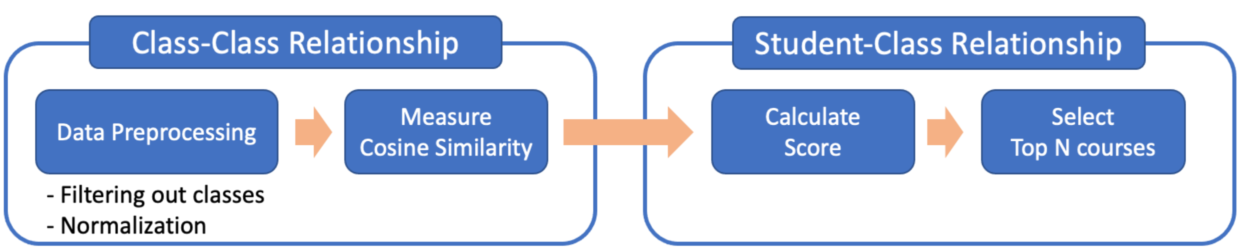


Figure 8. Brief overview of the cosine similarity neighborhood method for the course recommendation model.

Class-Class relationship

The class-class relationship measures how similar one class is to another. It looks for similar classes based on the classes that students have already taken or positively interacted with. Before measuring the relationship between classes, we preprocess our data by filtering out classes taken less than N students and normalizing student's class selection. Then, we measure the relationship between classes using cosine similarity. With these two steps, we construct a new class by class matrix containing the weights (relationships) between each of our classes where a perfect correlation between classes equals 1 and no correlation at all equals 0.

- Data was preprocessed according to the following steps:
  - **Filtering out classes taken less than N students:** Selecting N depends on the size of data. In our dataset, when N is 200, 24 classes remained among the original 400 classes, and the recommendation model performed the best. Notably, the fewer classes, the better the performance will be because the search space is smaller.
  - **Normalizing student's class selection:** The magnitude of all the student's class selection was calculated by taking the square root of the sum of the squares of all the student's class selection. This is done for all classes a student takes, noted here as *x*, *y,* and *z.*

$$magnitude = \sqrt{\{(x^2 + y^2 + z^2 + \ldots)\}}$$

We then create the new unit vector by dividing the class selection by the magnitude:

$$vector = \frac{x}{magnitude}, \frac{y}{magnitude}, \frac{z}{magnitude},,,$$

- **Measuring cosine similarity between classes:** The dot-product of the different class-vectors is then divided by the product of the normalized vectors from the previous step. We then calculate the normalized vector based on the Euclidean distance (L2-norm) of that vector, which means the square root of the sum of the absolute values squared. This is detailed in the following equation:

$$cos(\theta) = \frac{X \bullet Y}{||X|| \times ||Y||} = \frac{\sum_i X_i Y_i}{\sqrt{\sum_i X_i^2 \sum_i Y_i^2}}$$

Student-Class relationship

Given that the class-class matrix represents similarities among classes, we now need to identify students who have taken similar classes by calculating scores between each student and class. This creates a "neighborhood" of similar students defined by taking similar courses from the class-class relationship matrix. After calculating the student-class scores, recommendations can be generated by taking the top n courses with the highest score. To calculate the score, we use the following formula:

u= student
i = course
j = course taken
W = weights (0 or 1)
r = item-based matrix

$$S(u,i) = \frac{\sum_{j \in N} W_{ij} r_{ui}}{\sum_{j} |W_{ij}|}$$

The score for user u and item i is calculated by summing together all the weights for that item W_{ij} multiplied with the users rating for that item r_{ui}. We then divide by the sum of all the weights for that item W_{ij}. With this formula, we obtain a list of recommended classes. The highest score is the class that his/her most similar neighbors (students) have taken.

Performance Evaluation

Accuracy

We define accuracy as the number of courses actually taken in the top K recommendations divided by the total number of courses taken during the selected semester. For example, if provided 10 recommended courses and the student took 2 of those recommendations out of 3 total classes, then the accuracy would be 2/3 or 67%.

$$Accuracy = \frac{Number\ of\ Courses\ Taken\ in\ Top\ K\ Recommendations}{Total\ Number\ of\ Courses\ Taken\ of\ the\ Academic\ Semester\ of\ Interest}$$

We used both accuracy and modularity to gauge the performance of our model. We first compared accuracy over time from Second Year Fall term to Fourth Year Spring term. This is done to show performance for different semesters to help identify which semesters the model performs well or poorly. We do not recommend classes for the First Year, because they have no course-taking history, making this model unfit to apply to First Year students. This is known as the “cold start problem,” where recommendation systems are unable to perform well because new users have little data (Charu, 2016). We also compared modularity and average accuracy of four different but common majors, including Economics, Computer Science, Biomedical Engineering, and Political Science. This helps determine if the model performs similarly across majors. In order to test model generalization performance to more majors, we also test performance on a dataset containing students in the top 5 majors: Biology (BS), Economics (BS), Public Policy (AB), Biomedical Engineering (BSE), Computer Science (BS). Using modularity and our custom accuracy metric, we measure the performance of the neighborhood model for different semesters and majors.

Other Recommendation Models

In addition to the neighborhood method, another common set of recommendation algorithms are latent factor models (LFMs). LFMs are a state-of-the-art method for model-based CF (Charu, 2016). These models assume that there is an unknown low-dimensional representation of users and items with which the user-item affinity can be accurately modeled (Melville, 2010). Matrix factorization is a class of widely successful LFMs that attempt to find weighted low-rank approximations to the user-item matrix (Melville, 2010). In this project, we have attempted two types of matrix factorization based LFMs: singular value decomposition (SVD) and LightFM.

SVD

Singular value decomposition, SVD, is a very popular matrix factorization technique for recommender systems. However, it is most commonly used for explicit datasets where user feedback such as ratings is available (Melville, 2010). In the context of our problem, grades are used as a proxy for rating. Notably, grades are much different than course ratings, as they do not show a student's like or dislike of the class. However, in the absence of course ratings, we used grades instead. Thus, this model will tend to recommend courses where the students are most likely to receive a higher grade. SVD works according to the following equation: $R = U\Sigma V^T$, where *R* represents the predicted student grades, *U* consists of the current student grades, Σ is the diagonal matrix of singular values (weights), and V^T represents courses.

LightFM

LightFM is a hybrid matrix factorization model and combines both the concepts of content-based recommendation models and collaborative filtering (Melville, 2010). It learns the vector representations, also known as embeddings, for users and items to encode user preferences for items. Multiplying the encoded user matrix and item matrix together produces scores for every item for a given user. It is able to solve the "cold-start problem", where recommender systems struggle to make recommendations for new users due to the absence of past data. Hybrid models solve this issue by utilizing similarities based on user features to help generate recommendations for new users.

Results

Model Comparison

We compared the performance of the neighborhood model, SVD, and LightFM on the subset of data with the top 5 most popular majors using our accuracy metric. SVD and LightFM both performed consistently lower than the neighborhood-based method by about 7% (Table 3). This may be due to the limited number of course or student features in the dataset. Alternatively, the accuracy metric used to gauge model performance does not cater well to LFMs, resulting in poor performance. Furthermore, it is likely that the use of grades as a proxy for ratings could have drastically skewed recommendations. Lastly, it was impossible to analyze the vectors created in the latent space, making it difficult to determine why these models performed poorly.

Table 3. Average accuracies of neighborhood model, SVD, and LightFM, on the subset of data with the top 5 most popular majors when providing 15 recommendations.

k = 15	Second Year	Third Year	Fourth Year	Average
Neighborhood	14.4%	30.1%	25.2%	23.2%
SVD	13.9%	24.3%	17.4%	18.5%
LightFM	13.6%	11.3%	19.4%	14.76

We have also compared the advantages and disadvantages across the three methods shown in table 4. Though SVD and LightFM present some advantages over neighborhood methods, implementation and testing of these models revealed that the neighborhood model with cosine similarity performed much better than the other two models. Thus, we chose the neighborhood model as the final recommendation algorithm.

Table 4. Comparison of the three recommendation models.

	Neighborhood method	SVD	LightFM
Pros	● Much higher prediction accuracy ● More consistent performance	● Grade prediction ● Scalable ● Effective for sparse data	● Incorporate metadata of courses and students ● Able to solve cold start problem
Cons	● Requires large offline storage	● Very low prediction accuracy	● Very low prediction accuracy

Neighborhood Model Results

Using the neighborhood-based recommendation model, we compared the performance of the original dataset for economics majors with the performance of the dataset after filtering out less popular classes. After filtering out less popular classes, this model performed an average of 26.8% better than generating recommendations from the original dataset (Table 5). In every semester except Fourth Year Spring term, the model performed significantly better with the filtered data than with the unfiltered data (Table 5).

Table 5. Comparing accuracies of the filtered and unfiltered data for the neighborhood-based recommendation model for economics majors using 10 recommendations. The original data is unfiltered, while the filtered data is the subset where courses taken by less than 200 students from 2013 to 2020 were removed. This helps decrease the search space for classes and thus improves model performance.

k=10	2Y Fall	2Y Spring	3Y Fall	3Y Spring	4Y Fall	4Y Spring	Average
Original	25.4%	22.9%	9.1%	6.5%	14.4%	46.2%	20.8%
Filtering out less popular classes	60.8%	49.7%	51.1%	50.7%	38.6%	33.8%	47.6%
Change	+35.4%	+26.8%	+42%	+44.2%	24.2%	-12.4%	+26.8%

Next, comparing performance among the four different chosen majors, Computer Science performed the best with an average accuracy of 51.2%, and Political Science performed the worst at 45.3% (Table 6). The overall performance for the four majors was relatively similar. Performance for the subset of data with the top 5 majors was slightly lower at 38.2% (Table 6). These results indicate that the model performs better within a specific major rather than generalizing to a combination of majors.

To further compare majors and potentially identify reasons behind the difference in accuracy among majors, we calculate modularity for each major. The modularity of Computer Science is 0.4 and the modularity of Political Science is 0.3 (Table 6). Though we cannot determine if there is a causal relationship between modularity and accuracy, we see a clear positive relationship between them. This relationship indicates that majors with higher modularity may have higher accuracy and be more fit for the recommender system. For example, Computer Science majors may have clearer academic trajectories than other majors, which leads to higher modularity and thus better prediction accuracy than other majors. This concept could be further tested in the future.

Table 6. Comparison of average accuracy and modularity for four majors of choice and the subset of top 5 most common majors.

k=10	Economics	Computer Science	Biomedical Engineering	Political Science	Top 5 Major
Average Accuracy	47.6%	51.2%	50.7%	45.3%	38.2%
Modularity	0.32	0.40	0.38	0.30	0.25

Discussions and Limitations

Though the cosine similarity neighborhood method has relatively higher prediction accuracy than other methods, there are several limitations. First, the cosine similarity neighborhood method cannot consider multiple features such as grade. Second, since it needs to compute a very large item-item matrix, it is very expensive and inefficient to compute and recommend courses in actual applications, especially if it is used for real-time recommendation.

In terms of measuring model performance for recommender systems, it is difficult to gauge which evaluation metrics are fit for the problem, and thus, which models perform better. In practice, recommender systems can be assessed using A/B testing, but that is not feasible in our case. Furthermore, just because a student did not take a class, it does not mean that they did not like it. Likewise, if the student did end up taking the course, they still may not have liked it. These complications make it extremely difficult to determine if a recommendation engine is performing well. In our project, we chose to use a custom accuracy metric because we had data on the classes the students actually ended up taking. However, future work could implement other performance metrics or even A/B testing to better gauge model performance.

In addition, there is a limited number of courses students can take each semester. For example, the average student takes four courses each semester. Because our model attempts to accurately capture these 4 courses, it is relatively difficult to achieve a high accuracy given these minimal data points. Furthermore, students tend to lighten their load in their senior year and frequently take fewer classes. This means that using our accuracy metric, we expect a low overall accuracy. Lastly, there is an extreme amount of variability in course-taking patterns even amongst students of the same major. This makes it difficult to accurately capture such a wide range of classes. However, many of these challenges are innate to the problem, and despite these limitations, this neighborhood-based recommendation engine consistently performed better than other recommendation models tested.

Future Work

To further aid undergraduate students in making more informed decisions for major and course selection, additional data, including course descriptions, syllabi, professors, prerequisites, major requirements, and timing should be gathered and integrated into the current dashboards. To better track academic trajectories, an additional page should be added to the dashboards that presents course progression information in a more ingestible manner for students and advisors. This could show common courses to take next or courses that are commonly taken before higher level courses to help students plan course order. Future implementations could also identify combinations of classes that are commonly taken together leading to higher or lower grades, which could help with course load planning. Currently a prototype, the course and major planning dashboards can be further improved and fine tuned to be fully integrated with the current DukeHub registration system.

To improve the performance as well as the practical applications of the recommendation model, one potential direction is to investigate other representations of courses using advanced techniques, such as word embeddings, to identify cluster structures and similar classes across different domains. This could potentially provide new information that could improve model performance. To better evaluate model performance, other metrics such as relevance and diversity metrics in addition to our customized prediction accuracy could be used to better gauge the strengths and weaknesses of each model.

Overall, there are several improvements and additions that could be made to both the dashboards as well as the recommendation engine. However, our current work shows promising potential for the usability for a course planning tool, and future work can continue to build on these deliverables to develop a more thorough and robust tool.

Conclusion

Working towards the goals of tracking academic trajectories and developing a course recommender engine, we have successfully created three major deliverables. First, the integrated planner for major and course selection is easily accessible, user-friendly, and ingestible for both students and advisors. It allows for extensive filtering based on students' course history and personal preferences, which is especially useful for first year students who may not have a clear direction. Second, the visualizations of academic trajectories, including the network and alluvial plots, help identify curriculum flows for each major, while highlighting common prerequisites for courses between consecutive semesters. Lastly, the course recommender system provides more personalized suggestions by finding the neighborhood where students have taken similar classes over time. All of these products serve to provide more information to facilitate the major and course selection process for both the undergraduate students and the academic advisors at Duke University.

The lack of recommendation systems in the academic space as well as the clear need for increased guidance in course and major planning for undergraduate students emphasizes the potential for a course and major planning tool. This tool enables students to easily filter through the massive amount of classes and majors, choose classes that align with their desired pathways, and better plan their undergraduate curriculum. Advisors could also benefit from these products to better understand course-taking patterns as well as guide students through their undergraduate careers. Overall, there are several clear benefits in developing a system to aid students in course and major selection.

References

1. B. M. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Recommender Systems for Large-scale E-Commerce: Scalable Neighborhood Formation Using Clustering," in Proceedings of The fifth international conference on computer and information technology, 2002.
2. Charu, C. A. (2016). *Recommender Systems: The Textbook*.
3. Few, S. (2006). *Information dashboard design: The effective visual communication of data* (Vol. 2). Sebastopol, CA: O'reilly.
4. Jiang, W., Pardos, Z. A., & Wei, Q. (2019, March). Goal-based course recommendation. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 36-45).
5. Lee, Y., & Cho, J. (2011). An intelligent course recommendation system. *SmartCR*, *1*(1), 69-84.
6. P. Adamopoulos, "On Discovering non-Obvious Recommendations: Using Unexpectedness and Neighborhood Selection Methods in Collaborative Filtering Systems," in Proceedings of the 7th ACM international conference on Web search and data mining, New York, NY, USA, 2014.
7. P. Knees, D. Schnitzer and A. Flexer, "Improving Neighborhood-Based Collaborative Filtering by Reducing Hubness," in Proceedings of International Conference on Multimedia Retrieval, New York, NY, USA, 2014
8. Main, J. B., & Ost, B. (2014, January 24). The impact of letter grades on Student Effort, COURSE selection, and Major choice: A Regression-Discontinuity Analysis.
9. McFarland, D. A. (2006). Curricular flows: Trajectories, turning points, and assignment criteria in high school math careers. Sociology of Education, 79(3), 177-205.
10. Melville, P., & Sindhwani, V. (2010). Recommender systems. *Encyclopedia of machine learning*, *1*, 829-838.
11. Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, *57*(10), 1500-1509.
12. *Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical review E, 69(2), 026113.*
13. *Friedkin, N. E., & Thomas, S. L. (1997). Social positions in schooling. Sociology of Education, 239-255.*
14. *Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications.*
15. *Chase, I. D. (1991). Vacancy chains. Annual review of sociology, 133-154.*
16. *Sørensen, A. B. (1987). The organizational differentiation of students in schools as an opportunity structure. In The social organization of schools (pp. 103-129). Springer, Boston, MA.*
17. *Liu, Q., Xiong, Y., & Huang, W. (2014). Combining user-based and item-based models for collaborative filtering using stacked regression. Chinese Journal of Electronics, 23(4), 712-717.*
18. *Crucitti, P., Latora, V., & Porta, S. (2006). Centrality measures in spatial networks of urban streets. Physical Review E, 73(3), 036125.*
19. *Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications.*
20. *Scott, J. (1988). Social network analysis. Sociology, 22(1), 109-127.*

Appendix I

Data Dictionary

- Calculation ID: anonymized student ID
- Acad Plan: shorthand/abbreviation for major/minor/certificate/secondary
- Acad Plan Descr: description of acad plan (major/minor/certificate/secondary)
- Plan Type: denotes major/minor/certificate/secondary
- Degree: abbreviated type of degree (AB, BS, BSE)
- Descr_completions: type of degree (Bachelor of Arts, Bachelor of Science, Bachelor of Science in Engineering)
- Comp Term Descr: academic year and term of graduation
- Acad Year_completions: academic year of graduation
- Subject: subject of course
- Catalog: course number
- Descr_enrollment: name of course
- Grade: grade received in descr_enrollment
- Term Descr: academic year and term of enrollment in class
- Acad Year_enrollment: academic year of course enrollment
- Enrollment Start: first year student has registered for classes at Duke
- Term Year: year student is taking course
- Semester Term: term student is taking course (Fall, Spring, Summer 1, Summer 2)
- Class Year: indicates freshman through senior year (or more) and term (First Year Fall Term, First Year Spring Term, etc.)

Appendix II

1. First Year Fall Semester for BME

1	INTRO BIOCHEMISTRY I
2	ORD AND PRTL DIFF EQUATIONS
3	PROJECTS IN BIOMED ENG
4	COMP METH IN ENGINEERING (LEC)
5	MULTIVARIABLE CALCULUS
6	MODELS CELL AND MOL SYSTEMS
7	INTRO ELECTRIC, MAGNET, OPTICS
8	SPECIAL TOPICS
9	ADVANCED TOPICS
10	DESIGN DEVELOPING WORLD
11	MECHANICS OF SOLIDS
12	ORGANIC CHEMISTRY
13	TRNSPRT PHENOM:BIOLOGCL SYSTMS
14	STRUC/PROP OF SOLIDS
15	MOLECULAR BIOLOGY
16	QUANT PHYSIOLOGY BIOSTAT APPL
17	FUND OF BIOMAT AND BIOMECH
18	MOD DIAG IMAGING SYSTEMS
19	INTRO MEDICAL INSTRUMENTATION
20	SIGNALS AND SYSTEMS
21	INTRODUCTORY MECHANICS
22	CORE CONCEPTS IN CHEMISTRY
23	LINEAR ALGEBRA & DIFF EQUATION
24	ACADEMIC WRITING
25	FUND OF ELEC AND COMP ENGR

2. 3rd Year Spring Semester for BME

1	INTRO BIOCHEMISTRY I
2	SIGNALS AND SYSTEMS
3	ORD AND PRTL DIFF EQUATIONS
4	MECHANICS OF SOLIDS
5	MOLECULAR BIOLOGY
6	FUND OF BIOMAT AND BIOMECH
7	SPECIAL TOPICS
8	PROJECTS IN BIOMED ENG
9	DESIGN DEVELOPING WORLD
10	MODELS CELL AND MOL SYSTEMS
11	TRNSPRT PHENOM:BIOLOGCL SYSTMS
12	MOD DIAG IMAGING SYSTEMS
13	INTRO MEDICAL INSTRUMENTATION
14	FUND OF ELEC AND COMP ENGR
15	STRUC/PROP OF SOLIDS