A Sneak Peek at a Collaborative CS Project: What part do female students play?

Kunal Kapoor (200245869)

Dr. Tiffany Barnes (Ruth Akintunde)

Spring 2021

Abstract

We present an exploratory analysis of how different genders collaborate while working on programming projects in a Java Programming Course of a south-eastern research-intensive US university. In the projects we studied, when students collaborated on Github, they declared what aspects of the project they worked on by including a tag in each commit message. Some aspects of the projects involved intensive coding whereas others did not. Our study analyzed these tags to determine what parts of the project each gender is prone to working on. We also performed predictive analysis to determine if elements of commit messages and tags are significant predictors of gender and vice versa. Our findings from the class we studied show that gender is not a factor when it comes to what each student worked on. This could serve as a motivation for female high school students who are considering choosing a career in Computer Science.

Problem Statement

The purpose of our research was to investigate the role of gender in programming task delegation among college teams. In specific, our hypothesis centered around the idea that women do more testing/documentation and less implementation/programming when working on teams

with males. Furthermore, we explored how we could measure this and how we could prevent this inequity or encourage teams to divide tasks in a more equitable way.

Short Literature Summary

When exploring related works, we found that female students had less confidence in their ability to learn computer science (CS) concepts than male students in CS. In fact, they felt less confident than the male students in non-computer science majors. The study also found that female students who have supportive friends and family members are more likely to major in computer science [1].

We also found that women tend to have their pull requests accepted at a higher rate than men. In fact, women's pull requests are larger, indicating more lines edited. However, when a female's gender can be identified, they are rejected more often [2]. Seeing that a bias exists in how gender efforts are perceived when gender is known, we wanted to know if this bias in turn affects females' choice of programming tasks while working in teams consisting of both genders in a blended learning environment.

Methodology/Process

We collected the dataset containing commit messages directly from Github, which consisted of two files, Gender file and Commit Messages file. The Gender file provided by the course instructor contains mappings of author_IDs to students' gender. Missing values in the gender column were labeled "Unspecified" (U.S.) The Commit Messages file contained each commit made by a student during that semester and details such as the number of lines added, the datetime of the commit, etc. Using the author_ID, we merged the two files into one dataframe.

We dropped the columns typeChanged and commitID since they weren't relevant to our research question and sorted the dataframe by repo_name to make it easier to narrow our research on teams later on. We then classified the tags/messages that students used into 9 unique commit types (untagged, versioning, documentation, style, implementation, debugging, testing, skeleton and misc.). Using these commit types and various other features, we created various decision tree models to attempt to classify the gender of a contributor within a mixed team. Some of these models (KNN, Decision Tree), were quite accurate and achieved a balanced accuracy of ~80%. However, the most important features in these models had more to do with the # of lines added/modified/removed than the type of commit made.

Results/Contributions

I worked primarily with the pre-processing of the data, creating many of the visualizations present within the final paper and developing the ML models used to classify gender (apart from the Random Forest). I also compiled various other exploratory statistics such as the number of commits made by males/females in mixed teams, the day of the week with the highest average commit count, and so on.

The final, revised models that Ruth and I developed were able to classify gender within a mixed team environment based on commit details with ~80% accuracy. This shows that there are differences within commits between males and females. However, when we analyzed these differences, we found that they were not statistically significant. Therefore, we concluded that within the multi-gender teams we studied, we did not find any team with negative teamwork habits such as cooperation where females solely perform less programming intensive tasks.

Reflection

The scope of the project didn't significantly change from the proposed project as we were able to fill in extra time and resources with data exploration and analysis. In specific, I was able to explore popular datetimes of commits (Fig. 2), the number of lines added/modified/etc. by males and females (Table 1) and more. Furthermore, as I worked on the models, I began to focus more on these "extra" details such as datetime and lines added/modified/.. since they proved to be more significant within the models as compared to the feature of commit type.

An interesting discussion our team had dealt with the reasons behind why the computing industry is currently male dominated. Ultimately, we hypothesized that females tend to believe that they need a higher level of competence as compared to males in order to enter computing. Furthermore, a section within our related works touches on this point especially well. "The authors then concluded that they speculate the reassurance of one's competence by others may increase the probability that a woman musters the courage to enter into a male dominated major such as CS [1]. A study by Du et al. showed that males were exposed to programming at an earlier age than females. They also found that after a one hour coding tutorial, males were more likely to take a programming course than females even though females performed better than or equal to males in the programming quizzes after the hour of coding [3]."

We believe that showing the statistics of females contributing more to computing than may be commonly known (see related works), or discussing papers such as ours within a high school classroom setting will help to inspire confidence for females to pursue computing careers. Furthermore, it would be helpful to have current female computing professionals present the topic of computing to female students as compared to male computing professionals, since the field is currently male dominated.

I sincerely appreciated the opportunity to partake in academic research, especially in a topic that deals closely with my recent experience (CS undergraduate teamwork habits). I enjoyed learning the process of academic research, from devising a problem statement and exploring related works to writing a final paper worthy of conference submission. Lastly, I welcomed the opportunity to explore data science and machine learning models in a research/academic setting and I look forward to continuing research in a related field in the fall semester.

References

[1] Sylvia Beyer, Kristina Rynes, Julie Perrault, Kelly Hay, and Susan Haller. 2003. Gender Differences in Computer Science Students. 35, 1 (2003). https://doi.org/10.1145/792548.611930

[2] N. Imtiaz, J. Middleton, J. Chakraborty, N. Robson, G. Bai, and E. Murphy-Hill. 2019. Investigating the Effects of Gender Bias on GitHub. In 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE). 700–711. https://doi.org/10.1109/ICSE.2019.00079

[3] Jie Du and H. Wimmer. 2019. Hour of Code: A Study of Gender Differences in Computing. Information Systems Education Journal 17 (2019), 91–100