Examining the CSE Admissions Pipeline

CSE 512 Spring 2016 - Final Project

Sachin Mehta, Sam Crow, and Wesley Lee
fp-sacmehta-samcrow-wesleytlee

***Abstract*— Of the approximately 3000 students which take CSE 142 each year, 2000 (67%) continue to CSE 143 and only 750 (25%) end up applying to the CSE major. It is of interest to gain a better understanding of these dynamics: which students continue in this pipeline and how do they differ from students who do not? Our visualization serves as a framework to explore these questions and provide preliminary answers by combining flow charts of students through the CSE pipeline with statistical tests for mean differences in covariates between groups.**

***Index Terms*—Sankey diagram, Welch's t-test**

I. Introduction

Over the last several years, there have been many articles about the lack of gender and ethnic diversity in Computer Science at both the classroom and industry levels [1,2,3]. Underlying these articles is a desire to better understand which factors drive students to seek a career in computer science. Members of the UW Computer Science department, such as Stuart Reges and Alexander Miller, are interested in this problem on a local scale, focusing on the dynamics of student interest from the time students take their first CSE class (CSE 142) to the time they apply and are accepted/denied to the major. For our final project we focused on building a visualization that would allow for both the exploration of these dynamics and the preliminary testing of any resulting hypotheses.

Our visualization is based on anonymized data obtained from 3,365 students who took CSE 142 between the 2012 and 2014 school years. These students took surveys at the beginning of CSE 142 that asked questions about their backgrounds and interests in Computer Science. Demographic information about these students and grade information from CSE 142 and CSE 143 (if taken) are available. For students who applied to the CSE major, we have access to application scores given by the admissions committee as well as admission results. It should be noted that the data does not have information on all students who took CSE 142 during this range and has some missing values which we assume are missing at random when computing statistics.

Our proposed visualization will have two major properties. First, the ability to display the "flow" of students from CSE 142 through the entire CSE pipeline. Second, the ability to compare differences in covariates for two groups of individuals (and implicitly the ability to restrict the visualization to certain populations of interest). We hope the visualization will serve as a tool that allows researchers to more rapidly identify transitional stages (e.g. CSE 142 to CSE 143, CSE 143 to Application) of interest within the CSE pipeline and yield insights into potential factors that may be driving these behaviors. The visualization should be able to answer questions such as "Are individuals from certain groups more likely to drop out at certain stages of the pipeline than individuals from other groups?" and "What differentiates students who continue at a given stage from those who do not?"

II. Related work

As far as we can tell, there has been little work attempting to visualize multi-stage student data. The closest existing visualizations focus on a single transitional stage of interest. [4] utilizes a chord diagram to display transitions between majors and careers while allowing for filtering base on time period. [5] uses multiple line charts to track university admissions data (e.g. numbers of applicants, acceptances, enrollment) over time while allowing for filtering by interest, sex, and year.

A few challenges that differentiate our data and visualization from those existing are the longitudinal nature of our data for each student, and the large number of covariates available for each student. We choose to use Sankey diagrams as the base of our visualization and augment the diagrams with detailed statistics on the covariates.

Sankey diagrams are flow charts in which flow width is used to display flow quantity and annotations are used to describe individual flow components. They are named after Matthew Sankey, who used them in 1898 to track the flow of steam through a steam-engine and comment on the engines efficiency. Sankey diagrams are often used to display energy flows (see [6] for a review), but they also have been used in a wide variety of other applications as well. Most pertinently to our own application, [7] proposed using Sankey diagrams to help prospective students when deciding between multiple colleges. In his example, Sankey diagrams are used as a categorical version of parallel coordinates to display demographic information about students at any given school. Using fake data, [8] suggested that Sankey diagrams could be used to track the declared majors of students over multiple years.

Sankey diagrams have developed significantly since they were first introduced in the nineteenth century. For example, the use of color allows for the simultaneous display of multiple flows [9]. More recently, there has been work introducing interactivity into Sankey diagrams [10, 11]. Some techniques

that have been considered include linking between Sankey diagrams and other graphs, animating the flows in Sankey diagrams, and highlighting the paths corresponding to subsets of the flow. For our visualization, the primary desired interactivity component is the ability to compare groups of students corresponding to different flow components and/or Sankey diagrams.

III. Methods

The crux of our visualization are Sankey diagrams visualizing the flow of students from CSE 142 to applying to the major and the resulting admission decisions, as well as dropouts along this path.

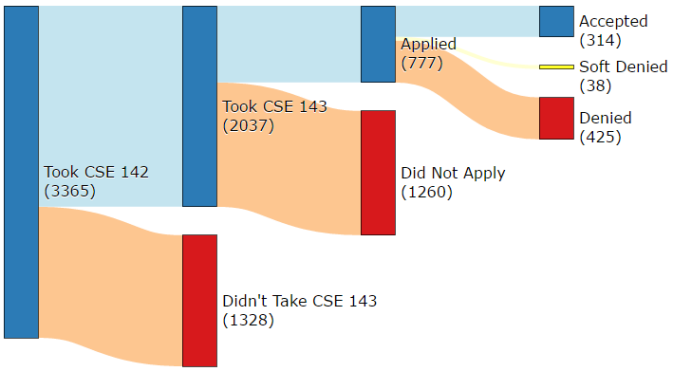

Figure 1: Sankey diagram of CSE pipeline.

In Fig. 1, a sample Sankey diagram from our visualization is provided. There are three stages to the Sankey diagram: whether or not to take CSE 143, whether or not to apply to the major, and the admissions decisions. Flow / node height encode the percentage of students (out of all those who took CSE 142) remaining at the corresponding stage of the diagram. Raw numbers are then encoded in parentheses under the node names, and hovering over the links causes a tooltip to appear which provides the percentage of students from the source node who go to the target node (e.g. what percentage of student who applied were accepted). Color and vertical position are used to emphasize the distinction between students who continued at each stage (blue, highest) and students who dropped out (red, lowest).

To enable the comparison of the CSE pipeline for two groups of students ("group 1" and "group 2"), we place two Sankey diagrams side-by-side and place a set of filters above each. Since there are a large number of potential filters, filters can be added with the "add a filter" tool and removed by clicking the corresponding close button, helping reduce the visual footprint of the filter option without affecting the flexibility of the potential filters. Heights between the two Sankey diagrams can be compared if interested in the overall percentage of students in group 1 at some stage compared to overall percentage of students in group 2 at some stage (e.g. percentage of women who take CSE 142 that are accepted compared to the percentage of men). We choose not to encode raw student counts as height to allow for the comparison of two groups with very different sizes.

An alternative to the side-by-side Sankey diagrams would have been to try and combine them into a single Sankey diagram and use color to differentiate between the two groups. However, we felt like this approach would have had the same issue with comparing groups of very different sizes.

We complement the Sankey diagrams with a pair of tools to select a link of interest (e.g. "CSE 143 → Applied to Major" to examine only students who applied to the major) for each of the two diagrams. Ideally, the visualization would have allow users to select a link by directly clicking on the Sankey diagram, but this functionality proved difficult to implement. Using one of these selection tools will cause a corresponding table at the bottom of the page to populate with summary statistics (mean and standard deviation) for all available covariates for the selected group of individuals. Tables are used rather than graphs since the covariates lie on very different scales and thus are hard to encode in a single graph.

However, since the raw summary statistics are less meaningful without a counterfactual, we place above these summary tables a bar plot displaying the approximate scores (test statistics) of the difference in mean between groups 1 and 2 for all covariates.

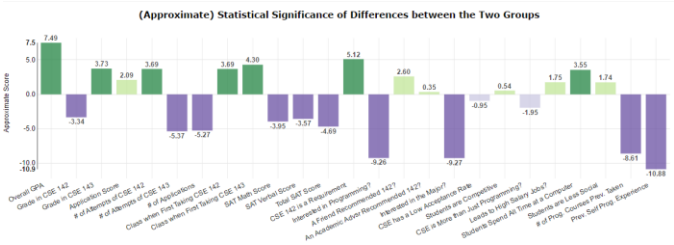

Figure 2: Boxplot displaying the (approximate) statistical significance of differences in mean between groups 1 and 2 for each covariate.

Statistically significant (see the next paragraph) differences are highlighted with darker colors than statistically insignificant differences to draw the user's attention. Positive differences (green) correspond to the covariate tending to be higher in group 1, while negative differences (purple) correspond to the covariate tending to be higher in group 2. Colors are just chosen not to coincide with the blue and red of the Sankey diagram, which otherwise may lead to faulty analogies. Again, we plot score rather than the raw difference in means between the groups due to the different ranges of the covariates. However, one can obtain the difference in means for any variable by hovering over the corresponding bar.

To calculate the score for each covariate, we use Welch's test for differences in mean. Suppose (μ_1,σ_1,n_1) denotes the mean, standard deviation, and number of observations for some covariate in group 1, and (μ_2,σ_2,n_2) describes the same covariate for group 2. Then the score is defined by

$$s = (\mu_1 - \mu_2)/(\sigma_1^2/n_1 + \sigma_2^2/n_2)^{0.5}. \quad (1)$$

Assuming that the covariate is normally distributed within each group, the score follows a t-distribution with complicated degrees of freedom under the null hypothesis of no difference. For large sample sizes n_1 and n_2, we can approximate this t-

distribution with the normal distribution. Lastly, we can apply the Bonferroni correction to account for the multiple comparisons issue (due to calculating t-tests simultaneously for many covariates) and maintain a familywise error rate (probability of making one or more type 1 errors) of 0.05.

Of course, these statistical tests should be taken with caution and really only serve as a way to draw the user's attention to potentially interesting covariates (as opposed to a rigorous testing tool). Unfortunately, it is quite possible to break the assumptions underlying the test; examples include comparing two groups that are not disjoint, comparing groups with small sample size, and having normality assumptions violated for certain covariates. Ultimately, we feel the ability of bar plot to draw the user's attention to certain covariates outweighs the risk of misuse.

IV. Results

Figure 3: Snapshot of final visualization, comparing all women and men who took CSE 142.

Let us consider a potential application of our visualization: examining the differences between the men and women in the CSE pipeline. There is already an about 2 to 1 imbalance of the sexes in CSE 142, but we mainly concern ourselves with whether or not this imbalance improves in the CSE pipeline.

After selecting the appropriate filter for each group, we will obtain the view shown in Fig. 3. By comparing the heights of the nodes corresponding to students who didn't take CSE 143 after CSE 142, we see that a significantly higher percentage of women do not end up taking 143. We can examine the box plot to see if there are any difference between the groups besides sex that might explain this disparity. As seen in Fig. 4, on average women are more likely to consider CSE 142 a required class, display less interest in majoring in CSE, and have less experience with programming entering CSE 142.

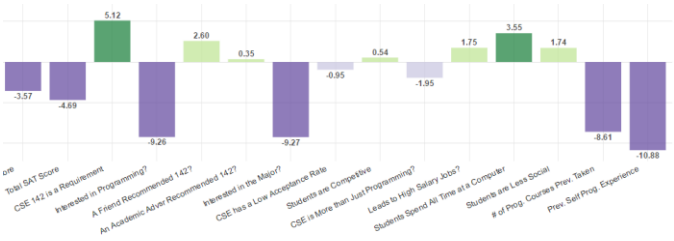

Figure 4: Zoomed-in portion of the boxplot from Figure 3. Green denotes that women had a higher average than men for that covariate and vice versa for purple.

This suggests the following non-sex-related hypothesis for this phenomenon: a higher proportion of the women are taking CSE 142 as a requirement for something else and do not go on to take 143.

Under this hypothesis, one would expect that if we only compared women and men who did end up applying to the major, they would not show similar differences among these covariates. Thus, we can test our hypothesis by selecting the stage "CSE 143 → Applied" under both Sankey diagrams. The boxplot and summary tables automatically update to reflect this selection.

Figure 5: Zoomed-in portion of the updated boxplot after filtering to only students who applied to the CSE major. Compare with Fig. 4.

From Fig. 5, we see that there are significant differences between the aforementioned covariates even when restricting our population of interest to applicants to the CSE major. This appears to be evidence against our hypothesis. Instead, the evidence is more consistent with the theory that a smaller proportion of women apply to the major because they simply enter the CSE pipeline with less predilection for the CSE major.

From here, we could continue testing and refining our theory, adapting our questions based on the observed results. For example, natural follow-up questions that one could answer with the visualization include "How do the women who apply to the major differ from the women who do not?" and "What factors are associated with students who initially state they are less interested in the CSE major eventually applying?"

V. Discussion

At the poster session Tuesday, interested users seemed to mainly fall into one of two categories. Some individuals had first-hand experience with the CSE admissions process and

were interested to see how their individual experience tended to align with their peers. These individuals tended to select many filters in order to make a group as similar to themselves as possible. They spent most of their time looking at the Sankey diagram and the summary tables, examining admission rates and mean GPA scores. They were also less likely to take advantage of the comparison tool since they had an implicit baseline to refer to (themselves).

Other individuals were more interested in questions of diversity. These users spent less time managing filters and more time comparing various groups with the boxplot. This usage was more in line with our expectations, but it was interesting to see what other users might try to use our visualization for.

By far, the most common piece of feedback we received / observed was difficulty interpreting the boxplot. At the time, covariates were not color-coded based on sign, so it often wasn't clear that scores were based on the group 1 mean minus the group 2 mean. Thus, it wasn't clear what a large positive value meant compared to a large negative value. Adding colors with a more prominent legend hopefully will lead to less confusion. Further improvements might include more closely connecting the boxplot to the Sankey diagrams from a visual perspective, such as adding a green tint to the left Sankey diagram and a purple tint to the right, with the hope of making the boxplot more intuitive to users. A more descriptive, qualitative label on the y-axis of the boxplot might also help.

We also received some feedback with regards to our set of filters. Comparing subsets (e.g. those who were accepted into the major and those who were not) for the same group of individuals was rather time consuming, since one would have to duplicate all the filters before being able to compare them. This would be remedied with an option to copy filters between the two groups. Regarding individual filters, one individual commented that it was a bit hard to finely control the numerical sliders corresponding to a continuous variable.

Lastly, in the Sankey diagram it was often of interest to calculate conditional percentages (e.g. percentage of CSE 143 students who applied to the CSE major), information that is only contained in the tooltip. Making this information more salient to the user seems desirable, if challenging.

VI. Future work

While we believe the visualization is a promising first step towards understanding the data at hand, it can definitely be improved. The CSE department has data on students even after their acceptance to the major, so the Sankey diagram could be extended to study the behavior of CSE majors in upper-level classes. An interesting challenge with upper-level classes is that the flow of students would be less regimented, with more options and classes that can be taken in any order.

A major challenge with this project was handling the data, so integrating the visualization with SQL would reduce this burden and bring more flexibility if the data source was updated or extra information on the students was available. From a methodological viewpoint, a better system for handling multiple attempts at a class or multiple applications in the Sankey diagrams would be highly desirable. While we try to handle this issue with filter options, a visual approach (possibly an extension of the Sankey diagram) may ultimately be more appealing and/or better suited for users.

Acknowledgment

We wish to thank Stuart Reges for bringing this interesting problem to the course and Alexander Miller for making the data available to us. We also wish to thank the course staff for their help and guidance for this project.

References

[1] K. C. Mason, "Computer science's diversity gap starts early," in *PBS NewsHour*, 2014. [Online]. Available: http://www.pbs.org/newshour/updates/teaching-coding-kids-key-closing-fields-diversity-gap/.

[2] M. Johnson, "The Computer Science Pipeline and Diversity: Part 1 - how did we get here?," in Google Research Blog, 2015. [Online]. Available: https://research.googleblog.com/2015/07/the-computer-science-pipeline-and.html.

[3] I. Najarro, "Addressing ethnic diversity in computer science," in Stanford Daily, 2015. [Online]. Available: http://www.stanforddaily.com/2015/01/16/addressing-ethnic-diversity-in-computer-science/.

[4] S. Devadoss, "Impact of major on career path for 15600 Williams College Alums," in Williams College, 2012. [Online]. Available: https://web.williams.edu/Mathematics/devadoss/careerpath.html.

[5] "Admissions counts," in Cornell University. [Online]. Available: http://irp.dpb.cornell.edu/tableau_visual/admissions.

[6] M. Schmidt, "The Sankey diagram in energy and material flow management," Journal of Industrial Ecology, vol. 12, no. 1, pp. 82–94, Feb. 2008.

[7] "The crux of presenting school admission data,". [Online]. Available: https://www.visart.io/the-crux-of-presenting-school-admission-data/. Accessed: Jun. 9, 2016.

[8] "Using Sankey diagrams to tell a story with data," in Duke Today, 2015. [Online]. Available: https://today.duke.edu/node/191539.

[9] M. Schmidt, "The Sankey diagram in energy and material flow management," Journal of Industrial Ecology, vol. 12, no. 2, pp. 173–185, Apr. 2008.

[10] H. Alemasoom, F. Samavati, J. Brosz, and D. Layzell, "EnergyViz: An interactive system for visualization of energy systems," The Visual Computer, vol. 32, no. 3, pp. 403–413, Nov. 2015.

[11] P. Riehmann, M. Hanfler, and B. Froehlich, "Interactive Sankey diagrams," IEEE Symposium on Information Visualization, pp. 233–240, 2005.