CS171 Final Project Process Book

Ben Shryock

Table of Contents

Initial Project Proposal

Submitted April 3, 2015

Background and Motivation

This project involves data collected from the last two years of Computer Science 51 (Spring 2014 and Spring 2015) at Harvard. I acted as a Head TF of the course during this time, and have been part of the teaching staff for four other courses during my time at Harvard. I am interested in education (with most of my experience being in Computer Science education at the university level), and I believe that CS51 has unutilized data that could lend insights into the experience of students and how they progress over the course of a semester.

Project Objectives

There are multiple populations that this visualization could benefit, including both students and CS educators.

For students of the course, I hope to answer questions related to performance (grades) and effort (time spent per problem set). This allows students to better understand how they're doing in the class and what steps they should take to improve their experience.

For CS educators, there are a number of interesting questions that can be investigated from this data. What is the relationship between time spent on a problem set and score achieved? What is the relationship between a student's final grade and their behavior in the course's online discussion forum? How much progress do students that initially describe themselves as less comfortable make compared to students that initially describe themselves as more comfortable make over the course of the semester? I hope to create exploratory visualizations that reveal any answers to these questions, in addition to the more explanatory visualizations discussed above.

Data

The data has been collected over the past two years of CS51. Thus far, it comprises the results of an introductory survey, grades per problem set, time spent per problem set, and statistics from the online discussion forum (Piazza). This is four separate data sources, at least some of which can be linked through common fields (*i.e.,* email address, HUID). About ~300 students have taken the course each of the past two years, generating thousands of data points across these different sources.

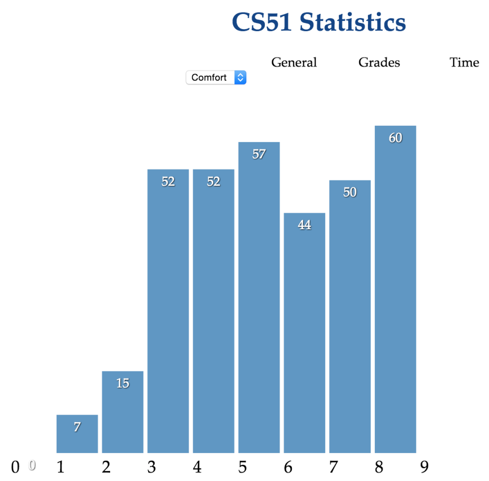
Data Processing

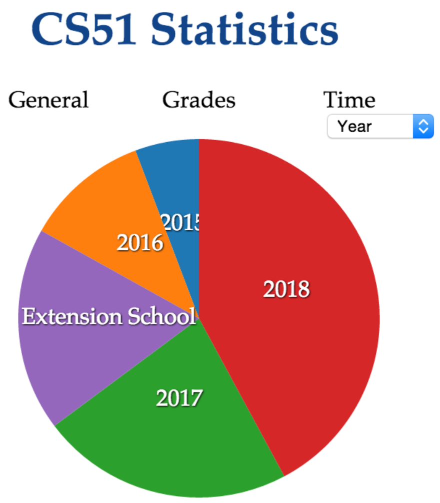
I've already performed most of the data cleanup for basic visualizations, such as histograms of time spent per problem set and grades per problem set. This has been done through the creation of Python scripts to perform general data wrangling and the removal of sensitive information. The outputted CSV or JSON files are then loaded using d3.

There will continue to be data processing necessary to hide any personally identifying information, particularly when dealing with sensitive information like grades. Further data processing will be necessary to connect these four separate data sources through shared attributes, unlocking a greater variety of comparative visualizations.

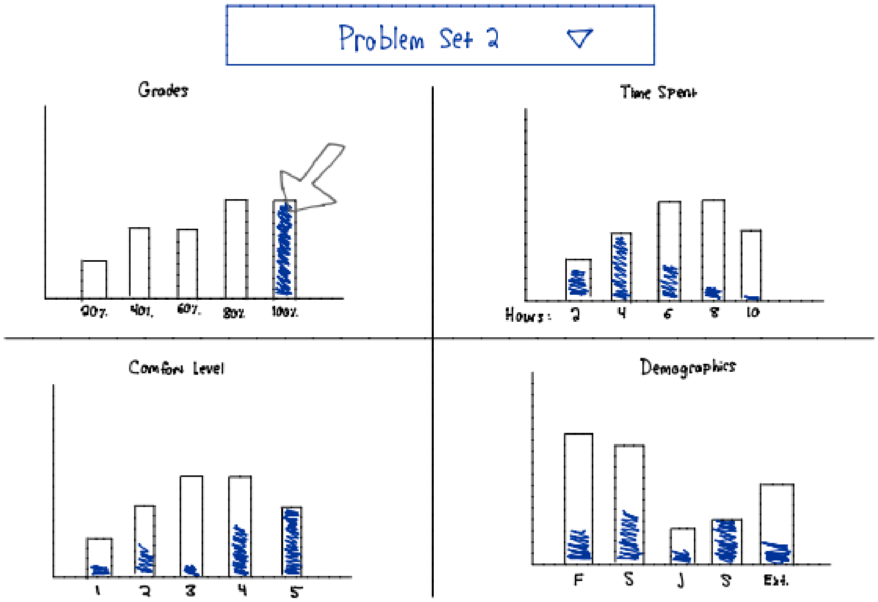
Visualization

The student facing designs will be relatively simple, designed to effectively communicate basic information about the course as a whole. I have already implemented the basics of these charts, a sample of which are shown below. The first is a histogram showing the self-described comfort level of students (on a 1-10 scale), taken from a survey given at the beginning of the course. The latter is a pie chart depicting the student makeup of CS51 2015 by year. (Note: these are not final products, and as such have elements like overlapping text and unclear labels)

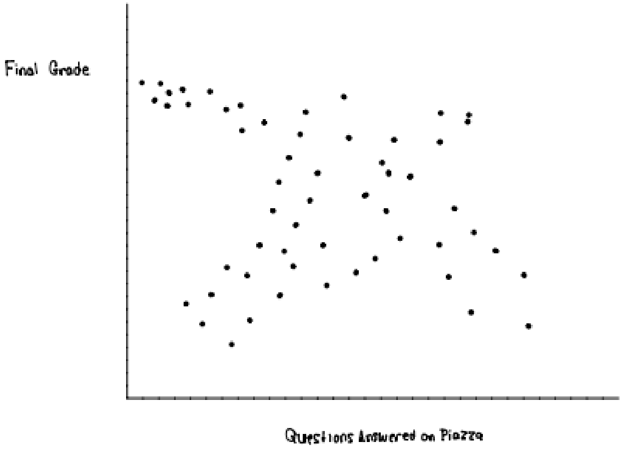




Though the other student facing designs are primarily histograms (problem set scores, time spent per problem set), there are more interesting potential visualizations intended for CS educators. A couple possibilities are sketched and explained below.



This visualization shows several charts in one screen. It initially provides a general representation of things like grades and time spent on a specific problem set alongside things like comfort level data and demographics data. Though this could function in a variety of ways, one idea is to have the “grades” histogram be the main visualization. When a bar in the grades histogram is moused over, the bars in the other graphs fill with color (become stacked bar graphs) to represent the data that corresponds to the moused over bar and the data that doesn’t. In the example shown in the diagram, when the 100% score is moused over in the grades diagram, the other graphs reveal that those students that received a 100% were likely to spend less time on the problem set, come from a higher comfort level, and are evenly distributed across the various class years. (Note: the feasibility of this graph is contingent on being able to link students between the separate survey, grades, time, and discussion forum data).



Another potential visualization, particularly useful for exploration, would be to examine the relationship between final grades and various discussion forum behaviors. For instance, are the students that frequently ask, view, or answer questions more likely to perform well in the course?

Another possible area of investigation is in the performance of low-comfort and high-comfort students on the first and last problem sets. Over the course of the semester, do the scores of the low-comfort improve? Do the students catch up to others that were initially more comfortable than themselves?

Must-Have Features

- A set of student-facing visualizations that answer the fundamental question of "how am I performing in this course" for at least 1 year of data
- A more interesting educator-facing visualization that provides insight into the performance and progress of students in an introductory CS course
- Some exploratory visualizations investigating the effect of various metrics on student grades

Optional Features

All Visualizations

- Incorporate multi-year data
- Incorporate transitions

Student-facing Visualization

- Add ability to input score/time spent, highlight appropriate histogram

Project Schedule

Deadlines

Sunday, April 5: Finish student-facing visualizations
Friday, April 10: Exploratory analysis of grade-related factors, linking data sources
Friday, April 17: *Milestone 1.* Educator-facing visualization draft
Friday, April 24: Finish educator-facing visualization
Friday, May 1: Set up website, finalize process book, complete peer assessment
Wednesday, May 5: *Project Due.* Submit project

Milestone 1 Update

April 17, 2015

Overview

The following few pages contain a discussion of the updates I've made to my final project (from plans to implementations) while preparing for the first milestone over the past two weeks. In general, I focus on only the things that have changed from the initial project proposal. Moving forward, I will update the process book on a more frequent basis.

Questions

As I continue to explore my data, I may revise the questions I'm seeking to answer. However, my visualizations currently focus on the question of "what factors relate to a student's performance in an introductory computer science course?".

Data

I have acquired data from three sources pertaining to Computer Science 51's 2015 offering. These include a welcome survey, student grades, and student behavior in our online discussion forum (Piazza).

Each of these data sources contains an email address, which I've used to connect the separate sources into singular data objects for each students. There were approximately 330 students that filled out the welcome survey. A number of these dropped the course (and some may have used different email addresses in these different data sources), which resulted in a total of 250 students for whom I have data from each of the three sources. Given our final enrollment numbers, this seems like a strong set of data.

Below is a list of the attributes I have extracted from each source:

Welcome Survey

- Class year (2015, 2016, 2017, 2018, or extension)
- Concentration
- Programming comfort level (1-10 scale)
- Operating System
- Whether or not the student has taken CS50

Grades

- Grades on PS1-PS7 and midterm

Discussion Forum

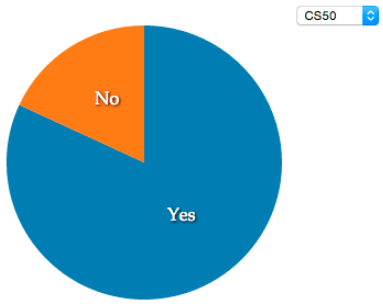
- Number of days on which the forum has been accessed
- Number of unique posts viewed
- Number of questions asked
- Number of answers given
- Number of contributions (includes questions, answers, and comments in followup discussions)

As mentioned in my initial project proposal, I also have access to data showing the amount of time students spent on each assignment. However, this data was only collected on a subset of the problem sets, only collected in 2014, and doesn't have a shared identifier with any of the other three data sources. As such, I may continue my project without utilizing this data.

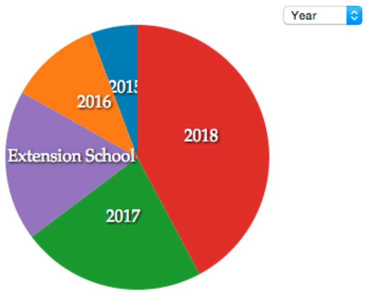
After using the students' email address to connect the three data sources, as much personally identifying information as possible was removed from the representation of a student. This is done with Python, and no identifying information should touch the browser.

Exploratory Data Analysis / Implementation

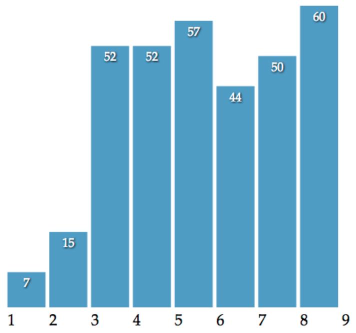
I began by separately analyzing the data in each of these sources. A sample of the created charts and their meanings are shown below.



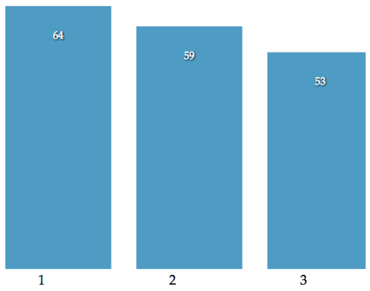
If a student has taken CS50



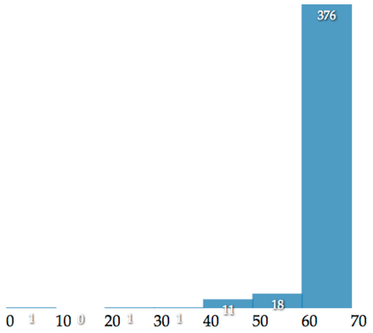
A student's class year



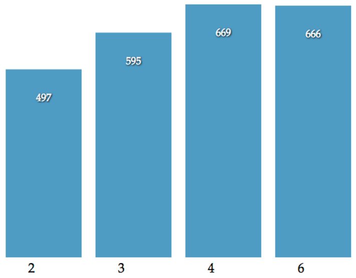
Comfort Level
(x: 1-10 scale, y: number of students)



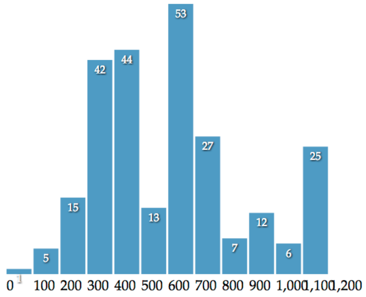
Average pset performance (x: pset number, y: average score out of 75)



Performance on a specific pset (pset 1 shown)
(x: score bucket, y: number of students in range)



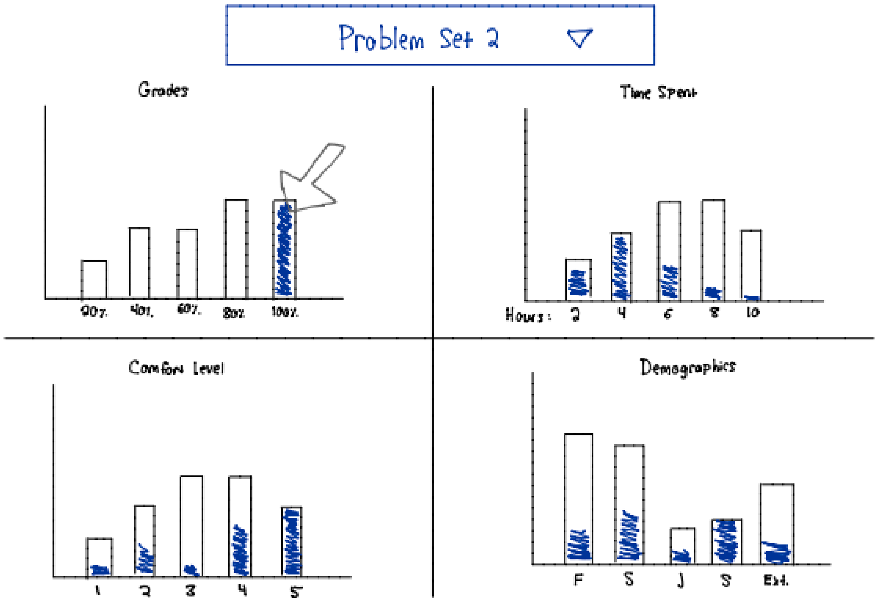
Average time spent on psets (x: pset number, y: number of minutes spent)



Time spent on a specific pset (pset 3 shown)
(x: minutes spent, y: number of students)

Design Evolution

I still am pursuing one of the ideas initially discussed in my project proposal.



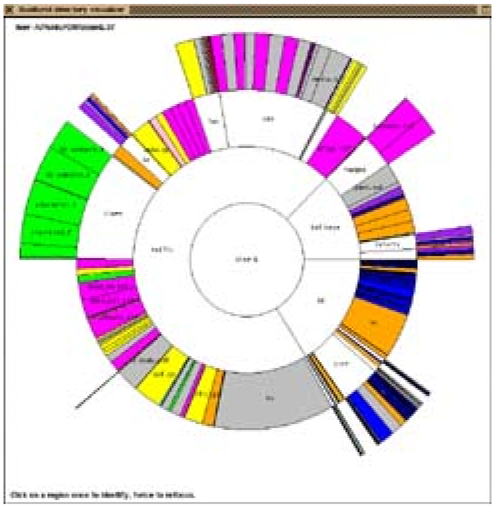
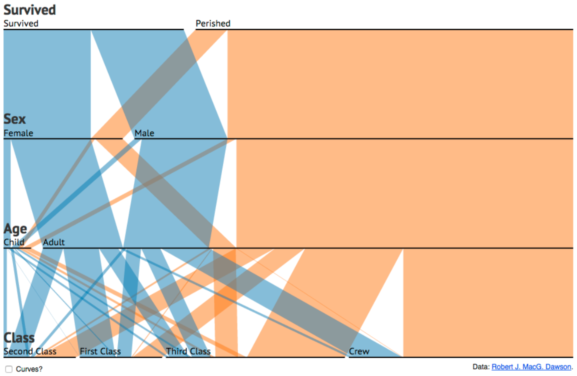
The general idea has not changed, and is reproduced below for ease of reference.

> “This visualization shows several charts in one screen. It initially provides a general representation of things like grades and time spent on a specific problem set alongside things like comfort level data and demographics data. Though this could function in a variety of ways, one idea is to have the “grades” histogram be the main visualization. When a bar in the grades histogram is moused over, the bars in the other graphs fill with color (become stacked bar graphs) to represent the data that corresponds to the moused over bar and the data that doesn’t. In the example shown in the diagram, when the 100% score is moused over in the grades diagram, the other graphs reveal that those students that received a 100% were likely to spend less time on the problem set, come from a higher comfort level, and are evenly distributed across the various class years.”

I have made two revisions to this idea in response to feedback from my advisor and design studio peer review. The first is to remove the central cross, making use of proximity to show the relationship between the graphs. The second is to allow for multiple bars in the primary chart to be selected instead of just one.

I feel this is a strong primary visualization due to its explanatory power. However, it does lack in visual appeal.

There are other visualization formats I plan to pursue. As in my initial project proposal, there is still exploratory analysis I'd like to do to assess potential correlations between the data attributes I've collected. Additionally, I've been considering other primary methods of visualization that could be more visually appealing that group bar charts. These include the sunburst and parallel sets visualizations (shown below).

	
Sunburst visualization	Parallel sets (visualization technique for multi-dimensional categorical data)