Prospectus

The Design of an Undergraduate Course on Open-Source Software Development

(Technical Topic)

The Root Source of Bias in Amazon's Abandoned Resume Scanner Tool

(STS Topic)

By

Emily Franklin

11/23/2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed: 

Approved: _____ Date_____

Benjamin Laugelli, Department of Engineering and Society

Approved: _____ Date _____

Daniel Graham, Department of Computer Science

Introduction

The University of Virginia undergraduate computer science curriculum "...emphasizes basic science, technical mastery, research opportunities and a firm grasp of scientific principles as well as strong communication skills and creative problem solving" ("CS Undergraduate Programs", 2021). Each of these tenets has been highlighted during my time at UVa, successfully preparing undergraduates for a career in the technology industry. However, the coursework could be further strengthened by an emphasis on emulating real-world scenarios. In the industry, communication is more than just verbal, but is demonstrated through meticulous code documentation. As such, students should gain experience describing all of the code that they write as well as be able to work on others' programs using only their provided documentation. One application of this idea would be the institution of a course in which students contribute to open-source software (OSS) projects.

However, this technical fix is incomplete. For any undergraduate curriculum in computer science, it is essential that all students, regardless of background, graduate feeling adequately prepared to take on the challenges they will encounter in the technology industry. However, variations in students' backgrounds and identities may lead to differences in learning outcomes and therefore feelings of preparedness. I argue that gender identity, specifically, impacts individuals' experiences both in education and in the workforce, and this must be taken into account when designing curricula. In a 2017 survey by GitHub, only three percent of the 5,500 randomly selected open-source developers identified as women ("Open Source Survey", 2017). To put this in perspective, at that time, 22.6% of the computer programmer workforce was female (Finley, 2017). This signifies that women may feel less welcome in the open-source

community and raises the question of whether women will inherently have the same experience as men in the proposed OSS development course.

Through the design of my technical project, I will be building a heterogenous network between students of varying gender identities, open-source codebases, professional developers, and a multitude of other human and non-human actors. Therefore, looking at the failure of a related network would be constructive. I will draw on the network composed during Amazon's development of an automated resume scanner tool that was ultimately abandoned after demonstrating bias against women. If I neglect to consider both the technical and social aspects of the OSS development course design, I risk the chance of women feeling unwelcome or unsupported in a course that interacts with a community that is predominantly male-led, which could cause them to benefit less than their male counterparts in the course. Additionally, the open-source community stands to miss out on diverse perspectives and contributions; in a 2018 survey by the Linux Foundation, 72% of companies report "...frequently using open source for non-commercial or internal reasons and 55% using open source for commercial products" (Olin, 2018). With the usage of open-source software on the rise, different groups must be represented in the development process.

Establishing an equitable undergraduate course experience is socio-technical in nature and, as such, requires proposals that address both technical and social aspects. My technical project will deliver a design for a new course in open-source software development and my STS project will provide a better perception of the consequences of women being underrepresented in the technology industry. An understanding of these aspects together will allow me to mitigate the sources of these negative effects in my course design and ensure a more equitable experience for students.

Technical Project

In the summer of 2021, I acted as a project management intern at the technology company Yext, which offers brand management solutions to businesses. Implementations like location-specific web pages, AI-powered site search engines, and online directory listings for clients were being launched constantly, so a key issue for the company was maintaining an inventory of previous projects in order to facilitate the development process moving forward. Therefore, two other interns and I undertook the task of reconstructing Yext's Consulting Inventory, a searchable database of completed implementations that previously had been out-of-date and rarely used.

Throughout the Consulting Inventory overhaul, I learned the importance of communication between teams to clearly define the requirements and to optimize the website for different users' needs. The class *CS 4750 - Database Systems* helped prepare me the most for the technical side of the project, which was designing how to store the data in a way that minimizes redundancy and is understandable logically. Most importantly, this task taught me how crucial documentation is both for current users and future developers. My team picked up the project where another developer had left off; unfortunately, not a lot of documentation had been established, so a lot of time was spent inefficiently determining how the website worked and how it could be programmatically updated. As a result, my team was careful to leave extensive documentation, including logs of all changes, ideas for future improvements, and company contacts for questions about different aspects of the Inventory.

The UVa Computer Science curriculum neglects to emphasize the importance of documentation, resulting from students infrequently working on existing codebases or projects.

Frequently, students write their programs from scratch and typically do not have to decipher others' code. As such, students are receiving inadequate practice for what software engineering and similar work will look like in the industry; individuals will often work on existing projects rather than establishing their own. Furthermore, even if the engineer is creating their own project, they must consider how others will be able to continue their work in the future. The computer science courses in their current form do not fully prepare students for an experiential learning event like mine. To remedy this, I will propose a course in which students work on open-source software (OSS) development projects. OSS is a category of software in which the code can be freely used for any purpose and is most often developed publicly and collaboratively. This proposed course would allow individuals to practice working on existing codebases, during which they must read logs and other documentation to gain context as well as actively converse with other developers. By implementing such curriculum changes, students will not only gain valuable skills in communication but develop a sense of community with other programmers, which is likely to facilitate their transition into the industry.

To organize this proposition, I will draw on the Association for Computing Machinery's 2020 Computing Curricula Report, which provides guidelines and recommendations for the structure of undergraduate computer science programs ("Curricula Recommendations", 2020). I will also use a successfully implemented OSS development course at Oregon State University as guidance (Budd, 2008). Additionally, I will use a journal article from the Communications of the ACM that outlines the most important learning outcomes for OSS student developers (Spinellis, 2021).

STS Project

In 2014, Amazon began work on a software tool to automate resume scanning in order to streamline its hiring process. The tool used a machine learning algorithm that applied Amazon's previous hiring patterns in determining the potential fit of new candidates. By 2015, the project was abandoned due to the algorithm discriminating against women, scoring resumes with terms like "women's rugby team" lower than male counterparts' (Goodman, 2018). This failure was attributed to bias in the training data of the algorithm; the algorithm had been trained using resumes submitted to Amazon over a 10-year period, during which a greater number of men than women were hired by the company. Therefore, the model recognized this pattern and began favoring male applicants over female applicants (Dastin, 2018). While this view on the project's outcome is true, it overlooks the root cause of the failure: women have always been underrepresented in the technology industry, and this issue must be addressed before any company can hope to create an unbiased hiring algorithm. If we continue to simply blame a skewed dataset when algorithms show bias, we neglect to understand why and how the datasets become biased in the first place. This reflection becomes crucial as other companies increasingly push to use machine learning in their recruitment processes (Bogen, 2019). I argue that the underrepresentation of women at Amazon over time in conjunction with the company's propensity to remove human actors where possible caused this project to fail. This became problematic to the network by introducing coded bias and attempting to amplify existing hiring patterns.

To support this argument, I will make use of Actor-Network theory, which frames the development of technology through the relationships between human and non-human actors that comprise a network (Cressman, 2009). Using Actor-Network theory to map out the network

composed during the development of Amazon's hiring algorithm will enable a deeper look into the history of interactions between groups, illuminating the factors that may have contributed to the project's failure. A main tenet of Actor-Network theory is that no single actor is more powerful or important than any other. I argue that undervaluing social factors, like gender diversity, introduced instability into the network over time. To reinforce this claim, I will analyze several gender bias lawsuits filed against Amazon by women and research firsthand accounts of their and others' experiences (Stempel, 2021). This will allow me to gain an understanding of possible barriers in career growth for women at Amazon as well as reasons for exiting. I will also examine the company's current and past demographic data and diversity initiatives ("Our Workforce Data", 2021).

Conclusion

The deliverable for my technical project will be a design for a new course in open-source software development with a detailed description of its intended learning outcomes. My STS project will be an analysis of the factors that led to the failure of Amazon's automated resume scanner; it will look especially at how the underrepresentation of women at Amazon caused the training dataset to become skewed. This analysis will be framed using Actor-Network theory to illustrate how a lack of emphasis on social factors in the network introduced instability.

The combined results of these projects will allow me to address the broader issue of providing an equitable learning environment in which students are enabled to succeed regardless of gender. As a result, a greater focus will be placed on addressing the factors that cause women to feel less welcome in open-source communities. The benefits of both the technical skills gained and the feeling of belonging in the computer science community will promote a successful

transition into the industry, which is one of the central goals of the University of Virginia computer science curriculum.

Word Count: 1,736

References

Amazon. (2021). *Our workforce data*. About Amazon. Retrieved October 24, 2021, from https://www.aboutamazon.com/news/workplace/our-workforce-data.

Bogen, M. (2019, May 6). *All the ways hiring algorithms can introduce bias*. Harvard Business Review. Retrieved October 23, 2021, from https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias.

Budd, T. (2008). *A course in open source development*. Retrieved October 22, 2021, from http://www.cs.trincoll.edu/~ram/hfoss/Budd-FOSS-Course.pdf.

Cressman, D. (2009). *A brief overview of actor-network theory: punctualization, heterogeneous engineering & translation*. Summit. Retrieved October 23, 2021, from https://summit.sfu.ca/item/13593.

Curricula recommendations. Association for Computing Machinery. (2020). Retrieved October 22, 2021, from https://www.acm.org/education/curricula-recommendations.

Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. Retrieved October 22, 2021, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Finley, K. (2017, June 2). *Diversity in open source is even worse than in tech overall*. Wired. Retrieved October 23, 2021, from https://www.wired.com/2017/06/diversity-open-source-even-worse-tech-overall/.

GitHub. (2017). *Open source survey*. Retrieved October 23, 2021, from https://opensourcesurvey.org/2017/.

Goodman, R. (2018, October 12). *Why Amazon's automated hiring tool discriminated against women*. American Civil Liberties Union. Retrieved October 22, 2021, from https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against.

Olin, E. (2020, December 20). *Corporate open source programs are on the rise as shared software development becomes mainstream for businesses*. Linux Foundation. Retrieved October 23, 2021, from https://www.linuxfoundation.org/press-release/corporate-open-source-programs-are-on-the-rise-as-shared-software-development-becomes-mainstream-for-businesses/.

Spinellis, D. (2021). Why computing students should contribute to open source software projects. *Communications of the ACM*, *64*(7), 36–38. https://doi.org/10.1145/3437254

Stempel, J. (2021, May 19). *Amazon.com faces five new racial, gender bias lawsuits*. Reuters. Retrieved October 24, 2021, from https://www.reuters.com/technology/amazoncom-faces-five-new-racial-gender-bias-lawsuits-2021-05-19/.

University of Virginia. (2021, June 2). *CS undergraduate programs*. University of Virginia School of Engineering and Applied Science. Retrieved October 23, 2021, from https://engineering.virginia.edu/departments/computer-science/cs-undergraduate-programs#accordion396960.