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

Ali Oran, Andrew Martin, Michael Klymkowsky, and Robert Stubbs

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 for departments' faculty and decision makers should be presentable in a highly-informative yet easily-interpretable manner, 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, which includes members of the campus community and external area experts. 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, EDM, Learning Analytics, LA, Higher Education, Data Visualization, Open-Source

I. Introduction

A. Motivation and The Academic Review Processes

In order to ensure academic excellence higher education institutions commonly engage in periodic and critical revision of their academic programs. While the terms describing these efforts vary slightly among institutions,

A. Oran and R. Stubbs are in Institutional Research group, A. Martin is in Ecology & Evolutionary Biology Department, and M. Klymkowsky is in Molecular, Cellular & Development Biology Department, at University of Colorado, Boulder.

An earlier version of the proposed methods was accepted for a speaker session in Association for Institutional Research (AIR) 2019 Forum in Denver, under the title: *"Where Did the Freshmen Go? Analyzing Students Progress Using Visualization"*.

“Academic Review and Planning” at University of Colorado Boulder [1], “Program Review” at Northwestern University [2] and at University of Washington [3], and “Academic Program Review” at Cornell University [4], the revision process and its objectives are very similar. We will use “Academic Review” as a general term for referring to these efforts, and the following as a common definition: “A regular review of colleges, schools and academic units designed 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 discipline 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 units 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 this phase, 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 (either by withdrawing from the course, or earning a grade for which the student doesn’t gain mastery of the subject). Ideally, the collected data, as well as its presentation, should provide clear and easily interpreted information to an academic unit so as to initiate productive discussions in the succeeding steps about central issues impacting students’ educational experiences. However, such efforts can be derailed if the information provided fails to convey the important patterns in the data to faculty and administrators. It is important to note that after research and teaching responsibilities faculty members, have a limited amount time to fully digest the provided information. At the same time, faculty and other decision-makers (administrators and departmental curriculum committees), may also lack familiarity with specialized data delivery and presentation platforms. Hence, the message and impact of student and academic program data can easily be lost in technical distractions and face the risk of not affecting the educational decision-making process. 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.

To assist these efforts, the advances in data science have given higher-ed institutions more means to analyze rich campus-wide data sources, which if carefully harnessed can yield valuable insights. Yet, the sheer amount of information being analyzed from these data sources can be overwhelming to many decision makers. In addition, different than the past, institutions are welcoming students from very different socioeconomic backgrounds, with diverse academic preparations and motivations. This diverse body of students are also following very different student paths after admission, depending on their interests, academic preparations, and financial means. Accordingly, academic reviews aimed for improving students' success still requires substantial efforts, first for analyzing the diverse data sets, and second for interpreting analyses results to take necessary actions. Hence, to help Higher-Ed faculty and decision-makers make better data-informed decisions, there is a continuing need for highly-informative yet easily-interpretable data-mining methods that can reflect the diverse student progress characteristics in academic units. But, how can we present insights from vast data sources for a maximum impact? How can we make its implications more readily discernible to campus administrators plus internal and external reviewers? A major driver of this study has been to answer this question to identify points of concern and target improvements. Among possible methods, visualization-based approaches can be ideal for conveying departmental data -the well-known saying 'A picture is worth a thousand words' perfectly capturing their descriptive capabilities.

B. Problems of Interest and the Academic Major Ecosystem

A critical aspect of a typical undergraduate degree program is the extent to which students can successfully complete its requirements within reasonable times. Unfortunately, a recent study by American Academy of Arts and Sciences [5] indicates that too few of students graduate and too few graduate in a timely manner. Accordingly, in the context of the academic program review, discussions often center around retention and graduation metrics, and the initiatives to improve them. A number of obstacles can reduce retention and graduation, such as budget cuts that reduce available resources to students' learning and advising, curricular plans that impose barriers to students' progress, various instructor and course effects that can impose academic ”bottlenecks”. These differences require us to analyze possible detrimental factors to students' success on the level of academic units. With a better understanding of students' progress and the gain and loss of students in an academic unit over time, common stumbling blocks that can be hindering students success could be identified, and potential solutions enacted. Yet, as discussed in the previous section, the conveyed information should be sufficiently-informative while easily-interpretable to faculty and administrators.

This type of an analysis on students' success should be able to reveal the characteristics of the group of students who fail to complete their studies in a major and of the group of students 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 in the same institution, or leaving the institution altogether) the analysis needs to distinguish these separate entry and leave groups. In addition, since the move-in and move-out decisions happen at any academic term the analysis should be able to convey these time-dependent student decision patterns clearly. Considering the multiple pathways into and out of a major and the time dependent characteristics of these pathways, we consider an academic major as an evolving ecosystem with multiple subgroups interacting in it. Fig. 1 illustrates this concept with inflows and outflows to some major, Major-A. In this ecosystem, in order to provide decision makers with valuable insights, we need to characterize each student subgroup's evolution and the differences from subgroups. This characterization will allow us to understand why' some subgroups are able to progress with successes and why some others cannot and eventually leave the major. Now, this question can be brushed aside as just an example of young students testing out different majors,

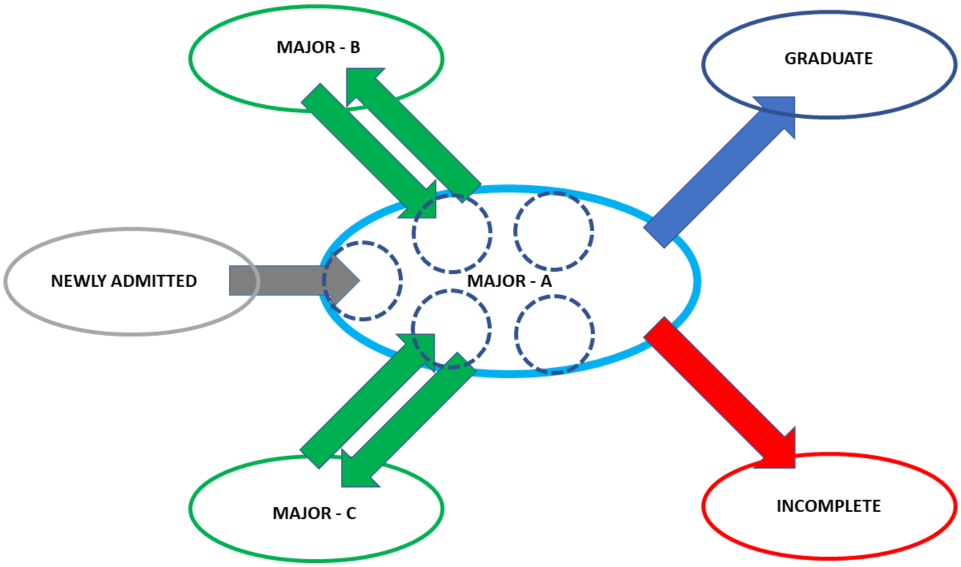


Fig. 1. An illustration of academic ecosytem, Major-A. Dashed-line circles representing subgroups in the ecosystem, and arrows representing student inflows/outflows to/from Major-A.

and accordingly showing random mobility patterns between majors and course groups. However, the real reasons may be quite different - a program may require courses that fail to engage (or seem relevant) to students, or are badly designed in a program's curriculum. While some might see such courses as "rites of passage", they can also be seen as "gatekeepers" - and their elimination or reform could enhance student retention and success. Also, some instructors' teaching methodology may be contributing to loss/gain of students from/to a major, especially in "gateway" courses. Accordingly, an academic unit ecosytem's evolution patterns in terms of "timeline", "origin-destination majors", and also "student group characteristics" should be clearly conveyed. The time of departures from a department can yield essential information about students' satisfaction and the problems they face while progressing in that department. For instance, the underlying reasons for a group of students leaving their major within their 1st year are likely to be quite different from the reasons of another cluster of students leaving later on. Other than that, students original and eventual destinations can also yield useful insights. Students switching to unrelated disciplines from their original majors (e.g. science to humanities) is quite a different pattern compared to students who switch between similar majors or to those who leave the university altogether. This type of an analysis can also provide a clear summary of which majors are more welcoming to students from other majors or from other institutions, and which do not allow for such transitions (either intentionally by high academic requirements, or unintentionally as a result of very distinct course structures). Finally, considering the diverse student bodies higher-ed institutions have been welcoming, the analysis should distinguish the student groups who might be effected 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 an academic unit ecosystem we will be answering 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?

Among the methods that can accurately answer these core questions, a stratified visualization-based approach can be one of the best options by being able to deliver a highly-informative yet easily interpretable analysis. Recent advances in visual methods have allowed us to find answers to these questions by developing visuals beyond the traditional visuals used in Higher-Ed (e.g. Line and Bar graphs, Pie charts etc.) with considerably little added coding effort. Here, we will provide the details to our approach and coding practices, after a brief review of relevant studies.

C. Data Mining Efforts in Higher Education & Related Work

Data mining is the process of discovering useful patterns from large amounts of data in an automatic or a semiautomatic way [6]. In this field, Educational Data Mining (EDM) can be defined as the area of scientific inquiry centered around the development of methods for making discoveries within the unique kinds of data that come from educational settings, and using those methods to better understand students and the settings within which they are taught and learn [7]. A similar term, Learning Analytics (LA) is also used for similar efforts, with slight differences in their approaches [8]. Independent of the terminology used, there has been a considerable increase in the use of these techniques in the context of Higher Education over the past two decades [9]. [9]. These techniques have given Higher-Ed institutions more analytical options to harness large campus-wide data sources for identifying possible areas for improvement and making data-informed decisions on a wide range of issues, from optimization of daily operations to improving of student engagement and learning outcomes. Accordingly, a very diverse group of studies have been proposed touching different issues in higher education (see surveys [10]–[13]). In light of these developments, it appears that most institutions have been making investments on both Descriptive and Predictive Analytics, either for improving student outcomes, or for more efficient delivery of programs or services [14].

The increasing use of analytics in Higher Education is also related to institutions' growing urgency to offset some of their recently realized challenges by harnessing the large campus-wide data sets. 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) [15]. 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) [15]–[18]. At the same time, there has been an increasing competition with educational institutions from other developed nations in attracting international students, whose out-of-state tuition have become essential for some universities, in an increasingly globally competitive field of higher education. Compared to the early 2000s, other countries, particularly Canada and Australia, have become educational destinations for a larger percentage of international students [19]. Some of these countries' future strategic plans aim to attract even more international students [20], [21], which may become an issue for some US higher education institutions that previously have relied on foreign students' out-of-state tuition. Finally, as a new generation of diverse students are entering our universities, possible factors that may be hindering students' success needs to be carefully identified. In these times of changing enrollment patterns, a better use of campus-wide data sources through analytics can give the decision makers the necessary insights to weather these times better[1]

While there have been a considerable number of studies on how to improve students' academic outcomes, there hasn't been a census to properly categorize the works, with each study/survey having introduced its own categorization. For our brief discussion, we distinguish them according to their goals. We refer to the group of studies whose aim is to accurately describe (summarize) student's academic performances as the "Descriptive Methods". And, we refer the group of studies whose aim is to accurately predict student's performances as the "Predictive Methods". Institutions have been using Descriptive Analytics for years to understand students' performances, and subsequently to take necessary action when necessary, such as reshaping their entering classes, refining policies and course requirements etc. In recent years, Predictive Analytics have been added to these efforts [22]. 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 [23]. Purdue University (Course Signals) [24], University of Phoenix [25], and Capella University [26] are just a few examples of the universities that have utilized such systems. Logistic Regression has been a common method for the predictions [25], [26], yet, more advanced methods from Machine Leaning have also been tried as well [27]. While Predictive Methods have been gaining more attention recently, the difficulty of accurately incorporating qualitative factors, such as student motivation and persistence which also influence student success, is still a major limitation for them.

Descriptive Methods form a wider group of studies that have been used in Higher-Ed for years. The challenge for data practitioners is how to choose the best method for analyzing the data of interest within limited time frames and afterwards presenting the findings to the decision makers in an impactful manner. 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. In addition, recent advances in data visualization have particularly simplified the once dreaded coding aspect of the visualization process, and have allowed visuals to become a more common method to convey data patterns. Accordingly, Higher-Ed institutions are also seeing a change from relying on traditional visuals (Line graphs, Pie charts etc.) to adoption of newer visuals for gaining better insights from the ever expanding campus-wide data sources, that includes Financial, Academic, and Personnel data. While some of the proposed visual methods remain experimental, others have seen actual use, and a greater discussion on these proven visuals is needed to ensure further progress.

Among Visual Methods, a distinction can be made according to methods' focus. Some visuals' focus is on visualizing students data in an aggregate way (Aggregate Visuals), while others focus' is on displaying each students data separately (Tracking Visuals). In practice, Aggregate Visuals can visualize characteristics of particular cohorts

[1] When this manuscript was being finalized, COVID-19 pandemic has brought an extra layer of uncertainty for student enrollment in higher education institutions.

or groups of students, e.g. visuals for average time-to-degree analysis of Chemistry majors accepted between 2010 to 2014. And, Tracking Visuals can visualize characteristics of each student, e.g. visuals for a particular student's course work over consecutive terms as a means of tracking the student's progress or other visuals to monitor the student's learning efforts. Accordingly, they have seen a wide usage particularly in on-line educational environments, 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 also to instructors for reaching out to students at necessary times. Mazza and Milani's GISMO [28], Bakharia and Dawson's SNAPP [29], Chiritoiu et. al's Students' Activity visualization tool, and Capella University's Competency Map [26], [30] are some of the well known examples of this group of tracking visuals.

In contrast, Aggregate Visuals can instead summarize the characteristics of cohorts, and accordingly can be quite useful in detecting possible issues within academic units' educational practices. This type of focus on cohorts and departments could yield valuable information when visuals reveal the variations of students' success and progress among different cohorts, and among different departments of the same university. Among Aggregate visuals, while the traditional data visualization methods (such as line plots, pie charts etc.) have been prevalent in Higher-Ed, with the recent advances in visualization software more informative visuals have recently gained attention interest by delivering more information. In this context, for understanding students' aggregate progresses and mobility patterns, Flow Diagrams are one of the promising new alternatives. From a general point, a Flow Diagram (or Chart) visually displays interrelated information such as events, steps in a process, functions etc., in an organized fashion, such as sequentially or chronologically. It can be constructed to visualize a variety of patterns such as manufactured products, currency moving between countries, paperwork progression through an organization [31]. 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 studies, including visualizing Energy Flows in the Energy Sector [32], Material Flows in the EU [33], Land Cover Dynamics in Urban Planning [34], and Temporal Visualization of Diabetes in Health Informatics [35].

However, the use of Sankey Diagram's for visualizing students' progress has been limited. To the best of our knowledge, the first study to use Sankey Diagrams in this context was in early 2014 by Orr et al., who analyzed the origins and destinations of students ever enrolled in Mechanical Engineering in a simple 2-column Sankey diagram [36]. Later in the same year, Morse, in his Masters thesis [37], proposed a multi-column Sankey diagram for visualizing students progress. Morse's work was similarly followed in Heileman et. al. in 2015 [38] to analyze a few particular hypothesis (termed myths) about students' success. These were followed by a couple of other studies in 2018 by Horvth et al. [39], and Basavaraj et al. [40] to understand students' progress in their institutions.

Following Morse's work, an informative visualization like this would have been expected to be embraced by institutions to understand their students' progress patterns. Yet, it has not been much noticed, and by those who have noticed it is still mostly considered an experimental visual. Hence, a detailed discussion of this new visual is needed by discussing the need for them, the questions they can answer, a comparison with other standard methods they can be replacing, possible software requirements for their development, and most importantly how they can be actually used in higher education institutions. This study attempts to fill this need by providing a discussion on these topics using our own experiences in developing highly-informative yet interpretable presentations for conveying students' progresses patterns to a group of departments undergoing academic review at our research intensive institution of more than thirty-thousand students. We will start our discussion, in II, by detailing how we address the core questions we had itemized in I-B. We will introduce the general principles of our approach in II-A, which will be followed by a brief discussion on the required data extraction and data analysis efforts in II-B. In II-C we will provide details on the proposed visualizations, and we will compare the traditional presentation methods and discuss their weaknesses (such as missing information, reproducibility, and ease of understandability), and discuss how Sankey-based visualizations can address those weaknesses. Finally, we will discuss how the Sankey-based visuals were used in practice, by detailing their use with a couple of case studies, and by discussing the software aspects. Most of our discussion through the paper will be independent of a specific software, but our use of open-source software R and its visualization package networkD3 will be detailed in the last section.

II. Identifying Students' Progress and Mobility Patterns

A. General Approach

We begin our discussion with the key questions we want to answer for academic units: who are the students entering the major, when do they enter, from where do they come from, and over the next semesters where do they go (how do they progress through the major). It is worth noting that each majors student population continuously evolves, with students moving in to it or leaving it at each academic term. For such a system, a longitudinal study, in which a cohort is followed for some time, in order to establish the frequencies of specific conditions and their correlation with environmental or other factors [41], will be convenient. In our analysis, the cohorts for a major will be defined as students who were part of that major at least for a semester and who had the same entry year to the institution, e.g. Biology-2010 cohort refers to the group of students who have entered the university in 2010-Fall semester and have majored in Biology at some semester after that. With this cohort definition, we will be able to answer the *"when"* question, through the analysis of multiple cohorts and their subcohorts and by observing their characteristics over the years. For answering *"where"* students come from and go to, we will be analyzing students enrollment majors at each term. Lastly, we will split the entry year cohort group into different subcohorts to answer the *"who"* question by observing the progress differences among different subcohorts. The splitting criteria depends on the analysis sought, for instance, splitting 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, such as Open option, Other-major and Transfer can reveal possbile disparities among those.

Our approach can be considered as a two step process:

1) Data Extraction and Data Analysis
2) Data Presentation

Data Extraction, requires the preparation of the needed data set, which is followed by Data Analysis to identify students' progress patterns. Data Presentation step involves developing the best method, in terms of informativeness and interpretability, for conveying the identified data patterns for the audience. Figure 2 summarizes these steps,

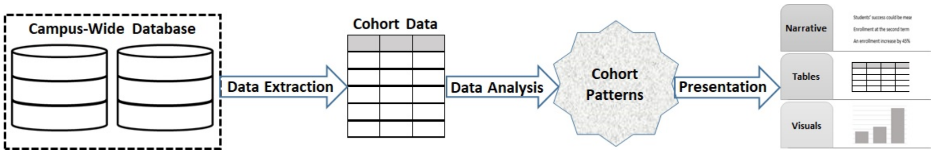


Fig. 2. General-level flow of Students' Progress and Mobility Pattern analysis.

and we will briefly detail the Analysis & Extraction steps in the next subsection, before going into details for the Presentation, the progress visuals.

As discussed, it was also critical to take into account that the analysis might be conveyed to an audience that had limited experience with the data sets and the visuals. Accordingly, we developed a list of guidelines, which we followed closely, in order to be able to develop a highly informative, yet easily interpretable analysis for our institution that can be used in the long term. The first two guided our final product, and the last two the process for managing the visuals in the long term:

1) Informativeness: The analysis should contain enough information to answer the questions of interest in a satisfactory manner.
2) Interpretability: The analysis should be easily interpreted by people not necessarily working along with campus-wide data.
3) Scalability: The analysis should be scalable so that it could be reproduced easily for different data sets for possible comparison of different cohorts and departments.
4) Ease of Cost & Time: The analysis' development, and succeeding updates should be as cost-effective as possible in terms of finance and time.

We will consider the trade-offs needed to be done to satisfy these guidelines.

B. Data Extraction and Analysis of Students' Progress

To start a data analysis effort, one needs to collect the necessary data from the right sources, a process known as *Data Extraction*. It involves standard database manipulation techniques, such as joining of different tables from the Campus-wide databases, and filtering of the data according to the selected criteria, for the cohort of interest. Cohorts can be defined according to problems of interests, and one can even extend the analysis to consider a collection of cohorts, with multiple entry years and majors, e.g. multiple cohorts representing different majors in a Natural Sciences department. Yet, the choice of cohort groups and the time frames eventually affects the complexity of the presentation in the succeeding step, with the growing number of different paths students take after admission (e.g. in terms of majors) being a limiting factor to convey all patterns in an interpretable format. Hence, a trade-off should be made between the information aspect and interpretatbility when defining the cohorts. Accordingly, in our study, student cohorts are defined based on the year students enrolled at the institution and whether they had declared a particular major during some semester in their studies (between entry and exit). For such cohorts, we keep track of cohorts majors at each semester to answer the *"where"* question. Hence, our extracted data will include the 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 for our analysis will require the following variables:

Cohort Data: [**Student-ID**, **Year-Term**, **Enrolled Major**, **Degree Date (if grad.)**, **Degree Major (if grad.)**] Once the data for the cohort is prepared, we move on to the *Data Analysis* step where we look for the patterns in it. For student's progress analysis, in the most simplest form, this boils down to analyzing the enrolled-majors at each academic term. More complex analyses could be incorporated, such as an analysis on course grades, or race/gender representation in cohorts; as long as complexity of the presentation can be managed. In our analysis, we only analyzed enrolled-majors over the semesters; and, this analysis already yielded a very diverse group of enrolled-majors in a few semesters, especially for cohorts from academic units with a large number of students. Accordingly, in order to further simplify the analysis results and the succeeding visuals, we recommend combining all enrolled-majors other than the initial major into the one category, called *"other major"*. Hence, students from the original cohort can be regrouped into the following groups at each academic term:

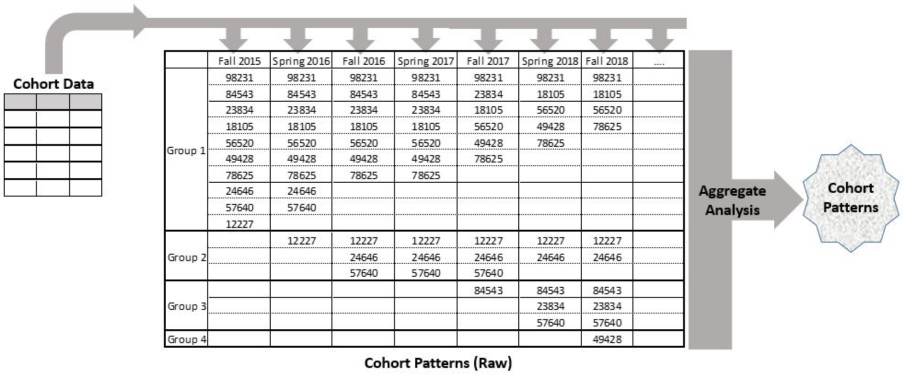
- Actively seeking degree in the major
- Actively seeking degree in an other major
- Graduated in the major
- Graduated in an other major
- Left the university

To illustrate these groups, we provide a simple example with a sample cohort of 10 students, who all enter the university in Summer-2015 and had the same declared major before starting their studies in Fall-2105, as detailed in Fig. 3 with random student IDs. This is also the group of students that are actively seeking degree in the major, and we name their group Group-1. At the end of each term, some students from Group-1 may leave the original major to another major, or may leave the university (without a degree), or may graduate. Accordingly, in Fig. 3 over the next several terms, student IDs are distributed according to their majors by the end of each term, with Group-1 consists of students continuing in the original major, Group-2 consists of students who changed majors (all alternative majors included in the same group), Group-3 consists of students who left the university without a degree, while Group-4 are students who have graduated (from any major).

After the patterns are identified in the raw data, we then apply an aggregate analysis, which yields information about the general characteristics of a cohort. We were 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 also the university, for a better understanding of retention and graduation characteristics. Other types of analyses could be pursued, such monitoring average GPA or race/gender representation in each group after admission. Figure 4 shows the table that summarizes the aggregate analysis on our sample data from Fig. 3.

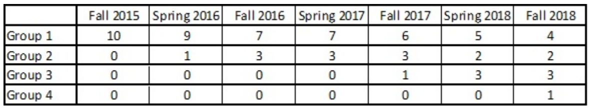
C. Visualizing Cohort Progress Patterns

After the analysis step, one needs to present the findings in the best way so that the progress and mobility patterns can be accurately and easily perceived, and their implications discussed. The table shown in Fig. 4 is in one option; it clearly summarizes the number of students in each group over time. Especially, for short time spans



	Fall 2015	Spring 2016	Fall 2016	Spring 2017	Fall 2017	Spring 2018	Fall 2018	....
Group 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				
	24646	24646						
	57640	57640						
	12227							
Group 2		12227	12227	12227	12227	12227	12227	
			24646	24646	24646	24646	24646	
			57640	57640	57640			
Group 3					84543	84543	84543	
						23834	23834	
						57640	57640	
Group 4							49428	

Fig. 3. A simple example of analyzing a sample cohort of 10 students' progress after their admission in Summer-2015. The random numbers depict distinct student-IDs and their distribution show each student's progress (belonging to certain groups) over consecutive semesters.



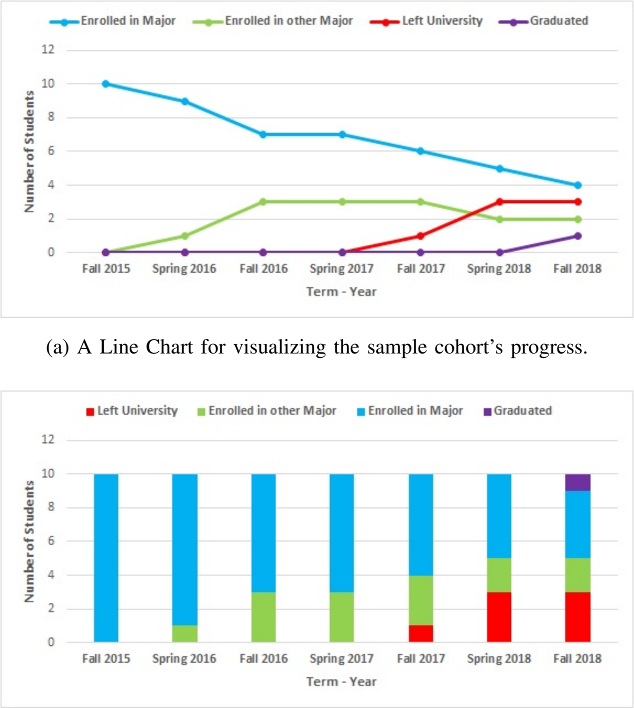
	Fall 2015	Spring 2016	Fall 2016	Spring 2017	Fall 2017	Spring 2018	Fall 2018
Group 1	10	9	7	7	6	5	4
Group 2	0	1	3	3	3	2	2
Group 3	0	0	0	0	1	3	3
Group 4	0	0	0	0	0	0	1

Fig. 4. Aggregate analysis for group sizes for the sample data from the previous figure.

of 4 or 5 semesters, it may be adequate for some institutional discussions. For longer time spans, such as tables presenting term-wise retention numbers, it may not be the most ideal form to present long longitudinal data in a table format. In addition, while the table can present 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 become an issue for a 10-student cohort with 4 groups, but for larger cohorts with more groups such wide and long tabular data can be difficult to digest. Often additional data tables are introduced, which makes the presentation more complex and difficult to interpret, particularly by those not having enough time or interest in understanding the data. Accordingly, using of visual techniques will be needed at least as a supplement, if not as a replacement, to these tables.

Some of the traditional graphical ways of presenting the cohort patterns are using the Line Charts or the Stacked Bars, shown in Fig. 5a and Fig. 5b respectively for our sample data set. Both clearly show the number of students who left their original major, left the institution, or graduated by the end of each term. Accordingly, these figures can convey the general patterns of student mobility, with different colors representing different groups and y-axis providing information about these groups' relative sizes compared to the original cohort. It can be also noted that the Stacked Bar Chart, by having bars of constant height, provides an easier to understand visual for conveying the relative size of each group, when compared to the Line Chart.

What is essentially missing in the Table, the Line Chart, and the Stacked Bar Chart, is the **flow information** about students moving among different groups. For instance, in our sample date set, between Fall 2017 and Spring 2018 we can 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



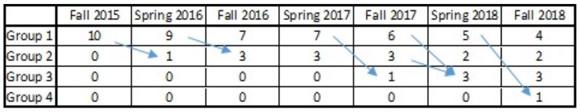
(a) A Line Chart for visualizing the sample cohort's progress.

(b) A Stacked Bar Chart for visualizing the sample cohort's progress.

Fig. 5. Traditional ways of visualizing the sample cohort's progress. In both figures, the cohort is analyzed in 4 groups: "Enrolled in the Original Major (Blue)", "Enrolled in a Different Major (Green)", "Left (Red)", "Graduated (Purple)"

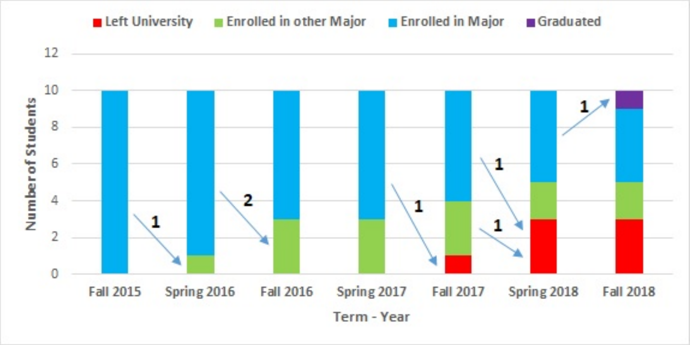
after having switched majors (i.e. after having studied in other majors). This missing flow information is valuable because it fleshes out our understanding of how different student cohorts interact with their programs, and how that changes over time. For instance, the flow information can reveal student satisfaction with a degree program and institution as a whole, as well as indicate how easily students can pursue other majors if they are not satisfied in their original major. A similar question of flow can also be asked about the graduates appearing as yellow patterns in Fall 2018. Again, the table or the figures can't provide the information about students graduation majors, that is whether from their original majors, or other majors. By going back to the raw data in Cohort patterns table in Fig. 3, by comparing student IDs, one can find out that one student has left the institution from the original major, and the other student from another major, that is after trying different majors. And, the graduate was from the original enrolled major.

Ideally, the presentation should contain enough information that going back to analyzing the raw data wouldn't be necessary for the audience. This could be done by adding more layers (groups) to these visuals or the table in Fig. 4. For instance, the Left University group could be expanded to have multiple subgroups according to students last majors, and the Graduated group can be expanded to have subgroups for each separate department that had granted degrees for this cohort. The drawback of this approach is that the more layers we add, the more difficult it is for the institutions decision makers to readily grasp the implications of the data. Especially for cohorts of hundreds of students, one might easily need more than a dozen groups to represent all the majors students were part of and graduated from. Eventually such complex presentations can easily cloud readers grasp of the underlying patterns. In addition, even with more groups, the readers would still need to keep track of the changes in the number of students at 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. As a second alternative, one could enhance tables and figures by superimposing flow information on them, as shown in Fig 6a and in Fig 6b. These additions provide us the essential information that was missing. Yet, the drawback is the reproducibility of these enhanced figures and tables in reasonable times. That is, superimposing another layer of information on top of another figure or table can easily double the time to produce them.



	Fall 2015	Spring 2016	Fall 2016	Spring 2017	Fall 2017	Spring 2018	Fall 2018
Group 1	10	9	7	7	6	5	4
Group 2	0	1	3	3	3	2	2
Group 3	0	0	0	0	1	3	3
Group 4	0	0	0	0	0	0	1

(a) Student flows from each group superimposed on the progress table.



(b) Student flows from each group superimposed on the Stacked Bar Chart.

Fig. 6. Superimposing student flows to provide extra information for the progress of the sample cohort.

One should note that Fig. 6b 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 10-student cohort's progress can be visualized as in Fig. 7. This figure is similar to the Bar Chart Fig. in 6b 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 one semester. With this extra layer of information, it becomes easy to convey student cohorts' progress over time in a clear manner, and accordingly identify the possible bottlenecks students face. For example, a large transfer of students out of a major can be an indicator of the presence of a particular course where students commonly fail or withdraw from. 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 for the academic review process.

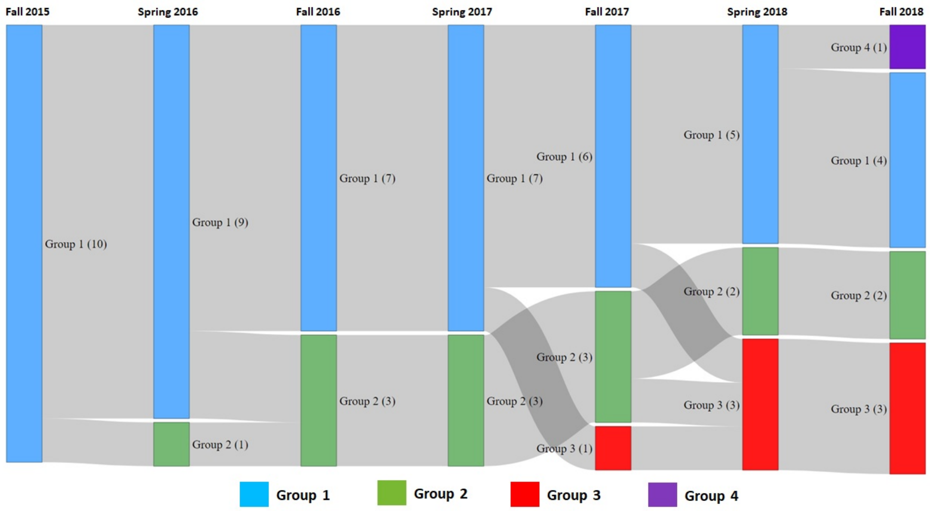


Fig. 7. Sankey for the sample cohort data of 10 students over several consecutive semesters, with each cohort population in parantheses.

D. Sankey in practice for Academic Review process

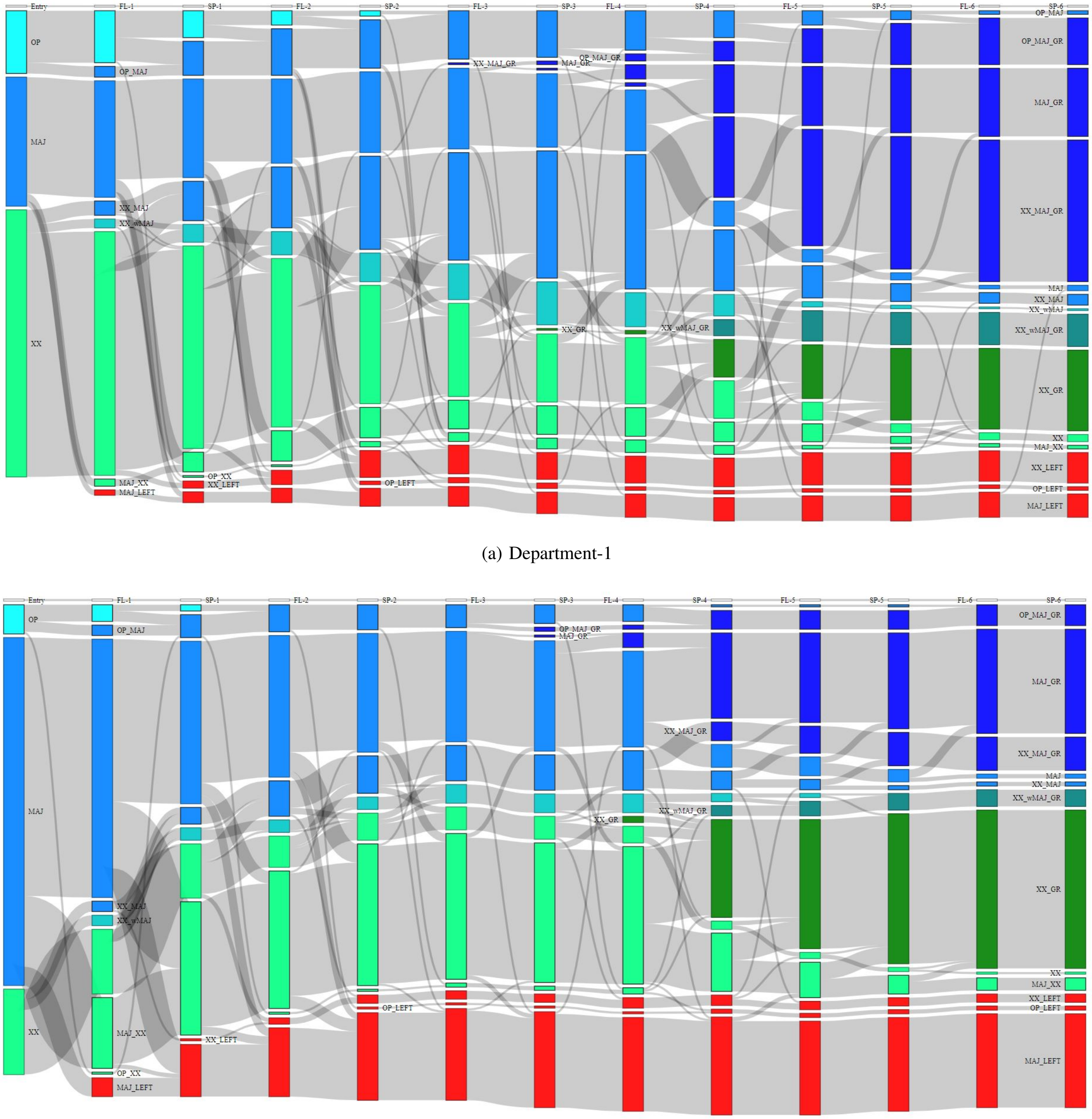
A primary purpose of the Academic Review process, as discussed in I-A, is to initiate productive discussions about central issues impacting students educational experiences. Accordingly, it requires data effectively presented so that its implications can be readily grasped, facilitating the framing of further questions and shaping efforts to improve students' educational outcomes. To help achieve this goal the Institutional Research (IR) office developed Sankey-based Students Progress Visualizations for use in the program review process, beginning with the 2018-19 Academic year. The development was carried out in closely 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 a department in terms of students mobility and progress. The contained information allows academic leadership to develop deeper appreciation of curriculum related practices and outcomes, including major degree requirements, and the effects of required courses and course sequences on retention and timely graduation. In addition, the nature of 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 in 2018, these Student Progress Visualizations have become one of the standard visual tools provided to departments for the 2019-2020 academic year (e.g. Fig. 8). 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 groups of students within a degree program. So far, we have developed stratified Students Progress Visuals for conveying the differences of progress among male/female students (e.g. Fig. 9), 1st-generation/non-1st-generation students, and under- represented-minority/non-under-represented-minority students.

Fig. 8a and 8b illustrate Students' 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. 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-the same colors from the sample Sankey diagram in Fig 7 and the earlier visuals in Fig. 5. In these departments, we noted that double majoring students were a considerable group. Hence, we added a new blue-green mixture color group to our visuals to represent these students' progress. For graduates, 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 an 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 at a later term.

Comparing these two figures, the differences in students' progress in the two majors is easily appreciated. Starting on the very left, the first major starts with a smaller group of students that have originally chosen to be in the major (blue), and over the next terms it welcomes undeclared students (light-blue), and a large group of students from the other majors (green). On the contrary, the second major starts with a large group of students that have originally chosen to be in the major (blue), but attracts fewer undeclared students (light-blue) or students from other majors (green). On the very right, we can notice the differences between the majors about 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). Yet, in the second major, more of its students leave the university and also more graduate from other majors. The middle parts of the figure allows us to understand how different student groups have progressed over the semesters. For instance, in both majors, we can 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.

In a second case study, we analyzed the progress patterns for male and female students in a third department, as illustrated in 9a for the males, and in 9b for the females. Again, starting from the very left end, we can note the differences between male and female students entry characteristics into the major. While there is a considerable number of undeclared (light-blue) male students who eventually choose this major, the number of female undeclared students is less. Yet, more females join this major after having starting the university in another major. On the very right, we can note that a larger proportion of males leave the university (red). And, as before, the leave rates stabilizes for both groups after the second year (by the 5th column).

Several other subtle patterns emerge after more detailed comparisons of these diagrams. For instance, we also use these figures to compare time to degree differences between departments, graduation rate differences between initially declared and undeclared students and more. 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 existing curricula, degree requirements, and other departmental practices for improving student success for a diverse group of students.



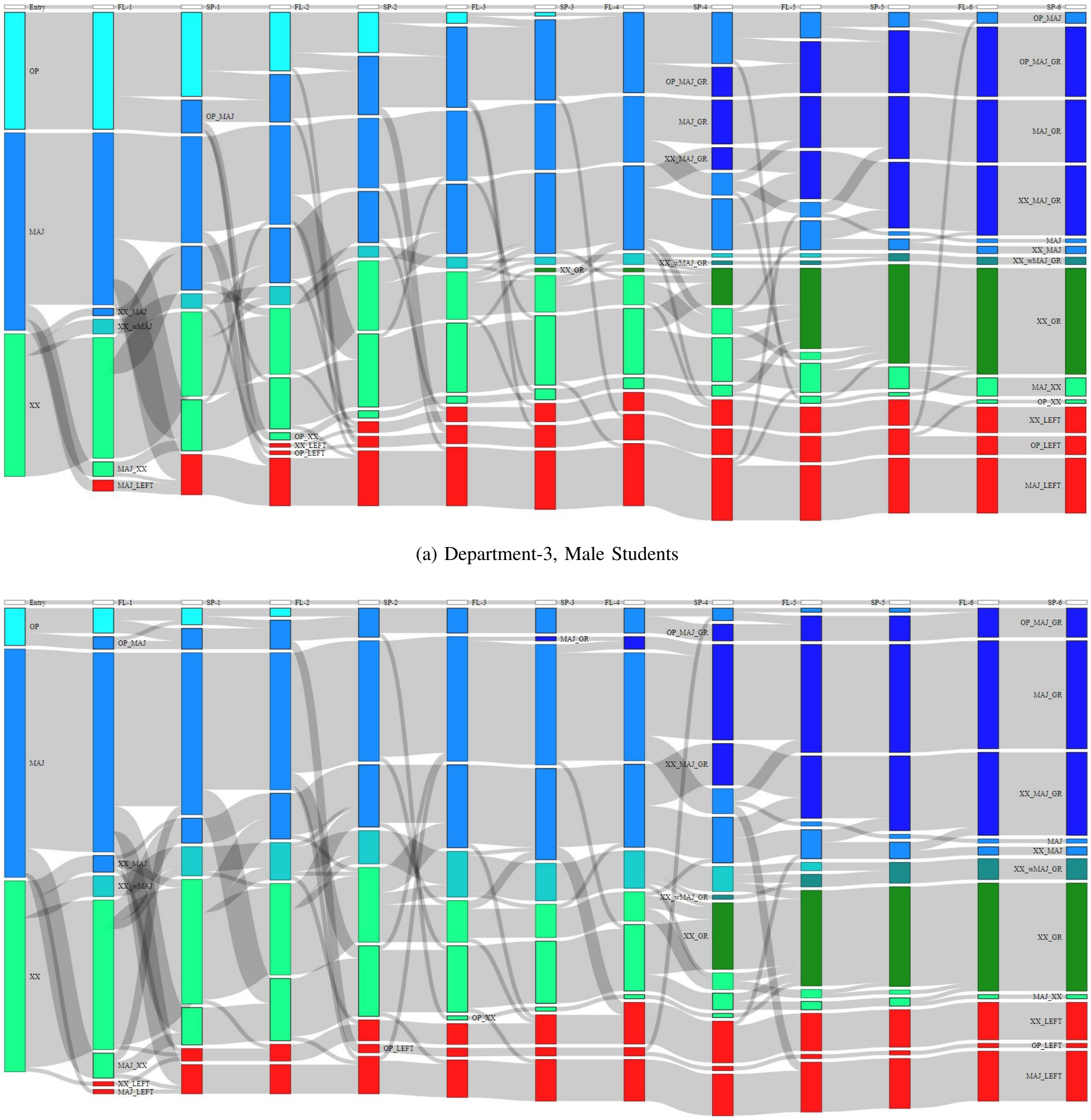
(a) Department-1

(b) Department-2

Fig. 8. Undergraduate students progress and mobility patterns over a 6-year period in two separate departments for a particular entry year.

E. Software Aspects

In this last section, we briefly discuss the software aspects of developing the Sankey-based visuals. To minimize the cost of our efforts, we tried to adhere to using open-source software, something made possible due to the rapid progress in this area. The chosen open-source software to do the data analysis and later to develop the Sankey visuals was R [42] and its package networkD3 [43] The abundant online discussion forums were extremely helpful in helping us improve our codes.



(a) Department-3, Male Students

(b) Department-3, Female Students

Fig. 9. Undergraduate students progress and mobility patterns over a 6-year period in a department for a particular entry year, for male and female student groups.

After the cohort data was extracted from the campus-wide database, it was imported into the R environment for the *Data Analysis* step to identify the students' progress patterns, yielding the tabular Cohort Patterns (Raw) structure shown in Fig 3. Depending on the software used, this tabular structure could be stored in different formats. In our codes, we used a list of dataframes to hold the data in the computer memory. For Sankey visuals, before the aggregate analysis we had to identify flows between different groups in each consecutive semesters. This was accomplished by finding the common student-IDs in different groups in each consecutive semesters. After identifying these student IDs, aggregate analysis followed, and yielded the flow information, that is, the number of students that were either staying in their groups, or moving to another one in the next 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 data set, a few extra efforts may be also necessary, primarily for removing unnecessary 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 networkD3 library allows a lot of flexibility for modifying many aspects of the visuals, such as coloring, spacing of bars, displayed texts, etc. [43]. This helped us finalize graphs' characteristics for the best presentation for our analysis, and it can be modified further depending on the department data and the analysis. 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 the group and flow information by hovering over the visuals with the mouse pointer. This allows the decision makers to perceive all essential cohort information using a single students' progress visual .html file. Further, these .html files can easily be incorporated into the university's web portal to provide the information to a larger audience.

III. Conclusion

In this study, we discussed the needs of developing a visual tool that could convey the progress and mobility patterns of students in an academic program at a higher-ed institution. For this discussion, we provided some of the essential questions that needed answering in higher-ed, the previous studies that approached these questions, and a detailed and illustrative discussion of our approach. Our efforts have resulted in a standard visual that is being used by several departments in their periodic academic reviews, and in other departmental studies. We also provided a couple of actual case studies to reflect the visual's practical use at our university.

References

[1] Academic review and planning. University of Colorado. [Online]. Available: https://arp.colorado.edu/

[2] Program review. Northwestern University. [Online]. Available: https://www.adminplan.northwestern.edu/program-review/

[3] Program review. University of Washington. [Online]. Available: http://grad.uw.edu/for-faculty-and-staff/program-review/

[4] Academic program review. Cornell University. [Online]. Available: https://irp.dpb.cornell.edu/academic-program-regulation/academic-program-review

[5] *American Academy of Arts & Sciences*, A Primer on the College Student Journey Report, 2016. [Online]. Available: https://www.amacad.org/publication/primer-college-student-journey

[6] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016, ch. 1.

[7] R. S. Baker, *Data Mining for Education*, 2005, vol. 19.

[8] G. Siemens and R. S. J. d. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, ser. LAK '12. New York, NY, USA: ACM, 2012, pp. 252–254.

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

[10] ——, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, no. 1, pp. 135 – 146, 2007.

[11] R. S. Baker and K. Yacef, "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.

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

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

[14] A. Parnell, D. Jones, A. Wesaw, and D. C. Brooks, *Institutions use of data and analytics for student success: Results from a national landscape analysis*, NASPA Student Affairs Administrators in Higher Education, AIR Association for Institutional Research, and EDUCAUSE Report, 2018. [Online]. Available: https://www.naspa.org/rpi/reports/data-and-analytics-for-student-success

[15] *SHEF: FY 2017 State Higher Education Finance*, State Higher Education Executive Officers Report, 2018. [Online]. Available: http://www.sheeo.org/sites/default/files/project-files/SHEEO_SHEF_FY2017_FINAL.pdf

[16] *Current Term Enrollment Fall 2012*, National Student Clearinghouse Research Center Report, 2012. [Online]. Available: https://nscresearchcenter.org/wp-content/uploads/CurrentTermEnrollment-Fall2012.pdf

[17] *Current Term EnrollmentFall 2015*, National Student Clearinghouse Research Center Report, 2015. [Online]. Available: https://nscresearchcenter.org/wp-content/uploads/CurrentTermEnrollment-Fall2015.pdf

[18] *Current Term Enrollment Fall 2018*, National Student Clearinghouse Research Center Report, 2018. [Online]. Available: https://nscresearchcenter.org/wp-content/uploads/CurrentTermEnrollmentReport-Fall-2018-3.pdf

[19] *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. [Online]. Available: https://p.widencdn.net/w9bjls/A-World-On-The-Move

[20] *International Education, A Key Driver of Canada's Future Prosperity*, Report, August 2012. [Online]. Available: https://www.international.gc.ca/education/assets/pdfs/ies_report-rapport_sei-eng.pdf

[21] *National Strategy for International Education 2025*, Report, April 2016. [Online]. Available: https://nsie.education.gov.au/sites/nsie/files/docs/national_strategy_for_international_education_2025.pdf

[22] *American Academy of Arts & Sciences*, A Primer on the College Student Journey Report, 2016. [Online]. Available: https://www.amacad.org/publication/primer-college-student-journey

[23] S. Lonn, S. J. Aguilar, and S. D. Teasley, "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, learning Analytics, Educational Data Mining and data-driven Educational Decision Making.

[24] K. E. Arnold and M. D. Pistilli, "Course signals at purdue: Using learning analytics to increase student success," in *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, ser. LAK '12. New York, NY, USA: ACM, 2012, pp. 267–270.

[25] R. Barber and M. Sharkey, "Course correction: Using analytics to predict course success," in *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, ser. LAK '12. New York, NY, USA: ACM, 2012, pp. 259–262.

[26] J. Grann and D. Bushway, "Competency map: Visualizing student learning to promote student success," in *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, ser. LAK '14. New York, NY, USA: ACM, 2014, pp. 168–172.

[27] A. Ahadi, R. Lister, H. Haapala, and A. Vihavainen, "Exploring machine learning methods to automatically identify students in need of assistance," in *Proceedings of the Eleventh Annual International Conference on International Computing Education Research*, ser. ICER '15. ACM, 2015, pp. 121–130.

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

[29] A. Bakharia and S. Dawson, "Snapp: A bird's-eye view of temporal participant interaction," in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, ser. LAK '11, 2011, pp. 168–173.

[30] [Online]. Available: https://www.capella.edu/blogs/cublog/measure-learning-with-capella-university-competency-map

[31] R. L. Harris, *Information Graphics: A Comprehensive Illustrated Reference*. New York, NY, USA: Oxford University Press, Inc., 1999.

[32] K. Soundararajan, H. K. Ho, and B. Su, "Sankey diagram framework for energy and exergy flows," *Applied Energy*, vol. 136, pp. 1035 – 1042, 2014.

[33] P. Nuss, G. A. Blengini, W. Haas, A. Mayer, V. Nita, and D. Pennington, "Development of a sankey diagram of material flows in the eu economy based on eurostat data," 2017.

[34] N. Cuba, "Research note: Sankey diagrams for visualizing land cover dynamics," *Landscape and Urban Planning*, vol. 139, pp. 163 – 167, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016920461500064X

[35] E. M. Hinz, D. Borland, H. Shah, V. L. West, and W. E. Hammond, "Temporal visualization of diabetes mellitus via hemoglobin a1c levels," in *International Conference on Technology Enhanced Learning, Milan*, 2004, pp. 1–8.

[36] M. K. Orr, S. M. Lord, R. A. Layton, and M. W. Ohland, "Student demographics and outcomes in mechanical engineering in the u.s." *International Journal of Mechanical Engineering Education*, vol. 42, no. 1, pp. 48–60, 2014.

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

[38] G. L. Heileman, T. H. Babbitt, and C. T. Abdallah, "Visualizing student flows: Busting myths about student movement and success," *Change: The Magazine of Higher Learning*, vol. 47, no. 3, pp. 30–39, 2015.

[39] D. M. Horvth, R. Molontay, and M. Szab, "Visualizing student flows to track retention and graduation rates," in *2018 22nd International Conference Information Visualisation (IV)*, 2018, pp. 338–343.

[40] P. Basavaraj, K. Badillo-Urquiola, I. Garibay, and P. J. Wisniewski, "A tale of two majors: When information technology is embedded within a department of computer science," in *Proceedings of the 19th Annual SIG Conference on Information Technology Education*, ser. SIGITE '18. New York, NY, USA: ACM, 2018, pp. 32–37.

[41] M. Suchmacher and M. Geller, "Chapter 1 - study type determination," in *Practical Biostatistics*, M. Suchmacher and M. Geller, Eds. San Diego: Academic Press, 2012, pp. 3 – 15. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B978012415794100001X

[42] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[43] J. Allaire, C. Gandrud, K. Russell, and C. Yetman, *networkD3: D3 JavaScript Network Graphs from R*, 2017, r package version 0.4. [Online]. Available: https://cran.r-project.org/package=networkD3