Using Venn, Sankey, and UpSet Diagrams to Visualize Students' Study Progress Based on Exam Combinations

Alexander Askinadze, Matthias Liebeck, Stefan Conrad
Heinrich Heine University Düsseldorf, Germany
{askinadze, liebeck, conrad}@cs.uni-duesseldorf.de

ABSTRACT: Educational dashboards allow educators to gain insights about their students and their learning progress. It is essential to understand why students may drop out of the university. In our educational dashboard, we used a combination of Venn, Sankey, and UpSet diagrams to perform an in-depth analysis by investigating the effects of individual courses and their combinations. We present our visualizations based on student data from a computer science course at a German university.

Keywords: student visualization, Venn diagram, Sankey diagram, UpSet diagram

1 INTRODUCTION

Educational institutions are collecting more and more data that can be analyzed. Learning Analytics (LA) is a research field that deals with the analysis of such data and is defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs." (Siemens, 2010)

Arnold and Pistilli (2012) presented the "Course Signals" (CS) system, a successful example of the use of learning analytics at Purdue University. Based on available data in the CS system, an algorithm predicts students' risk levels. The predicted risk levels can be used by instructors, for example, to intervene by posting a traffic signal indicator (green, yellow or red indicating the likelihood of being successful or unsuccessful) on a student's dashboard or by e-mail messages. Their evaluation showed that the usage of CS delivered significantly higher retention rates than without.

While predictive models can be used to identify at-risk students and intervene as mentioned above, Charleer, Klerkx, and Duval (2014) mention that such predictive models may be black boxes and give users no insight into the reasons for the decisions made. This motivates us to use learning analytics dashboards that apply information visualization techniques to display students' data so that educators may gain deeper insight and become empowered to make own decisions, rather than relying on automated decisions (Verbert, Duval, Klerkx, Govaerts, & Santos, 2013).

With our work on a dashboard for educators in an academic context, our research belongs to the research area of LA. We want to provide a better insight into the overall study progress of different student cohorts by proposing new visualizations that show which exams students passed jointly. Students can drop out of studies for many reasons (Sagenmüller, 2018). Among other things, it may be due to certain mandatory exams. The proposed visualizations, especially for the analysis of drop-out students, may help to find potential causes of drop-out associated with the curriculum. The core questions that we want to answer are: 1) How can we visualize which exams or exam combinations our graduates or drop-out students passed until a selected semester?; 2) How can we use the visualizations to identify certain exams that are difficult for students?; 3) How can we compare the study progress of different cohorts of students in one visualization?

2 RELATED WORK

In the past, many visualization techniques have been applied to present in an educational dashboard. Charleer, Klerkx, and Duval (2014) report that most dashboards consist of basic visualizations, e.g., bar and line charts or scatter plots. Gray, Teahan, and Perkins (2017) list further common visualization techniques, such as pie and donut charts or tables, as well as some advanced visualization techniques related to the information visualization community, such as tag clouds, stream diagrams, heat maps, sunburst diagrams, aster plots, bubble diagrams, and radar plots. Sankey diagrams are another visualization technique that has been used in the learning analytics context. They were originally intended to visualize the energy efficiency of a steam engine. Sankey diagrams can be used to display study progress over several semesters. Morse (2014) investigated cohorts of students and utilized Sankey diagrams to visualize how the students changed their major, graduated or dropped out throughout multiple semesters. Similarly, Heileman, Babbitt, and Abdallah (2015) used Sankey diagrams to debunk myths about student progress by visualizing student cohorts from different majors as being enrolled, having graduated or stopped studying.

In contrast to previous research, we need visualizations that show which exams students have passed until a given semester. This would be possible, for example, with simple bar charts, but then the information about which exams were passed jointly would be lost. In addition, the time component that shows in which semester the respective exams were passed, would be lost. Therefore, we rely on and combine the strengths of Venn, Sankey, and UpSet diagrams as an advanced visualization technique to perform an in-depth analysis by investigating the effects of individual courses and their combinations on the overall study progress in the next chapters.

3 METHOD

In this chapter, we discuss the three proposed visualization techniques in general and show their application on real data in the following chapter. Usually, visualizations depict courses or modules individually. We now focus on our core questions and investigate which exams are passed jointly. For our educational dashboard, we implemented a drop-out analysis method which allows us to analyze students who have completed their studies without a degree until a certain semester by selecting multiple courses. For the analysis of the drop-outs in the first semesters, we recommend selecting the courses that are scheduled for the first semester in the curriculum.

Table 1: Pros & cons of Venn, UpSet, and Sankey diagrams in our dashboard.

	Venn	UpSet	Sankey
shows jointly passed exams until the selected semester	x	x	x
temporal information (shows information per semester until selected semester)			x
can compare different cohorts in one diagram			x
visualization scales well for more than three exams		x	
exam combinations are displayed clearly and intuitively	x		

One approach we have not seen before in the context of LA are Venn diagrams which can display, for example, intersections, unions, differences, and symmetric differences. With Venn diagrams, we want to investigate which exams are passed before students drop out. For each exam combination, the corresponding intersection set shows the number of students who passed all exams in this intersection. Venn diagrams are easy to understand for a small number of exams. However, they are impractical for a large number of analyzed exams (> 3) since the number of combinations grows exponentially with the number of exams. We use the implementations of Venn diagrams from the JavaScript library D3[1] (Bostock, Ogievetsky, & Heer, 2011) and venn.js[1]. D3 is extendable and uses modern web technologies, such as SVG and canvas, for the interactive visualization of complex data.

Therefore, we propose to additionally use the visualization technique UpSet (Lex, Gehlenborg, Strobelt, Vuillemot, & Pfister, 2014). For every module combination, UpSet also visualizes the number of jointly passed exams with a bar chart for which the legend below indicates which combination corresponds to each bar, as visualized in Figure 1 (b). The UpSet diagram is sorted in descending order, allows to hide specific combinations (which is not possible in the Venn diagram) (Khan & Mathelier, 2017), and can display a high number of exam combinations since the diagram grows horizontally and can still be scrollable. A disadvantage of UpSet over Venn diagrams is that all intersections for a specific module do not need to be directly side by side due to the sort order and are, therefore, more difficult to interpret. Therefore, we show both visualizations side by side in our dashboard to combine their respective strength. For the implementation, we relied on upset.js[1].

Although Venn and Upset diagrams are suitable visualization techniques for displaying combinations of exams cumulatively up to a particular semester, they cannot be used to show study progress for different semesters. Additionally, they are not able to visualize different cohorts, e.g., graduates and dropouts, in one figure at the same time. We suggest using Sankey diagrams for both purposes. Each node in a Sankey diagram is a combination of exams in a specific semester, as visualized in Figure 1 (c). When analyzing many exams, the visualization of the names for each exam combination is challenging. Therefore, we decided to use a binary encoding. For n different exams, there are 2^n possible combinations. Below the visualization is a legend, which assigns the binary numbers to their corresponding combination of passing (1) and failing (0) exams. For example, the notation 1_100 represents those students that only passed the first of three selected exams at the end of the first semester. For the implementation, we used the D3 extension d3-sankey[1]. Since Sankey diagrams display detailed temporal data, they are powerful, yet difficult to understand, especially for a high number of exams and a time span of multiple semesters. We summarized the advantages and disadvantages of the three visualization methods in Table 1.

4 ANALYSIS / APPLICATION ON REAL DATA

We present our visualizations based on students' progress data of different cohorts of a German computer science degree program. The courses of the first-semester curriculum are Calculus I, Linear Algebra I, and Programming. All of these courses are mandatory for the bachelor's degree. Since already 25% of all dropouts occur until the end of the second semester and we are interested

[1] github.com/d3/d3; github.com/benfred/venn.js; github.com/chuntul/d3-upset; github.com/vasturiano/d3-sankey

in helping students as early as possible in their studies, we start by investigating students that drop out before their third semester. For the analysis, we filtered out all students who did not have a single exam attempt and for whom no study history data is therefore available. Figure 1 (a-c) shows passed exams of students that dropped out of their studies until the end of their second semester in Venn, UpSet and Sankey diagrams, respectively. However, by examining the passed exams, it is not clear at all whether the students even tried to pass the exams. Therefore, we propose to additionally display exam attempts, as depicted in Figure 1 (d).

Our visualizations show that students are less likely to pass the mathematical exams (and especially both of them) than the computer science course Programming. Dropouts that are able to pass a math exam are, to a large extent, able to also pass the other two exams. From the Venn diagram in (a), we can see that most of the dropouts do not pass a single course. Compared with the UpSet diagram of exam attempts (d), we notice that the majority of them at least tries to pass the Programming exam. Also, we see that more than half of the students that try to pass the Programming exam do not even try to pass one of the math exams. Different reasons might account for this behavior, for example, that it may be easier to meet the examination requirements for Programming or that math subjects appear too difficult for the students and they, therefore, do not even register for the exam. In any case, the math exams seem to be the most significant obstacles.

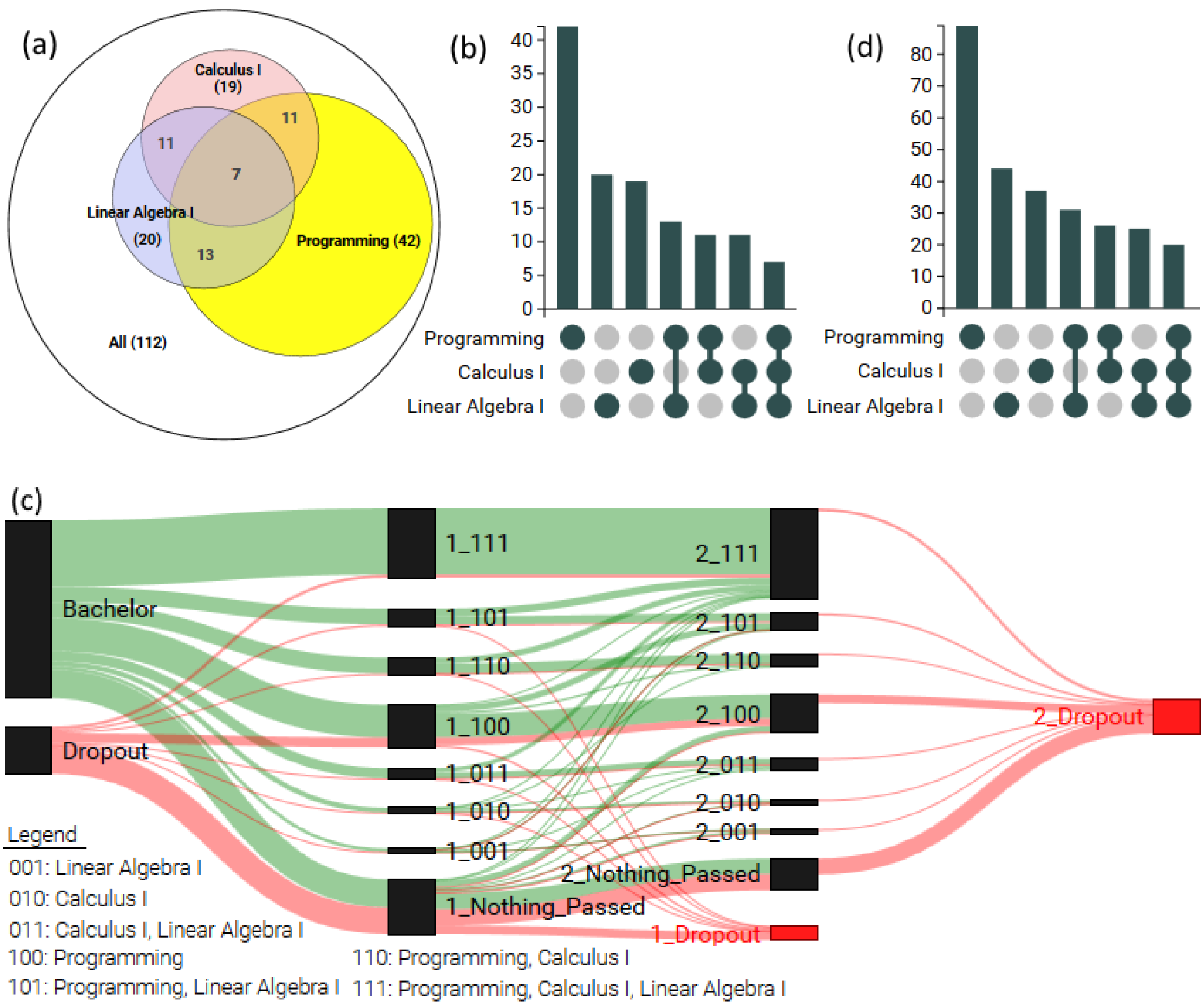

Figure 1: Venn (a), UpSet (b) & Sankey diagrams (c) showing jointly passed exams of students who dropped out of studies till the end of the second semester. UpSet (d) diag. shows exam attempts.

The Sankey diagram (c) visualizes the study progress of students that dropped out until the end of the second semester as well as graduates. We can see that many students who have passed all three freshman courses after the first semester will most probably not drop out until the end of the second semester. Also, most of the dropouts before the end of the second semester did not pass any of the freshman courses or only the Programming course. Since teachers have little chance of helping students who drop out due to personal reasons, it now makes sense to focus in detail on the courses and compare the behavior of dropouts with that of graduates. We can see that most of the students (including graduates) who have only completed the Programming exam at the end of the first semester will not have passed the math exams until the end of the second semester. This suggests problems with the math exams for which solutions should be found.

5 CONCLUSION

In this work, we presented three visualizations that are suitable to illustrate combinations of exams. The first two visualizations, the Venn diagram and the UpSet diagram, can be displayed side by side as they complement each other. Together, they can deliver new informative insights into the study behavior of dropouts. However, they cannot visualize temporal study progress over multiple semesters. Therefore, we proposed the usage of Sankey diagrams that are able to visualize the study progress by combining passed and failed exams. The presented visualization methods allowed us to gain new insights into the data that we would not have seen without them. In the future, we want to evaluate our visualization on different data, make our dashboard more interactive, and use Sankey diagrams to visualize the temporal progress of course exercises.

REFERENCES

Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on LAK*. ACM, 267-270.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D³ Data-Driven Documents. *IEEE Transactions on Visualization & Computer Graphics*, (12), 2301-2309.

Charleer, S., Klerkx, J., & Duval, E. (2014). Learning Dashboards. *Journal of Learning Analytics*, *1*(3), 199-202.

Gray, C. C., Teahan, W. J., & Perkins, D. (2017). Understanding our Analytics: A Visualization Survey.

Heileman, G. L., Babbitt, T. H., & Abdallah, C. T. (2015). Visualizing Student Flows: Busting Myths About Student Movement and Success. *Change: The Magazine of Higher Learning*, 47(3), 30–39.

Khan, A., & Mathelier, A. (2017). Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC bioinformatics*, *18*(1), 287.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE transactions on Visualization and Computer Graphics*, 20(12), 1983-92.

Morse, C. (2014). *Visualization of Student Cohort Data With Sankey Diagrams via Web-Centric Technologies* (master's thesis). University of New Mexico, Albuquerque, New Mexico.

Sagenmüller, I. (2018). Student retention: 8 reasons people drop out of higher education. Retrieved January 13, 2019, from u-planner.com/blog/student-retention-8-reasons-people-drop-out-of-higher-education

Siemens, G. (2010). 1st International Conference on Learning Analytics and Knowledge 2011. Retrieved January 16, 2019, from https://tekri.athabascau.ca/analytics/

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, *57*(10), 1500-1509.

ACKNOWLEDGEMENTS: This work was partially funded by the IST-Hochschule University of Applied Sciences.