Identifying Students' Progress and Mobility Patterns in Higher Education Through Open-Source Visualization

Ali Oran
Brigham & Women's Hospital, aoran@ieee.org

Andrew Martin, Michael Klymkowsky, Robert Stubbs
University of Colorado Boulder, andrew.martin-1, michael.klymkowsky, robert.stubbs@colorado.edu

Abstract - **For ensuring students' continuous achievement of academic excellence, higher education institutions commonly engage in periodic and critical revision of its academic programs. Depending on the goals and the resources of the institution, these revisions can focus only on an analysis of retention-graduation rates of different entry cohorts over the years, or survey results measuring students' level of satisfaction in their programs. They can also be more comprehensive, requiring an analysis of the content, scope, and alignment of a program's curricula, for improving academic excellence. The revisions require the academic units to collaborate with university's data experts, commonly the Institutional Research Office, to gather the needed information. The information should be highly informative yet easily interpretable, so that the review committee can quickly notice areas of improvement and take actions afterwards. In this study, we discuss the development and practical use of a visual that was developed with these key points in mind. The visuals, referred by us as "Students' Progress Visuals", are based on the Sankey diagram and provide information on students' progress and mobility patterns in an academic unit over time in an easily understandable format. They were developed using open-source software, and recently began to be used by several departments of our research intensive higher-ed institution for academic units' review processes. Our discussion includes questions these visuals can address in Higher-Ed, other relevant studies, the data requirements for their development, comparisons with other reporting methods, and how they were used in actual practice with actual case studies.**

Index Terms - Educational Data Mining, Learning Analytics, Open-source Data Visualization

INTRODUCTION

I. Motivation and The Academic Review Processes

To ensure academic excellence higher education institutions commonly engage in periodic and critical revision of their academic programs. While the names describing these efforts vary slightly among institutions (Academic Review and Planning at University of Colorado Boulder [1], Program Review and Academic Program Review in other institutions), the revision process and its objectives are very similar: A regular review of colleges, schools and academic units to identify academic program strengths and weaknesses and to provide constructive options for program development and modification" [1]. The review efforts include review committees that can be comprised not only of campus constituents but also of disciplinary experts external to the institution. The efforts commonly begin with academic units engaging in self-studies during which they address a series of planning queries to solicit strategic information and to document the unit's organizational qualifications. Topics include role and mission, centrality, outcomes, and diversity goals [1]. In this information gathering phase, academic programs also collaborate with other departments, such as Institutional Research, to collect the necessary information. Following that, the internal and external reviews commonly follow. And in the final phase, recommendations for unit improvements are proposed addressing the identified issues.

The initial phase of information gathering is particularly critical because subsequent analyses and conclusions are based on that information. Depending on the strategic goals, constraints and resources of the institution and its academic units, this step can incorporate various analyses. Common approaches include the analysis of retention and graduation rates of different entry cohorts over the years, along with surveys measuring students' level of satisfaction with their programs. More comprehensive analyses may involve examination of the content, scope and alignment of programs' curricula, assessment of the impact of specific courses (and instructors) on retention and predicting the likelihood of students encountering courses characterized by high levels of failure. Ideally, the collected data, as well as its presentation, should provide clear and easily interpretable information to an academic unit to initiate productive discussions in the succeeding steps about central issues impacting students' educational experiences. However, such efforts can be derailed if the provided information fails to convey the important patterns in the data to faculty and other decision-makers (administrators and departmental curriculum committees). Accordingly, in this phase, the query of the needed data sources, the assembly of the correct amount of information and developing accurate and easily

understandable metrics for the faculty and administrators would be required to ensure the succeeding discussions could focus on the needed areas of improvement.

The sheer amount of information being analyzed can be overwhelming. Characterizing the diverse body of students following very different paths after admission can be challenging. Hence, there is a continuing need for highly informative yet easily interpretable methods that can reflect the diverse student progress characteristics in academic units. How can we present insights from vast data sources for a maximum impact? How can we make its implications more readily discernible to Higher-Ed decision makers? In this study, we discuss our efforts to answer these questions by developing a visualization-based approach, which can be ideal for conveying complex data patterns.

II. Related Work

Following the recent advances in data science field there has been a considerable interest for utilizing new data techniques in High-Ed research over the past two decades [2]. A very diverse group of studies have been proposed touching different issues [see 3–5]. These new approaches have the potential to allow institutions to harness large campus-wide data sources to identify areas of possible improvement and for making data-informed decisions, e.g., optimization of daily operations, improving student engagement and learning outcomes.

The interest in the use of new data techniques is heightened by the growing urgency to offset some of Higher-Ed's recently realized challenges. The most pressing challenge has been the decline in state funding to higher education in the past couple of decades, particularly during the Great Recession. While state appropriations have increased since the low point of 2012, as of 2017, only six states have reached or surpassed their pre-recession levels in 2008, as reported in the State Higher Education Finance Report by The State Higher Education Executive Officers (SHEEO) [6]. Another challenge has been the changing patterns in enrollment numbers in the past decade. While during the Great Recession Higher-Ed institutions in general saw continuous increases in enrollment numbers, since 2011 these numbers have been decreasing in general, as reported in studies by SHEEO, and by National Student Clearinghouse Research Center (NSCRC) [6–8]. At the same time, there has been an increasing competition with educational institutions from other developed nations in attracting international students. Compared to the early 2000s, other countries, particularly Canada and Australia, have become educational destinations for a larger percentage of international students [9] and this may become an issue for some US higher education institutions that have relied on foreign students' out-of-state tuition. Lastly, as a new generation of diverse students are entering our universities, possible factors that may be hindering their success needs to be carefully identified. In these times of changing enrollment patterns, a better use of campus-wide data sources through analytics has the potential to provide decision makers valuable insights.

Recent studies that have focused on improving students' academic outcomes can be categorized according to their approach under two major groups: "Descriptive Methods", and "Predictive Methods". Descriptive methods are the studies whose approach is based on accurate characterization of students' academic performances (through descriptive statistics) and seeking solutions by evaluating those characteristics. They have been used for years to identify possible issues, and subsequently to take necessary action when necessary, such as reshaping entering classes, refining policies and course requirements etc. Predictive Methods have become more widespread in recent years in many Higher Ed institutions [9]. Among the predictive methods, Early Warning Systems (EWS), aiming to identify students who might have a high likelihood of academic failure by harnessing campus wide data sources, are one of the most known group of studies [10]. Purdue University (Course Signals) [11], University of Phoenix [12], and Capella University [13] are just a few examples of the universities that have utilized such systems. Logistic Regression has been a common method for the predictions [12, 13], yet, more advanced methods involving Machine Leaning have been tried as well [14]. While predictive methods have been gaining more attention recently, the difficulty of accurately incorporating qualitative factors, such as student motivation and persistence (both of which influence student success) is still a major limitation.

Among Descriptive Methods, given the nature of human cognition, visual methods can be more effective for conveying the results of the analysis than other methods, such as written narratives or numerical tables. Among Visual Methods, a distinction can be made according to the methods' focus. Some visuals focus on visualizing students' data in an aggregate way (Aggregate Visuals), while others focus on displaying each student's data separately (Tracking Visuals). Tracking visuals can visualize characteristics of each student, e.g., visuals for a student's course work over consecutive terms for tracking the student's progress. Tracking visuals have seen a wide usage particularly in on-line education, where student-instructor interaction is quite different than the traditional campus-based institutions. In such environments, they can provide a good summary of critical information to students for adjusting their study practices, and to instructors for reaching out to students at necessary times. Mazza and Milani's GISMO [15], Bakharia and Dawson's SNAPP [16], Capella University's Competency Map [17] are some of the best known examples of this group of tracking visuals.

In contrast, Aggregate Visuals can summarize the overall characteristics of cohorts, e.g., average time-to-degree visuals of Chemistry majors accepted between 2010 to 2015, and accordingly can be quite useful in detecting possible issues within academic units. Among Aggregate visuals, traditional data visualization methods (such as line plots, pie charts etc.) have been prevalent in Higher-Ed. Recent advances in visualization software mean that newer visualization techniques have the potential to convey more information more impactfully to decision makers. In this context, for understanding students' progress, Flow Diagrams are one of the promising new alternatives. A Flow Diagram (or Chart) visually displays interrelated information such as events, steps in a process in an organized fashion,

such as sequentially or chronologically. Among Flow Diagrams the Sankey Diagram is ideal to visualize measurable processes. Its primary advantage stems from its visualization of the flows of a process using lines with variable thickness which are proportional to the magnitude of the flows. Accordingly, it has been widely used in a very diverse group of fields.

However, the use of Sankey visuals has been limited in Higher-Ed. To the best of our knowledge, the first study to use Sankey Diagrams in Higher-Ed was in early 2014 by Orr et al., who analyzed the origins and destinations of students enrolled in a Mechanical Engineering program using a simple 2-column Sankey diagram [18]. Later in the same year, Morse, in his Master's thesis [19], proposed a multi-column Sankey diagram for visualizing students' progress. Morse's work was followed in Heileman et. al. in 2015 [20] to analyze a hypotheses, termed myths, about students' success. These were followed by two studies in 2018 by Horvth et al. [21], and Basavaraj et al. [22] to understand students' progress in their institutions.

Following Morse's work, an informative visualization like Sankey might have been expected to be embraced by institutions. Yet, it has not been much noticed, and by those who have noticed it is still often considered an experimental approach. Hence, more discussion of this new visual is needed by reviewing how they can be used in Higher-Ed institutions, the questions they can answer, a comparison with other visual methods they can be replacing, and possible software requirements for their development. In this study we attempt to fill this need by providing a discussion on these questions by summarizing our own experiences in choosing to develop Sankey visuals for conveying students' progress patterns to a group of departments undergoing academic review. We start our discussion by detailing the core questions we originally aimed to answer before choosing to develop Sankey visuals.

Problems of Interest in Students' Success

A critical aspect of a typical undergraduate degree program is the extent to which students can successfully complete its requirements within a reasonable time. Unfortunately, a recent study by American Academy of Arts and Sciences [23] indicates that too few students graduate and too few graduate in a timely manner. Accordingly, in the context of academic program review, discussions often center around retention and graduation metrics, and the initiatives to improve them. A number of obstacles impact retention and graduation rates, such as budget cuts that reduce available resources for student advising, curricular plans that impose barriers to students' progress, and instructor and course effects that lead to academic "bottlenecks". Operational and curricular differences among departments require us to analyze these factors on the level of academic degree programs.

Ideally, such an analysis should be able to reveal the differing characteristics of the group of students who fail and who graduate. With students having multiple paths into a major (e.g. starting in the major, or starting in another major and later moving to the major), and also multiple paths out of a major (e.g. leaving to another major or leaving the institution altogether) the analysis also needs to distinguish these separate entry and leave groups. These diverse student pathways can be brushed aside as just an example of young students testing out different majors, and accordingly showing random mobility patterns between majors. Yet, the real reasons may be quite different - a program may require courses that fail to engage students or are badly designed or delivered. While some might see such courses as "rites of passage", they can also be seen as "gatekeepers"; their elimination or reform could enhance student retention and success. Also, some instructors' teaching methodology may be contributing to a loss of students from a major, especially in "gateway" courses. These possible issues can be approached by analyzing origin-destination majors. Students switching to very different disciplines from their original majors (e.g. STEM major to humanities) will display a different pattern compared to students who switch between similar majors or to those who leave the university altogether. In addition, origin-destination major analysis can identify majors that are more welcoming to students, and those that do not allow for such transitions. As students change majors in any academic term. the analysis should be able to convey time-dependent patterns as well. A term-wise analysis of students' entry and departure patterns can yield essential information about students' satisfaction and the problems they may be facing at a particular term - e.g., the underlying reasons for students leaving their major within their 1st year are likely to be quite different from those of students leaving later on. Lastly, considering the diverse student bodies higher-ed institutions have been welcoming, the analysis should be able to distinguish the student groups most affected by departmental barriers. In this regard, there will be a need for stratification of student data to detect those at-risk student groups. In summary, for improving students' success in an academic unit (and within its offered majors), our analysis will seek answers to the following questions:

1. Who are the students leaving/entering the major?
2. When are the students leaving/entering the major?
3. Where are the students leaving to/entering from?

Identifying student progress patterns

I. General Approach

An academic unit's student population continuously changes at each term. For analyzing such a system, a longitudinal study, in which a cohort is followed for some time, in order to understand cohort's specific characteristic, will be convenient. We define the cohort of students according to their majors (or their affiliated academic unit) and their entry year to the institution. For identifying the stumbling blocks within an academic unit, we need to understand the progress of all the students' who were ever affiliated with that unit. Accordingly, when defining the cohort of a major we include both the students who originally start at that major and the students who start in another major but who later become part of that major. For instance, Biology-2010 cohort will refer to a group of students who have entered the university in 2010-Fall semester and have majored in Biology at some semester after that. This cohort definition will allow us to answer the "When" question, by analyzing cohorts'

characteristics over the subsequent terms after entry. For answering "Where" students come from and go to, we keep track of students' end-of-term majors at each term. Lastly, we split the cohort into subcohorts to answer the "Who" question to observe the progress differences among students in the cohort. Here, the defining (splitting) criteria for the subgroups depends on the analysis sought. For instance, splitting cohorts according to gender and race can allow for an analysis of possible progress disparities among race and gender groups; or splitting according to entry-major characteristics can reveal possible disparities among those. In our study, with our interest in analyzing origin-destination majors, we defined the subcohorts of a cohort based on students' entry-majors, which we defined as: Open option entry, Original major entry, and Other major entry. According to this definition, the Biology-2010 cohort would have three subcohorts: Biology-2010 with open option entry, Biology-2010 with Biology entry, and Biology-2010 with another major entry.

Note that the choice of subcohorts eventually affects the complexity of the final presentation. The number of different paths (different majors) students take after admission increases at each term and thus becomes a limiting factor for conveying all end-of-term major patterns for each subcohort in an interpretable format. This particularly becomes challenging when analyzing academic units with large numbers of students. Our choice of defining three subcohorts and grouping all end-of-term majors other than the original major into one category, "Other major", was selected in order to manage this issue. Based on our experiences with increased complexity, we established the following guidelines to ensure that our final analysis would be:

1. **Informative**: The analysis should contain enough information to answer questions of interest.
2. **Interpretable**: The analysis should be easily interpreted by people not necessarily working with data.
3. **Scalable**: The analysis should be scalable so that it could be reproduced easily for different size data sets for comparing different cohorts and departments.
4. **Easy to generate**: The analysis' development and succeeding updates should be as cost-effective as possible in terms of finance and time.

Our longitudinal study is then carried out in three steps: Data Extraction, Data Analysis and finally Data Presentation, which involved developing the best method, in terms of informativeness and interpretability, for conveying the data patterns found in the analysis step.

II. Data Extraction and Analysis of Students' Progress

Data Extraction step involves the use of standard database manipulation techniques, e.g. joining of data-tables, and filtering of the data, to acquire the termwise data sets for the cohorts of interest. As discussed in the previous section, we keep track of end-of-term majors for each student in the cohort for each term until graduation or leaving the institution. Accordingly, in the most minimal sense, the cohort data required the following variables. Cohort Data: [Student-ID, Year-Term, End-of-Term Major, Degree Date (if graduated), Degree Major (if graduated)]. As our analysis looks into subcohorts based on entry-majors, in addition to the variables above, a student's entry-major is also needed.

Once this cohort/subcohort data is prepared, the Data Analysis step follows. For student's progress analysis, in the most simplest form, this boils down to analyzing students' end-of-term majors at each academic term. More complex analyses can also be incorporated, e.g., analysis on course grades, or race/gender representation in cohorts, as long as the complexity of the presentation can be managed. Even when analyzing end-of-term majors, to avoid growing complexity because of numerous different majors, we grouped student's majors. The grouping was done similar to subchorts, i.e., we grouped majors according to being the original major or not. These groups were also defined separately for students who have graduated. A fifth group was defined for students who have left the university (without a degree). Overall, each student in a subcohort was placed into one of the following groups at each semester:

1. Actively seeking a degree in the original major
2. Actively seeking a degree in an other major
3. Graduated in the original major
4. Graduated in another major
5. Left the university

To illustrate how these concepts work in practice, we provide a simple example in Table I, a sample cohort of 10 students with random studentIDs, starting in Fall-2015 in the same major (since all students are starting in the original major, this cohort consist of only one subcohort). Initially in Fall-2015, all the students are actively seeking a degree in the original major, hence they are in the first group. At the end of each term, students' status may change: leaving the original major for another major, leaving the university (without a degree), graduating, or even returning to their original majors. Accordingly, in Table I, over several terms, student IDs are distributed according to students' status by the end of each term, with group numbers as defined above.

After these groups are identified for each student in the cohort, we then apply an aggregate analysis, which yields information about the general characteristics of the cohort. We are primarily interested in the number of students belonging to each group at each term, so that the change of those numbers over time reveals student mobility patterns in and out of a major and the university, Table II summarizes the aggregate analysis on our sample data from Table I. Again, other types of analyses can be pursued, such as average GPA or race/gender representation in each group.

TABLE I
A 10-STUDENT SAMPLE COHORT AND STUDENTS' GROUPS OVER TIME

	Fall15	Spr16	Fall16	Spr17	Fall17	Spr18	Fall19
Grp.1	98231	98231	98231	98231	98231	98231	98231
	84543	84543	84543	84543	23834	18105	18105
	23834	23834	23834	23834	18105	56520	56520
	18105	18105	18105	18105	56520	49428	78625

	56520	56520	56520	56520	49428	78625	
	49428	49428	49428	49428	78625		
	78625	78625	78625	78625			
	24666	24666					
	57640	57640					
	12227						
Grp.2		12227	12227	12227	12227	12227	12227
			24666	24666	24666	24666	24666
			57640	57640	57640		
Grp.3							
Grp.4							49428
Grp.5					84543	84543	84543
						23834	23834
						57640	57640

TABLE II
AGGREGATE ANALYSIS FOR GROUP SIZES FOR THE SAMPLE DATA

	Fall15	Spr16	Fall16	Spr17	Fall17	Spr18	Fall19
Grp.1	10	9	7	7	6	5	4
Grp.2	0	1	3	3	3	2	2
Grp.3	0	0	0	0	0	0	0
Grp.4	0	0	0	0	0	0	1
Grp.5	0	0	0	0	1	3	3

III. Visualizing Cohort Progress Patterns

After the analysis step, one needs to present the findings in the best way so that the patterns can be accurately and easily perceived, and their implications be discussed. Table II is one option; it summarizes the number of students in each group over time and may be ideal for analyzing short time spans of a few terms, but it may not be effective when presenting long term longitudinal data. In addition, while a table like Table II presents the actual numbers in each group (or the relative numbers), providing actual and relative numbers together will require either more rows or columns. This may not be an issue for a small 10-student cohort, but for large cohorts analyzed for long time spans such tabular data can be difficult to digest. Accordingly, some graphical technique will be needed at least as a supplement, if not as a replacement, for these tables.

Two traditional graphical techniques for presenting such cohort patterns are Line Charts and Stacked Bars, shown in Figures 1 and 2 respectively for our sample data set. Both clearly show the number of students in each group and can convey the general patterns of student mobility with different colors representing different groups and the y-axis providing information about these groups' relative sizes compared to the original cohort. By having bars of constant height, the Stacked Bar Chart provides an easier to understand visual for conveying the relative size of each group, when compared to the Line Chart. Yet, what is essentially missing in Table II, Figures 1 and 2 is the flow information about students moving among different groups. In our sample set, between Fall 2017 and Spring 2018 we notice that two students from the original cohort have left the university. Yet, it is unclear from the table or the figures whether these students left the university after having studied only in their original major, or after having switched majors (i.e., after studying in other majors). This missing flow information is valuable because it helps us understand how student cohorts interact with their programs, and how those interactions change over time. A similar question can be asked about the graduates appearing as purple patterns in Fall 2018. Again, the table or the figures do not provide information about students' graduation majors, that is are they graduating from their original majors, or from other majors. By going back to the raw data in Table. I, by comparing student IDs, one can find out the answers: one student has left the institution from the original major, and the other student from another major, that is after trying a different major; the graduate in Fall 2018 graduated from their original major.

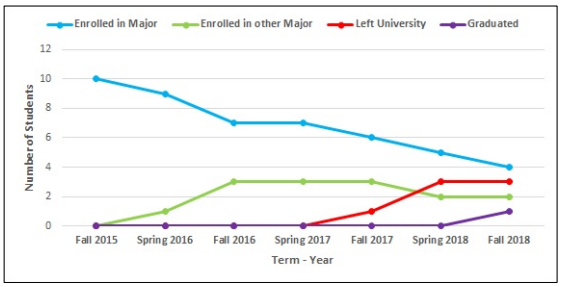

FIGURE 1
A LINE CHART FOR VISUALIZING THE SAMPLE COHORT'S PROGRESS.

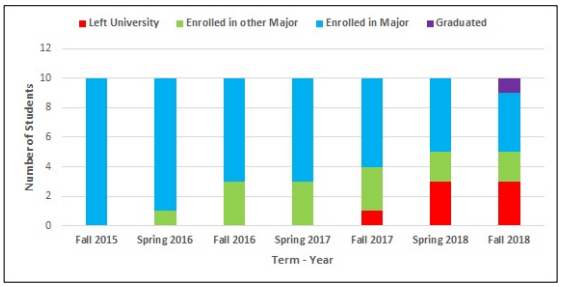

FIGURE 2
A STACKED BAR CHART FOR VISUALIZING THE SAMPLE COHORT'S PROGRESS.

Ideally, the final presentation should contain enough information that going back to analyzing the raw data will not be necessary for the decision makers. This could be done by adding more layers (groups) to these visuals or the table. For instance, the group-5, Left University, could be expanded to have multiple subgroups according to students last majors. Yet, adding more stratification will make both

the table and the figures more complex. In addition, even with more groups, one would still need to keep track of the changes in the number of students in each group to understand the flow of students from one group to another. This effort of trying to keep track of the changes in the number of students between different groups over several semesters would be an incredible drag (better wording?). One could think about enhancing tables and figures by superimposing the flow information on them. Yet, superimposing another layer of information on top of another figure or table can easily double the time needed to produce them.

One should note that a stacked bar chart with superimposed flow information is in fact a primitive Sankey diagram. Accordingly, one can use Sankey-based visuals from the beginning to avoid reproducibility issues while still including the flow information. With this approach, the sample 10-student cohort's progress can be visualized as in figure 3. This figure is similar to the Bar Chart in figure 2 in that each column represents the cohort data at a particular semester, and from left to right we see the changes in the student cohort as time progresses. The extra information are the lines connecting these columns, whose thickness represents the number of students moving from one group to another at the end of each semester. With this extra layer of information, it becomes possible to convey a cohorts' progress over time in a clearer manner, and accordingly identify the bottlenecks, e.g., a large flow of students out of a major after a particular term can indicate the effect of a course (or an instructor) associated with students' decision to change major. Now that we have introduced the Sankey-based students' progress visuals, we proceed to discuss how this type of visuals were used in practice at our institution in the academic review process.

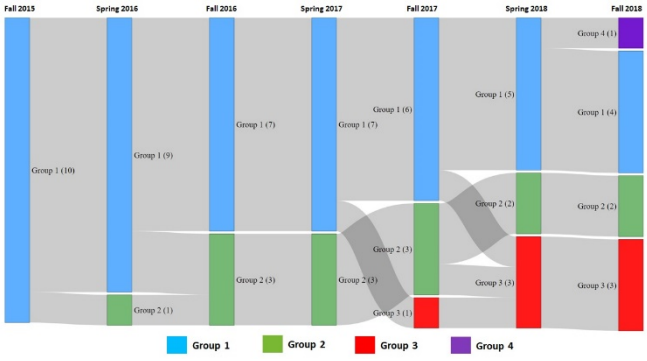

FIGURE 3
SANKEY VISUAL FOR VISUALIZING THE SAMPLE COHORT'S PROGRESS

IV. Sankey diagrams in the context of Academic Review

As discussed, a primary purpose of the Academic Review process, is to initiate productive discussions about central issues impacting students' educational experiences. Accordingly, it requires information clearly presented so that it facilitates the framing of further questions and the shaping of efforts to improve students' educational outcomes. To help achieve this goal the Institutional Research (IR) office developed Sankey-based Student Progress Visualizations for use in the program review process. The development was carried out in close collaboration with some of the faculty members involved in the review process. Having a good balance of informativeness and understandability, these visualizations provided the faculty, reviewers and administrators a compact, all-in-one overview of an academic unit in terms of students' progress. In addition, the standardized Sankey diagrams allowed faculty to compare their students' progress directly with students from other academic units to highlight similarities and differences. Following their positive reception, these Student Progress Visualizations have become one of the standard visual tools provided to departments for the 2019-2020 academic year. The IR office has also started providing customized (stratified) versions in response to specific requests from faculty and administrators. These stratified visuals show the progress of different cohorts of students within a degree program. So far, we have developed stratified Student Progress Visuals for conveying the differences of progress among male/female students, 1st-generation/non-1st-generation students, and underrepresented-minority/non-under-represented-minority students.

Figure 4 shows student progress visuals for two Natural Science departments, reflecting the progress and mobility of two undergraduate cohorts admitted into the departments in the same year for the next 6 years. The three different colored bars on the left are the three subcohorts we defined earlier. We used blue-shades for students who were enrolled at the department major, green-shades for students who were not enrolled in that particular major but in another major, and red-shades for students who have left the university. In these departments, we noted that double majoring students were a considerable group. Hence, we added another blue-green mixture color group to represent those students' progress. We used darker blue and green colors for students who had graduated in their respective groups, i.e., dark-blue for students having graduated from the major, dark-green for students having graduated in another major, and dark-green-blue for students graduating with double majors. Lastly, a light-blue group was added to represent students who had not chosen a specific major by the time they were admitted to the university, but who later chose to join into the major (i.e., open-major students).

Comparing these majors (top and bottom panels of figure 5), the differences in students' progress in the two majors is easily appreciated. Starting on the very left, the first major begins with a smaller group of students that have originally chosen to be in the major (blue), and over the next terms it welcomes open-major students (light-blue), and a large group of students from other majors (green). On the contrary, the second major starts with a large group of students that have originally chosen the major (blue), but it attracts fewer open-major students (light-blue) or students from other majors (green). This difference in attracting students from other majors was quite informative. On the very right, we notice the differences between the majors with respect to students graduating or leaving the university. The first department has a small group of students leaving the university (red), and the proportion of students graduating

from the major (dark-blue) is higher than the proportion of students in another major (dark-green). In the second major, more of its students leave the university and more graduate from other majors. The middle sections of the figure allows us to understand how different student groups have progressed over the semesters. For instance, in both majors, we note that the loss of students (red) happens mostly in the first 2 years (by the fifth column), with the second department's loss stabilizing after the second year, but the first's slightly growing more.

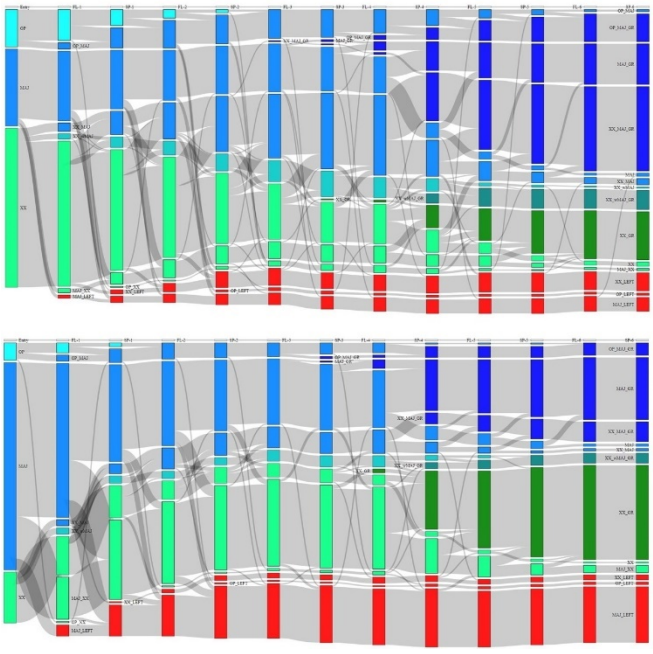

FIGURE 4
SANKEY VISUALS COMPARING STUDENTS' PROGRESS IN TWO NATURAL SCIENCE DEPARTMENTS OVER A 6-YEAR PERIOD

In a second case study, figure 5 shows the progress visuals for female and male students in a third department, over 6 years. Again, starting from the very left end, we can note the differences between male and female students' entry characteristics into the two majors. While there is a considerable number of undeclared (light-blue) male students who eventually chose this major, the number of female undeclared students is less. Yet, more females join this major after having started at the university in another major, which is an interesting difference. On the very right, we note that a larger proportion of males leave the university (red) compared to females. Also, as in figure 4, the drop rates stabilize for both groups after the second year (by the 6th column). Several other subtle patterns can be noticed after a more detailed comparison of these diagrams. For instance, we used these figures to compare time to degree differences between departments, graduation rate differences between initially declared and undeclared students and for comparing several other time-dependent metrics. By varying the entry cohort according to different criteria (e.g., gender, first-generation status etc.), even more patterns can be revealed. In general, these progress visuals serve as excellent platforms for faculty discussions focused on identify obstacles and improving existing courses and curricula, degree requirements, and other departmental practices for achieving better student success.

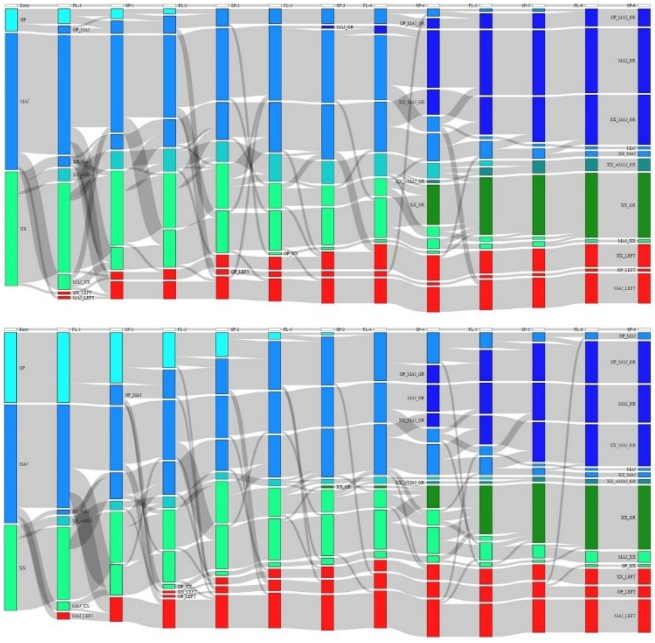

FIGURE 5
SANKEY VISUALS COMPARING FEMALE AND MALE STUDENTS' PROGRESS IN A DEPARTMENT OVER A 6-YEAR PERIOD

V. Software Aspects

To minimize the cost of our efforts, we used open-source software. We used R [42] and its package networkD3 [43] to do the data analysis and later to develop the Sankey visuals. After the cohort data was extracted from the campus-wide database, each student group (i.e., the five groups defined on pg. 5) was identified by analyzing end-of-term majors. Depending on the software, this information could be stored in different formats. In our codes, we used a list of dataframes to store the data. For building Sankey visuals, we had to identify flows between different groups in each consecutive semester. This was accomplished by finding the common student-IDs in different groups in each consecutive semester. After identifying these student IDs, aggregate analysis was followed, and yielded the flow information, that is, the number of students that were either staying in their group or moving to a different group in the subsequent semester. When using the networkD3 package, this extra information should be stored as a dataframe, with the first two columns representing group numbers, and the third representing the value of the flow, that is the number of students. Depending on the particular dataset, a few extra efforts may be also necessary, for instance removing empty groups (groups with no students).Based on our experience, the technical difficulties are manageable for most IR Analysts with basic programming skills.

One thing that may not be evident from the visuals is

that the networkD3 library allows a lot of flexibility for modifying many aspects of the visuals, such as coloring, spacing of bars, displayed texts, etc. This helped us standardize our visuals' characteristics for the best presentation, and it can be modified further depending on the department data and the analysis sought. In addition, the library can produce the Sankey plots in the .html format, which provides interactive features for the user, i.e., being able to get group and flow information by hovering over the visuals with a mouse pointer. This allows the decision makers to perceive all essential cohort information using a single visual .html file. Further, these .html files can be easily incorporated into the university's web portal to provide the information to a larger audience via the internet.

CONCLUSION

In this study, we provided a summary of our efforts for developing a Sankey-diagram based visual tool that could convey students' progress patterns at an academic unit at a higher-ed institution. We discussed the essential questions that we were interested in answering, the similar previous approaches, and a detailed and illustrative discussion of our approach. We provided a couple of actual case studies to showcase these visual's practical use at our university.

REFERENCES

[1] "Academic Review and Planning Advisory Committee (ARPAC)" https://www.colorado.edu/facultyaffairs/arpac Web. Accessed: January 20, 2022.

[2] Romero C. and Ventura S., "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12–27, 2013.

[3] Baker R. S. and Yacef K., "The state of educational data mining in 2009: A review and future visions," *JEDM— Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009..

[4] Romero C. and Ventura S., "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 6, pp. 601–618, Nov 2010.

[5] Mohamad S. K. and Tasir Z., "Educational data mining: A review," *Procedia - Social and Behavioral Sciences*, vol. 97, pp. 320 – 324, 2013, the 9th International Conference on Cognitive Science..

[6] SHEF: FY 2017 State Higher Education Finance, State Higher Education Executive Officers Report, 2018.

[7] Current Term EnrollmentFall 2012, NSCRC Report, 2012. https://nscresearchcenter.org/wp-content/uploads/CurrentTermEnrollment-Fall2012.pdf. Web. Accessed: January 20, 2022.

[8] Current Term EnrollmentFall 2018, NSCRC Report, 2018. Web https://nscresearchcenter.org/wp-content/uploads/CurrentTermEnrollmentReport-Fall-2018-3.pdf. Web Accessed: January 20, 2022.

[9] A World on the Move Trends in Global Student Mobility, Institute of International Education (IIE), Center for Academic Mobility Research and Impact Report, October 2017. https://p.widencdn.net/w9bjls/A-World-On-The-Move. Web Accessed: January 20, 2022.

[10] Lonn S., Aguilar S. J., and Teasley S. D., "Investigating student motivation in the context of a learning analytics intervention during a summer bridge program," *Computers in Human Behavior*, vol. 47, pp. 90 – 97, 2015

[11] Arnold K. E. and Pistilli, M. D. "Course signals at Purdue: Using learning analytics to increase student success," in *Proc. of the 2nd Int. Conf. on Learning Analytics & Knowledge*, ACM, 2012, pp. 267–270.

[12] Barber R. and Sharkey M., "Course correction: Using analytics to predict course success," in *Proc. of the 2nd Int. Conf. on Learning Analytics & Knowledge*, ACM, 2012, pp. 259–262..

[13] Grann J. and Bushway D., "Competency map: Visualizing student learning to promote student success," in *Proc. of the 4th Int. Conf. on Learning Analytics & Knowledge*, ACM, 2014, pp. 168–172...

[14] Ahadi A., Lister R., Haapala H., et al, "Exploring machine learning methods to automatically identify students in need of assistance," in *Proc. of the 11th Annual Int. Conf. on International Computing Education Research*, ACM, 2015, pp. 121–130..

[15] Mazza R. and Milani C., "Gismo: A graphical interactive student monitoring tool for course management systems," in *Int. Conf. on Technology Enhanced Learning*, Milan, 2004, pp. 1–8.

[16] Bakharia A. and Dawson S., "Snapp: A bird's-eye view of temporal participant interaction," in *Proc. of the 1st Int. Conf. on Learning Analytics and Knowledge,* 2011, pp. 168–173.

[17] Competency-Based Education, What Does Learning Look Like? https://www.capella.edu/blogs/cublog/measure-learning-with-capella-university-competency-map Web Accessed: January 20, 2022.

[18] Orr M. K., Lord S. M., Layton R. A., et al, "Student demographics and outcomes in mechanical engineering in the U.S." *Int. Journal of Mechanical Engineering Education*, vol. 42, no. 1, pp. 48–60, 2014.

[19] Morse C., "Visualization of student cohort data with sankey diagrams via web-centric technologies," MS Thesis, 2014.

[20] Heileman G. L., Babbitt T. H., and Abdallah C. T., "Visualizing student flows: Busting myths about student movement and success," *Change: The Magazine of Higher Learning*, vol. 47, pp. 30–39, 2015.

[21] Horvth D. M., Molontay R., and Szab M., "Visualizing student flows to track retention and graduation rates," in *Proc. of 22nd Int. Conf. on Information Visualization (IV)*, 2018, pp. 338–343.

[22] Basavaraj P., Badillo-Urquiola K., Garibay I., et al "A tale of two majors: When information technology is embedded within a department of computer science," in *Proc. of the 19th Annual SIG Conf. on Information Technology Education,* ACM, 2018, pp. 32–37.

[23] American Academy of Arts & Sciences, A Primer on the College Student Journey Report, 2016. https://www.amacad.org/publication/primer-college-student-journey Web Accessed: January 20, 2022.

AUTHOR INFORMATION

Ali Oran, Sr Research Scientist, Brigham & Women's Hospital. (Ali Oran was previously with University of Colorado Boulder)

Andrew Martin, Professor, Department of Ecology and Evolutionary Biology University of Colorado Boulder

Michael Klymkowsky, Professor, Department of Molecular, Cellular & Developmental Biology, University of Colorado Boulder

Robert Stubbs, Director of Institutional Research, University of Colorado Boulder