

Academic Dishonesty using GitHub

Patrick Pichardo

University of Texas at Austin

Department of Computer Science

Prof. Carol Ramsey

March 3, 2024

Academic Dishonesty using GitHub

By Patrick Pichardo, The University of Texas at Austin, Prof. Ramsey

Introduction:

CS313E lays the foundation for advanced problem-solving through coding, making it essential for any student aiming to pursue a software-intensive career to thoroughly understand the material. While numerous online resources such as Chegg, Course Hero, ChatGPT, and Gemini can be utilized for academic dishonesty, the focus of today's discussion is on a platform that lacks a paywall, making it the most accessible. This short paper aims to investigate how GitHub can facilitate academic dishonesty on CS313E assignments.

Gathering data:

To examine the extent to which students might use repositories titled similarly to the CS313E class (e.g., "CS313E", "MitraCS313E", "TeymourianCS313E", "RamseyCS313E", "Elements of Software Design"), I will be reviewing the most recently posted repositories within the past three year (from 02/2021-02/2024). I will conduct a selection of 10-15 repositories from each search term and compare the frequency with which we encounter an assignment from our CS313E class.

Conditions: there must be an exact match to our assignment's python files (e.g., ImageFill.py counts, Graph.py doesn't count even though it's a similar BFS/DFS assignment)

Logic: I'll be using Boolean logic for if a condition is met (e.g., True/False, 1/0)

Data:

Figure 1 Descriptive Statistics: CS313E

Variable	Obs	Mean	Std. dev.	Min	Max
numberspiral	15	.6666667	.48795	0	1
employeesa~s	15	.3333333	.48795	0	1
workinghard	15	.6666667	.48795	0	1
josephuspr~m	15	.3333333	.48795	0	1
expression~e	15	.6	.5070926	0	1
bstcipher	15	.1333333	.3518658	0	1
bstandvalt~e	15	0	0	0	0
imagebucke~l	15	.2	.4140393	0	1
topologica~t	15	.4666667	.5163978	0	1

For Figure 1, we observe the statistics for GitHub repositories with 'CS313E' in the title.

The table presents data on all nine assignments for this semester, with 15 observations for each.

'Number Spiral' and 'Working Hard' had the highest average occurrences, with a mean of approximately 0.6667, indicating that they were present in two-thirds of the repositories observed for 'CS313E'.

Figure 2 Descriptive Statistics: Elements of Software Design

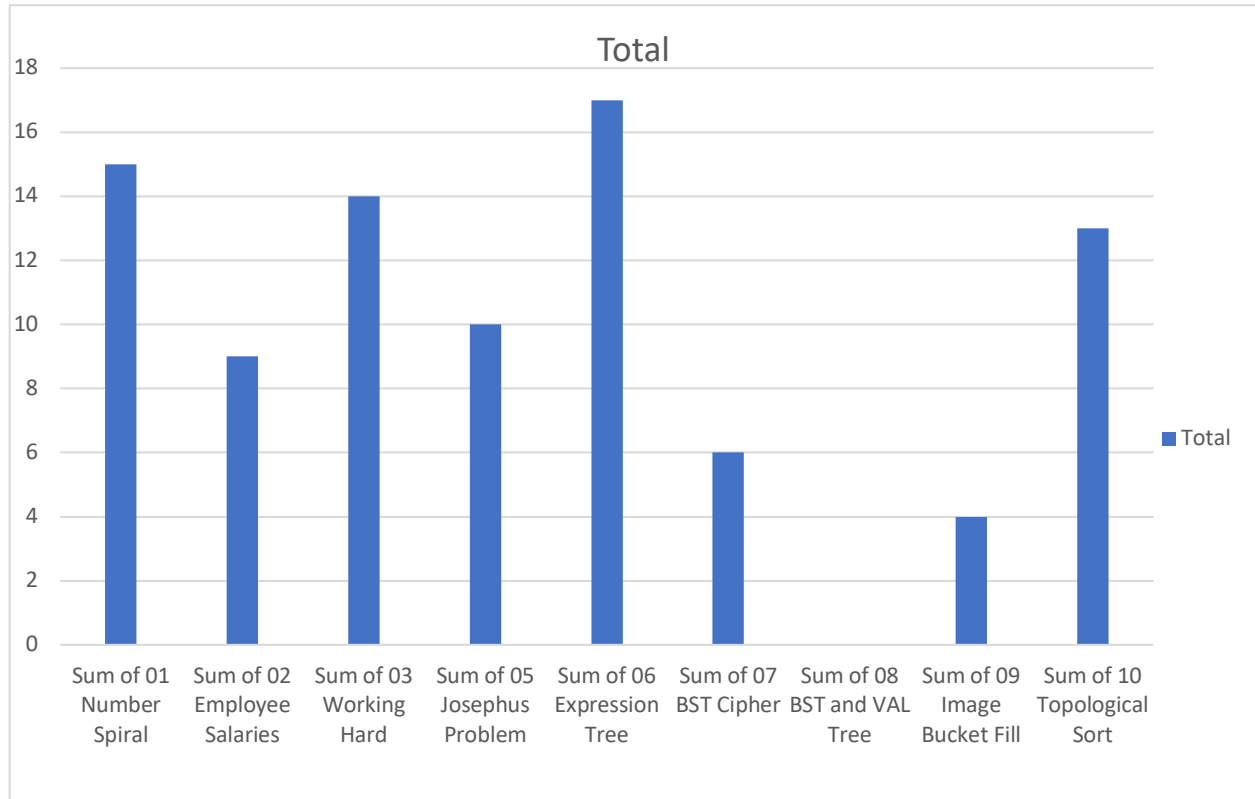
Variable	Obs	Mean	Std. dev.	Min	Max
numberspiral	15	.3333333	.48795	0	1
employeesa~s	15	.2666667	.4577377	0	1
workinghard	15	.2666667	.4577377	0	1
josephuspr~m	15	.3333333	.48795	0	1
expression~e	15	.5333333	.5163978	0	1
bstcipher	15	.2666667	.4577377	0	1
bstandvalt~e	15	0	0	0	0
imagebucke~l	15	.0666667	.2581989	0	1
topologica~t	15	.4	.5070926	0	1

In Figure 2, we review the statistics for GitHub repositories titled 'Elements of Software Design.' This table includes all eight assignments for the semester, each with 15 observations.

'Expression Tree' exhibited the highest mean, at approximately 0.5333, indicating its presence in

approximately half of the repositories associated with 'Elements of Software Design.' The overall lower mean values, compared to those in Table 1, suggest there was more null data present when randomly selecting repositories named 'Elements of Software Design'.

Figure 3 Combined Data Histogram



In Figure 3, the data have been aggregated to calculate the total sum of occurrences. This sum is represented in a histogram that displays the distribution of occurrences across the dataset. The histogram reveals that 'ExpressionTree', 'NumberSpiral', 'TopologicalSort', and 'WorkingHard' are the most frequent, appearing in 13 to 17 instances out of all 30 observations. 'EmployeeSalaries', 'BSTCipher', and 'Josephus' fall in the middle range, with 6 to 10 instances out of 30. 'BST and AVL Tree', and 'Image Bucket Fill' are the least common, with only 0 to 4 instances out of 30.

Figure 4 Combined Frequency Table

Frequency per Search	Freq.	Percent	Cum.
0	9	30.00	30.00
1	5	16.67	46.67
2	1	3.33	50.00
4	4	13.33	63.33
5	3	10.00	73.33
6	6	20.00	93.33
7	2	6.67	100.00
Total	30	100.00	

In Figure 4, the data has been aggregated to determine the frequency of observations. We have determined that out of all 30 repositories, 21 of them contained at least one file that matched the assignment from CS 313 E. This indicates a 70% occurrence rate of repositories with at least one matching file.

Figure 5 Total Amount of Repositories

repopromp	totalr~s
TeymourianCS313E	0
RamseyCS313E	0
MitraCS313E	0
Elements of Software Design	57
CS313E	119

The search for repositories using the terms 'MitraCS313E', 'TeymourianCS313E', and 'RamseyCS313E' returned no results. This lack of findings was consistent even after modifying the search parameters to include variations of the instructors' names with the class identifier

'CS313E'. No repositories were associated with the individual names of the instructors along with the course code.

The search term 'Elements of Software Design' yielded 57 repositories. The broad search term 'CS313E' resulted in 119 repositories, which likely encompasses a wide range of course-related content, including assignments, notes, and examples. It is likely that both of these search terms overlap due to their relation to each other. Both of these search terms were also filtered by python.

Notably, a single repository associated with Professor Mitra was identified. It has been uniquely categorized under 'CS313E'.

This data was collected on excel as well as the figure 3; the figures 1, 2, 4, 5 were calculated using: [StataBE](#). All data can be found from: [GitHub - Repositories](#).

Conclusion:

The main goal of this brief paper was to explore the potential misuse of GitHub for academic dishonesty in CS313E assignments. It also aims to provide foresight into which assignments are more readily available on GitHub. These insights could either help us prepare for upcoming assignments that may have easy access or modify existing assignments to safeguard academic integrity. The research adopted a Boolean approach, which may have inadvertently excluded assignments that were similar but not identical to those in the course, indicating an avenue for further refinement and depth in future research. Due to time constraints, the scope of this study was kept broad, reflecting the typical search behavior of a student seeking undue advantages through GitHub. It's hoped that the findings presented will offer valuable perspectives for future academic planning and integrity measures.