Spring 3-14-2019

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

Ali Oran
Institutional research, ali.oran@colorado.edu

Andrew Martin
Ecology and Evolutionary Biology, andrew.martin-1@colorado.edu

Michael Klymkowsky
Molecular, Cellular & Developmental Biology, michael.klymkowsky@Colorado.edu

Robert Stubbs
Institutional research, robert.stubbs@colorado.edu

Submitted to: manuscript No.
(will be inserted by the editor)

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

Ali Oran · Andrew Martin · Michael Klymkowsky · Robert Stubbs

Received (version as of): **March 14, 2019** / Accepted: date

Abstract The critical revision of the content, scope and alignment of curricula is essential for improving students success at Higher-Ed institutions. The effort is never trivial, as students are accepted from multiple sources (e.g. high schools, community colleges, other institutions) with different academic preparation, socioeconomic background, and motivation. In addition, students follow diverse paths either according to their interests, or according to the necessities, such as academic or financial requirements. Some graduate from their entry majors in 4 years, some need more time, some transfer to different disciplines, and some leave their universities. Accordingly, providing Higher-Ed decision makers with an accurate summary of these diverse student characteristics is a necessity to help them make better data-informed decisions for improving students success. In this regard, any data mining methodology that can convey valuable patterns from student data sets in clear and informative fashion will be valuable. In this study, we discuss the development and use of such a visual tool based on the Sankey Diagram. It presents students progress and mobility patterns in an easily understandable format, was developed using open source software, and was used by several departments of a research intensive Higher-Ed institution of more than thirty-thousand students during their academic review process. This paper provides a general discussion about how these visuals could be used in Higher-Ed institutions by discussing problems that can be addressed, detailing the data-needs, the development methods, comparisons with other reporting methods, and how they were used in actual practice.

A. Oran
Institutional Research, CU Boulder; E-mail: ali.oran@colorado.edu

A. Martin
Ecology and Evolutionary Biology, CU Boulder; E-mail: andrew.martin-1@colorado.edu

M. Klymkowsky
Molecular, Cellular & Dev. Biology, CU Boulder; E-mail: michael.klymkowsky@Colorado.edu

R. Stubbs
Institutional Research, CU Boulder; E-mail: robert.stubbs@colorado.edu

Keywords Educational Data Mining · EDM · Learning Analytics · LA · Higher Education · Data Visualization · Open-Source

1 Introduction

Data mining is the process of discovering useful patterns from large amounts of data in an automatic or a semiautomatic way [1]. 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 [2]. Along with the advances in Data Mining, EDM has also seen considerable growth over the past two decades [3]. During this period, a very diverse group of studies have been proposed touching different problems in higher education using EDM (see surveys [4–7]).

The recent surge of interest in EDM studies wasn't only driven by simple interests of trying out the recently developed Data Mining techniques to Higher-Ed problems, but more by urgent needs to offset challenges Higher-Ed institutions have faced recently by harnessing 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) [8]. 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) [8–11]. These changing enrollment patterns are making it harder for Higher-Ed decision makers to develop long-term plans. Another challenge has been the 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 [12]. Moreover, in the near future, one might expect to see the continuation of these trends as some of these countries' future strategic plans aim to attract even more international students [13,14]. Finally, as a new generation of students are entering our universities, some proven practices that have previously ensured student success are in need of adjustments for to best service a new and changing student population.

To help address these challenges, recent advancements in Data Mining have given Higher-Ed institutions needed analytical tools to harness the campus-wide data sources for identifying possible areas for improvement and making data-informed decisions. Accordingly, in a nationwide study in 2018, it was noted that most institutions were making investments on both Descriptive and Predictive Analytics, either for improving student outcomes, or for more efficient delivery of programs or services [15]. Among these data mining tools, Data Visualization has

been one of the primary methods of approaching the problems and conveying the patterns in the data. In this context, the well-known saying 'A picture is worth a thousand words' perfectly summarizes the importance of these visual aids. With the recent advances in Data Visualization methods, along with the emergence of Big Data technologies, in the past decade Data Visualization has become even more essential for Data Mining problems. Following these advances, Data Visualization software carries significant new capabilities, allowing their users to go beyond traditional visuals and to bring more insightful analyses to their work. Besides the expanded capabilities, Data Visualization software has improved on the user-interface and simplified the once dreaded coding aspect of the visualization process. Currently, data practitioners from a wide variety of professions are able to utilize a variety of new visualization methods to bring much needed data insights in brief development time spans. Accordingly, Higher-Ed institutions are also seeing a change from only relying on traditional visuals (Line graphs, Pie charts etc.) to trying new visuals for gaining better insights from the ever expanding campus-wide data sources, such as Financial, Academic, and Personnel data. In this context, some of the proposed methods remain experimental and it is yet to be established which will be effectively deployed at a Higher-Ed institution. However, some others have seen actual use, and a greater discussion on these proven visuals is needed to ensure further progress.

Motivated by the advances in data visualization, we have been primarily involved in developing practical open-source visualization tools, that can be used by relevant stakeholders at Higher-Ed institutions to understand students' progress and mobility patterns after they are admitted. Understanding these patterns over time and across different academic units could help institutions enable better decisions for using limited resources effectively, and improving student success. In this regards, Graduation and Retention Rates are commonly used as measures of institutional effectiveness. Yet, too few of our students graduate and too few graduate on time as noted in a recent study by American Academy of Arts and Sciences [16]. Thus, the development of analytic tools that can help administrators, departments, and faculty make better decisions to improve rates of student success is an urgent need at most of our institutions. In this light, developing these tools with practical open source software is preferred in order to benefit the widest range of institutions in terms of their budget to support data analysis efforts. In this study, we detail the efforts of developing this type of an advanced visualization tool that is based on the Sankey diagrams to summarize the progress of students through a curriculum in an easily understandable way. These visuals have been developed for several departments at a research intensive higher-ed institution of more than thirty-thousand students, that were going through their Academic Review process, and were used in actual practice. In general, they can be deployed within any Higher-Ed department to identify students' progress and mobility patterns in a clear and concise way, when needed. Identification of these patterns can help departments notice various aspects of an academic program that might need improvements, such as locations of student bottlenecks, poor designed curriculum, and performance of transfer students. While the prospect of facilitating better discussions and actions within departments makes Sankey-based visuals valuable to Higher-Ed practitioners, they haven't been studied in detail to be generally accepted. In this study, we notice this missing information and provide a discussion of the Sankey-based visuals by detailing the data-needs, the development meth-

ods, comparisons with other standard reporting methods used in Higher-Ed, and providing some of our own results.

We start our discussion by reviewing relevant studies. Afterwards, we will introduce the essential problems commonly discussed in Higher-Ed in regards to students' progress. We will discuss ways to answer those problems, and discuss presentation methods for the analysis. In that regard, 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 the academic review process, the software aspects, and finally discussing Sankey visuals for two different majors.

2 Related Work

There have been a considerable number of studies in EDM regarding students' academic performances, and their progresses through majors. In addition, another considerable number of studies have approached similar problems under the Learning Analytics (LA) community, which have slight differences in their approaches compared to the EDM [17]. In order to provide our readers with a tractable review of similar works in this vast number of studies, we find it useful to start our discussion with a general categorization of these works, and briefly discuss each group, before covering the relevant Sankey-based approaches in detail. Unfortunately, there hasn't been a census to properly categorize these works, and each study/survey has introduced its own categorization. For our brief discussion, we start by distinguishing 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 in the future as the "Predictive Methods".

Institutions have been using Descriptive Analytics for years to understand students' performances, and subsequently to take necessary action when needed, such as reshaping their entering classes, refining policies etc. In recent years, Predictive Analytics have been added to these efforts [18]. 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 [19]. Purdue University (Course Signals) [20], University of Phoenix [21], and Capella University [22] are just a few examples of the universities that have utilized such systems. Logistic Regression has been a common method for the predictions [21, 22], yet, more advanced methods from Machine Leaning have also been tried as well [23]. 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.

On the other hand, Descriptive Methods constitutes a wider group of studies, some being used in Higher-Ed as proven methods for years. The challenge for data practitioners is choosing the best method for analyzing the data of interest within limited time frames and afterwards presenting the findings to the decision

makers with an impactful presentation. In this context, Visual Methods could be more effective for conveying the results of the analysis than others because of the nature of human cognition. Compared to narratives, or numerical tables they can more easily get the attention of decision makers, and also convey more information. Accordingly, for describing students' progresses pattern visual methods could be ideal.

Among Visual Methods, a distinction could be made according to methods' focus. We refer to those methods whose focus are visualizing students' data under an aggregate analysis as the "Aggregate Visuals". Whereas we define the "Tracking Visuals" as those methods whose focus are visualizing each student's data separately. In practice, Aggregate Visuals can visualize characteristics of particular cohorts or groups of students, e.g. visuals for average time-to-degree of Chemistry majors that was accepted between 2010 to 2014. And, Tracking Visuals can visualize characteristics of each student, e.g. visuals for a particular student's course progress over time as a means of tracking the student's progress or other visuals to monitor the student's activity. 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 [24], Bakharia and Dawson's SNAPP [25], Chiritoiu et. al's Students' Activity visualization tool, and Capella University's Competency Map [22,26] are some of the well known examples of this group of visuals.

In contrast, Aggregate Visuals can instead be used to 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 different departments at the same university. Among Aggregate visuals, while the traditional data visualization methods (such as line plots, pie charts etc.) have been prevalent in almost every Higher-Ed institution, with the recent advances in visualization software more informative visual tools have recently become feasible. In this context, for understanding students' aggregate progresses and mobility patterns, 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, 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 [27]. 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 [28], Material Flows in the EU [29], Land Cover Dynamics in Urban Planning [30], and Temporal Visualization of Diabetes in Health Informatics [31].

However, the use of Sankey Diagram's for visualizing students' progress through an Educational Institution has been very 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 [32]. Later in the same year, Morse, in his Masters thesis [33], proposed a multi-column Sankey diagram for the progress of students, and provided a technically detailed discussion of the possible data wrangling needs of such an effort. Heileman et. al.'s 2015 study [34] improved upon Morse's work to develop more diverse Sankey diagrams that were used to debunk some myths regarding student progress. In 2018, similar Sankey visuals were again used to understand students progress, first at Budapest University of Technology and Economics [35], and second at Department of Computer Science at University of Central Florida [36]. Notwithstanding these efforts, from a general perspective many aspects of using Sankey diagrams for students' progress in Higher-Ed have yet to be realized. Here we attempt to fill this void by providing a general-level discussion of the use of Sankey Diagrams in Higher-Ed based on our experiences in developing the necessary tools for a group of departments undergoing academic review at our institution.

3 Problems of Interest in Higher-Ed

An important challenge for institutions of higher education is how to enable students to complete their studies and earn their degrees within reasonable times. Success is commonly measured by graduation and retention rates, and accordingly, institutions seek to maximize these rates. Several factors conspire to reduce these rates, including, but not limited to, budget cuts that reduce available resources, curricular structures that impose barriers to student success, various instructor and course effects that can impose academic "bottlenecks", and changes in student characteristics. Additionally, there are differences among departments and divisions within universities and colleges -different acceptance criteria, graduation requirements, and course sequencing- that influence graduation and retention. One strategy that may contribute to increasing graduation and retention rates and generally improving students' education experience is to make student cohort data available to departments so that educators can better perceive how students progress through their curriculum. With a better understanding of the gain and loss of students in a department over time, common stumbling blocks that could be hindering students academic progress could be identified, and potential solutions enacted. In this regard, in Higher-Ed one of the most important questions of interest for college administrators is:

1. Why are the students leaving their departments?

While this question can be brushed aside as just an example of young students testing out different majors, 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 or presented. 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.

To answer our initial question accurately, one should also consider two related aspects. First, 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. Second, students eventual destinations can also yield actionable insights for departments and curricular designers. For instance, students switching to unrelated disciplines to their original majors (e.g. science to humanities) display a different pattern compared to students who switch between similar majors or who leave the university altogether. In this regard, this type of an analysis can also provide a clear summary of which majors are more welcoming to students from other majors, and which do not allow for such transitions (either intentionally by high academic requirements, or unintentionally as a result of very distinct course structures). Also, a similar analysis about the transfer students' entry majors into the institution (from another 2- or 4- year institution), and their subsequent progress patterns can be used to identify more suitable course entry points for them. Accordingly, we also consider the following questions whose answers can yield useful insights for Higher-Ed decision makers:

2. When are the students leaving their majors?
3. Where are the students leaving to?

Our goal is to provide accurate answers to these three questions to determine *"Students' Progress and Mobility Patterns in Higher Education".*

4 Identifying Students' Progress and Mobility Patterns

Data practitioners need to pay attention to how to convey the results of an analysis to an audience that may or may not have much experience or even interest in the analyzed data. Failure to understand the expectations and the experiences of the recipients of the analysis makes even the best data studies useless, and results in a waste of resources. Accordingly, over the course of our meetings we developed guidelines for our work, which we followed closely in order to be able to connect with our audience, the faculty members and administrators, and have our resulting analysis be useful for them. In these guidelines, we paid close attention to the following factors:

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

With these guidelines in mind, we approach the *Identification of Students' Progress and Mobility Patterns* problem through a two step process:

1. Data Extraction and Data Analysis
2. Data Visualization

Data Extraction, requires the preparation of the needed data set, which is followed by Data Analysis to identify the patterns. Data Visualization step involves developing the best visual presentation for the identified data patterns for the audience.

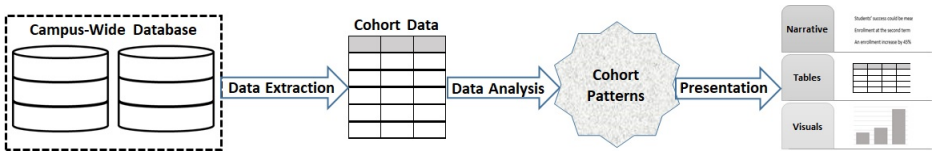


Fig. 1 General-level flow of Students' Progress and Mobility Pattern analysis.

In our work, it involved developing the Sankey-based visuals. Figure 1 summarizes these efforts, and we detail each step in the succeeding subsections.

4.1 Analyzing Students' Progress

The first effort in this process is *Data Extraction* which yields the data for the cohort of interest, and involves standard database manipulation techniques, such as joining of different tables from the Campus-wide databases, filtering of the data according to the needs, etc.. In general, the cohort should be defined according to faculty's interests. In our study, we were particularly interested in the progress of students who entered the university in a particular year and who have declared their first major in a particular major. Note that this group included students who were undeclared majors in the beginning of their freshman year, but eventually chose the same first major. One can extend the analysis to consider multiple cohorts, with different entry years and first majors. Cohort Data is the longitudinal data of this cohort from the time of entry to a particular time. We have chosen the last semester as the final semester, yet different time-frames could be set depending on the interests. The cohort data should include the end-of-term majors for each student in the cohort for each term until graduation. Accordingly, in the most minimal sense, the cohort data could be an Excel or .csv file with the following columns (variables):

"Student ID", "Year/Term", "Major"

Additional information such as "Degree Date", "Enrollment Status" can be useful as well. Once the data for the cohort is prepared, we move on to identifying the patterns in it. For student's progress analysis, this pattern analysis boils down to correctly identifying the separate groups student cohorts belong at each semester (or quarter). These separate groups can be the different majors the cohort have chosen over time, or it can be a collection of majors (e.g. a group representing all majors under Natural Sciences). The choice of these groups eventually affects the complexity of the presentation in the succeeding step, and a trade off will need to be made between the detail of information provided in the presentation versus interpretability of the presentation.

We provide a simple example of a sample cohort of 10 students in Fig. 2 which summarizes the aforementioned efforts. We work on this example in the subsequent sections to have a clearer discussion. In this figure, Group 1 is the group of students who were actively seeking a degree in some major (e.g. Chemistry). Group 2 is the group of students who left that major for another major in the same university by the end of a given term. Group 3 is the group of students who left the university altogether, and Group 4 is the group of students who graduated (from any major) by the end of a given term.

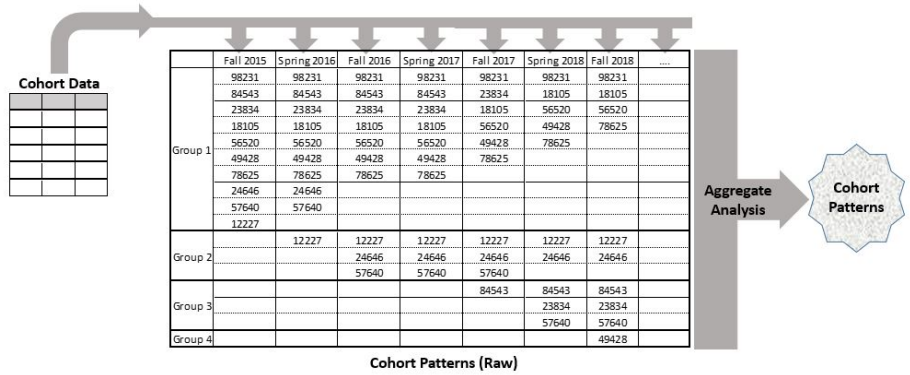


Fig. 2 A simple example of analyzing a sample cohort data of 10 students' progress over several consecutive semesters.

Depending on faculty's interests, other groups could be included to characterize students' educational progress at a deeper level. For instance, graduates can be grouped according to the majors that granted them their degrees, or the entry cohort could be split into groups that reflect their genders, or SAT/ACT scores. Once these patterns are identified in the raw data format, it comes down to applying needed aggregate analysis and afterwards choosing the best method to present the results in the most effective fashion.

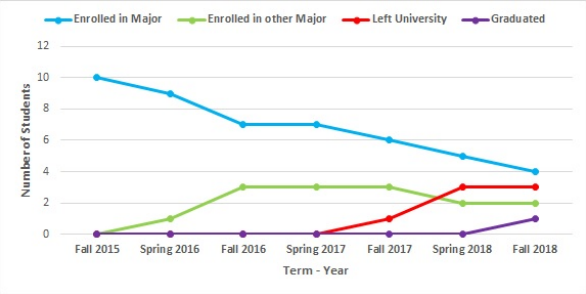
4.2 Visualizing Cohort Progress Patterns

An aggregate analysis can yield a wide variety of information about the general characteristics of a cohort. In our case, we are primarily interested in the number of students belonging to each group at each semester/quarter, so that the change of those numbers over time reveals student progress and mobility patterns. Figure 3 shows the table that summarizes the aggregate analysis on our sample data from Fig. 2.

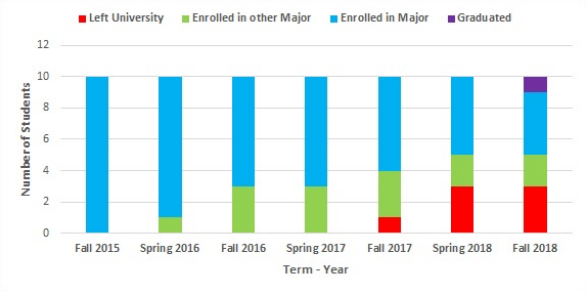
	Fall 2015	Spring 2016	Fall 2016	Spring 2017	Fall 2017	Spring 2018	Fall 2018
Group 1	10	9	7	7	6	5	4
Group 2	0	1	3	3	3	2	2
Group 3	0	0	0	0	1	3	3
Group 4	0	0	0	0	0	0	1

Fig. 3 Aggregate analysis for the sample data from the previous figure.

After identifying the cohort numbers at each term for each group, 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. 3 is one option; it clearly summarizes the number of students in each group over time. In fact, for short time spans of 4 or 5 semesters, its may well be adequate for institutional discussions. However, for longer time spans, such as the 12 semesters needed to analyze "6-year graduation rates", it is not a good option. While it presents the actual size of each group, it doesn't provide information



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



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

Fig. 4 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)"

about groups' relative sizes. While this may not be an issue for a 10-student cohort with 4 groups, for larger cohorts with more groups such tabular data needs to be supplemented by other columns of data (or a second table) providing information about relative cohort sizes (e.g. percentages). In return, this would just make the presentation material more complex making accurate interpretations problematic. Accordingly, the use of visual techniques might be the better option to have a clearer presentation.

Some of the traditional graphical ways of presenting the cohort patterns could be using the Line Charts or the Stacked Bars, shown in Fig. 4a and Fig. 4b respectively, for our sample data set. Both figures clearly show the number of students who have left their original major, left the institution, or who have graduated by the end of each term. Accordingly, from these figures one can understand the general patterns when students change their major, when they leave the institution, when they graduate, and each of 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 growth of each group within the original cohort, when compared to the Line Chart.

However, one thing that is commonly missing in all these, the Table, the Line Chart, and the Stacked Bar Chart, is the flow information about students moving from one group to another. For instance, in our sample date set, between Fall 2017 and Spring 2018 we can notice that two more students from the original

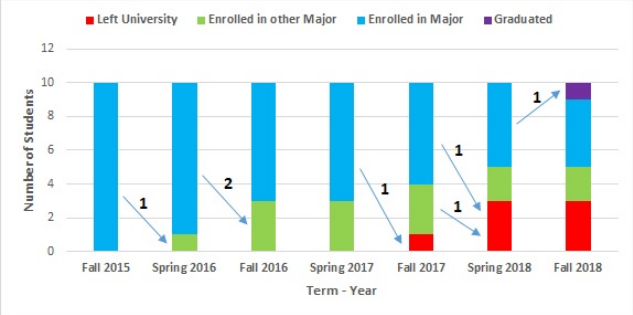
cohort have left the university. Yet, it is unclear from neither of these figures or the table whether those two students were still enrolled in their original majors, or had they already switched majors and left the school after trying out different majors. Yet, this missing flow information is essential for our presentation because it completes our understanding of how different student groups evolve over time, which in return gives us useful insights about the students and the institution, such as student satisfaction about the institution or how easily students could pursue other majors if they are not satisfied in their first major. A similar question of flow can also be asked about the graduates appearing as yellow patterns in Fall 2018. Again, these figures or the table can't provide the information about where the students graduated from, that is whether from their original majors, or other majors. Now, going back to the raw data in Cohort patterns table in Fig. 2, by comparing student IDs, one can find out that one student has left the institution from his/her original major, and the other student from another major, that is after trying different majors. And, the graduate was from Group-1, that he/she graduated from the original major he/she had enrolled in.

Ideally, the final presentation should contain enough information so that going back to analyzing the raw data wouldn't be necessary. One way to include this missing valuable information is by adding more layers (groups) to these visuals or the table in Fig. 3. For instance, the "Left University" group can 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 institution's 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, particularly for people unfamiliar with student data. As an alternative, one could enhance tables and figures with the flows information superimposed on them. That is, one can add new information about the flows on top of the existing figure and the table, as shown in Fig 5a and in Fig 5b. These additions provide us the essential information that was missing in the previous presentations. 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 just double the time to produce them.

One should note that Fig. 5b is in fact a primitive Sankey diagram. Accordingly, one can use Sankey diagrams from the beginning to avoid reproducibility issues, and accordingly the sample cohort's progress could be visualized as in Fig. 6. This figure is similar to the Bar Chart Fig. in 5b 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,

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

and accordingly identify the possible bottlenecks students face. Now that we have introduced the Sankey for visualizing the progress of our sample cohort, we proceed to discuss how these type of visuals were used in practice at our institution for the academic review process.

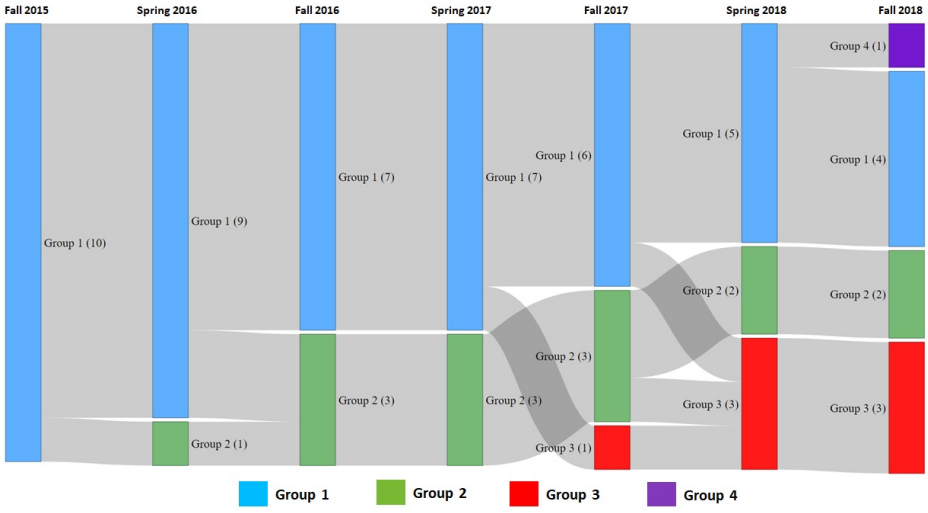


Fig. 6 Sankey for the sample cohort data of 10 students over several consecutive semesters, with each cohort population in parantheses.

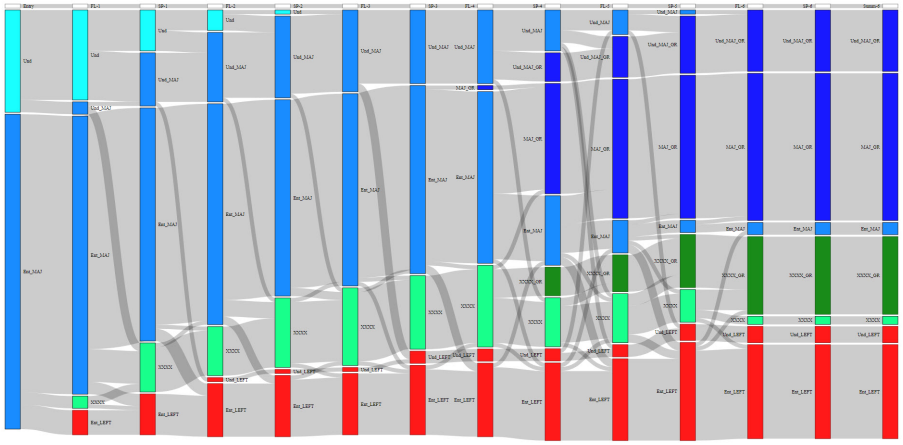
4.3 The Academic Review process and Sankey in practice

The Academic Review process is 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 [37]. The programs participate in this review on a seven-year cycle, and the process includes review committees that are comprised of campus constituents and discipline experts external to the institution. It is a similar effort to what many other Higher-Ed institutions go through under similar names, eg. "Program Review" at Northwestern University [38] and at University of Washington [39], and "Academic Program Review" at Cornell University [40].

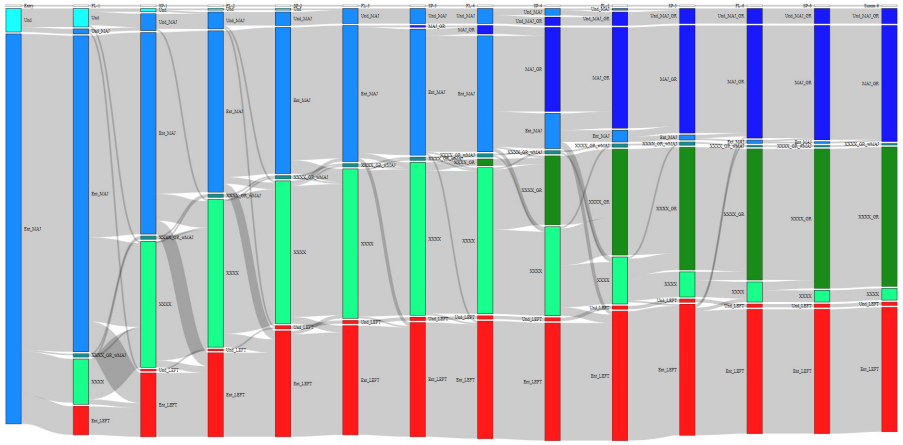
As a first step, academic programs engage in self studies during which they address a series of planning queries. These queries are designed to solicit strategic information and to document the units organizational qualifications. Topics include role and mission, centrality, outcomes, and diversity goals. Programs own personnel prepare the reporting, yet, separate standardized unit data profiles from the Office of Planning, Budget, and Analysis also complement the self-study. This step is followed by the Internal Review, the External Review, and lastly by the Academic Review and Planning Advisory Committees evaluation of the self-study, internal, and external reviews and with making recommendations for unit improvement [37].

Theoretically, a review process should initiate productive discussions in departments about central issues impacting students' educational experiences. Accordingly, it requires the query of the correct data sources, the assembly of the right amount of information and accurate and easily understandable metrics for the faculty, so that discussions could focus on the needed areas of improvement. To help achieve this goal, for the 2018-19 academic review cycle the Institutional Research office developed a Sankey-based visualization platform for identifying the student progress and mobility patterns for Natural Science departments. This process was carried out in closely collaboration with some of the faculty members involved in the review process. The Sankey-based presentations served as an alternative to traditionally used visuals and metrics that addressed the retention and graduation rates of academic units, but were unable to convey some critical information altogether in a single visual, as discussed in the previous sections. Having a good balance of informativeness and understandability, these visualizations provide the faculty a compact, all-in-one summary of a department in terms of student mobility and progress. Academic leadership can use this critical information to arrive at a deeper appreciation 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 allows faculty to compare their students' progress directly with students from other academic units to highlight similarities and differences.

In Fig. 7a and in 7b we provide two Sankey-based visuals from our work, reflecting the progress and mobility patterns of two cohorts of undergraduate students, admitted into two Natural Science departments in the same year, over several semesters. Similar to Fig 6, we used Blue-shades for students who were enrolled at the particular major, Green-shades for students who have switched majors, and Red-shades for students who have left the university. In Blue and Green shades, the darker colors (towards right side) represent the group of students graduating in



(a) Department-1



(b) Department-2

Fig. 7 Undergraduate progress and mobility patterns in two separate departments for a particular entry year.

each group. And, Light Blue was used for "undeclared students" that are students who had not chosen their majors in their first semester. Between these figures, the differences in patterns are quite obvious, even though both departments are within the Natural Sciences. One of the primary difference is in the departments' loss of students to other departments. While the proportion of entry cohorts who left the university without a degree was similar in both departments (Red-Shades), the loss of students to other departments is quite different (Green-shades), with only a small portion from the first department's cohort moving to other majors, where as more than half of second department's cohort eventually graduating from a different department. Another important difference is in the proportion of "Un-

declared" students (light-blue). While a quarter of department-1's entry cohort is initially undeclared, in department-2 this group is considerably smaller.

There are several other subtle patterns that become clear after a more detailed look at the diagrams, such as time to degree differences between departments, graduation rate differences between initially declared and initially undeclared students, etc. By varying the entry cohort according to different criteria (e.g. gender, first-generation status etc.), more patterns can be observed for understanding the progress of different groups of students at a Higher-Ed institution. In general, these visuals can act as good starting points for faculty discussions focused on existing curricula, degree requirements, and other departmental practices for improving student success for a diverse group of students.

4.4 Software Aspects

In this last section, we briefly discuss the software aspects of developing the Sankey-based visuals. In our work, we tried to adhere to using open-source software to minimize the cost of our efforts. With the advances in visualization techniques and software in the past couple of decades, this choice didn't hinder our efforts in any way. In fact, the chosen open-source software to develop the Sankey visuals, R [41], and the needed package networkD3 [42], has provided us with abundant online discussion forums that were extremely helpful whenever we had to improve our codes. Compared to some other propriety software, having this type of open discussion forums were clearly an advantage.

After the cohort data was extracted from the campus-wide database, it was imported into the R environment for the *Data Analysis* step, which was identifying the student progress patterns. As discussed in section 4.1 this pattern analysis boils down to correctly identifying the separate groups that student cohorts belong to at each semester, and yielded the tabular "Cohort Patterns (Raw)" structure shown in Fig 2. 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. When we were working with traditional visuals, such as Line Graphs and Bar Charts, the aggregate analysis follows, and afterwards the visuals show these aggregated patterns. For Sankey visuals, before the aggregate analysis we must to identify flows between different groups in each consecutive semesters. This can be accomplished by finding the common student-IDs in different groups in each consecutive semesters. After identifying these student IDs, aggregate analysis follows, and yields flow information, that is, the number of students that are either staying in their groups, or moving to another one in the next semester. When using 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.

These are the basic efforts needed to generate Sankey visuals for visualizing students' progress. Depending on the particular data set, a few extra efforts such as cleaning the data, removing unnecessary groups (groups with no students) might be needed. Yet, from our experience, the technical difficulties are manageable for most IR Analysts with basic programming skills. And, networkD3 allows lots of flexibility allowing further modifications on many aspects of the visuals, such as coloring, spacing of bars etc. [42].

5 Conclusion

In this study, we detailed our efforts at our institution to develop a visual tool based on the Sankey diagram that could convey student progress and mobility patterns to faculty, and other stake holders. This visual can illustrate the diverse paths students follow through a Higher-Ed institution after admission, in a clear, and informative way. In this regard, it can be used as an alternative to traditional reporting tools, such as Tables, Line Graphs, and Bar Charts, all of which miss the essential flow information between different groups students belong at each semester. For a general discussion, we initially discussed similar works in literature, then introduced the problems in Higher-Ed that can be addressed by the proposed visuals, and finally detailed our approach. For clarity, we provided the reader with a sample cohort data, and compared Sankey-based visuals with other reporting tools to demonstrate Sankey's advantages. We also discussed how these visuals were used for an academic review process and detailed some of the observed patterns to show possible data exploration opportunities Sankey can provide.

References

1. 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.
2. R. S. Baker, *Data Mining for Education*, 2005, vol. 19.
3. 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.
4. ——, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, no. 1, pp. 135 – 146, 2007.
5. 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.
6. 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.
7. 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.
8. *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
9. *Current Term Enrollment Fall 2012*, National Student Clearinghouse Research Center Report, 2012. [Online]. Available: https://nscresearchcenter.org/wp-content/uploads/CurrentTermEnrollment-Fall2012.pdf
10. *Current Term EnrollmentFall 2015*, National Student Clearinghouse Research Center Report, 2015. [Online]. Available: https://nscresearchcenter.org/wp-content/uploads/CurrentTermEnrollment-Fall2015.pdf
11. *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
12. *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
13. *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
14. *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

15. 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
16. *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
17. 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.
18. *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
19. 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.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. [Online]. Available: https://www.capella.edu/blogs/cublog/measure-learning-with-capella-university-competency-map
27. R. L. Harris, *Information Graphics: A Comprehensive Illustrated Reference*. New York, NY, USA: Oxford University Press, Inc., 1999.
28. K. Soundararajan, H. K. Ho, and B. Su, "Sankey diagram framework for energy and exergy flows," *Applied Energy*, vol. 136, pp. 1035 – 1042, 2014.
29. 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.
30. 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
31. 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.
32. 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.
33. C. Morse, "Visualization of student cohort data with sankey diagrams via web-centric technologies," 2014.
34. 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.

35. 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.
36. 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.
37. Webpage for university's academic review process. University of. [Online]. Available: XXXX
38. Program review. Northwestern University. [Online]. Available: https://www.adminplan.northwestern.edu/program-review/
39. Program review. University of Washington. [Online]. Available: http://grad.uw.edu/for-faculty-and-staff/program-review/
40. Academic program review. Cornell University. [Online]. Available: https://irp.dpb.cornell.edu/academic-program-regulation/academic-program-review
41. 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/
42. 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