THE MR SCRAPER AT GENERAL ATOMICS COMMONWEALTH COMPUTER REASEARCH, INC.

AN ANALYSIS OF TIKTOK'S CONTENT GENERATION ALGORITHM THOUGH STAR'S ETHNOGRAPHY OF INFRASTRUCTURE

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Kevin Cooper

November 1, 2022

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.

ADVISORS

Rider Foley, Department of Engineering and Society

Briana Morrison, Department of Computer Science

INTRODUCTION

Lasting success is driven by productivity. For large website-focused companies, this requires driven employees who are talented collaborators. It is tedious to have to send documents back and forth whenever a teammate makes a change. Programs like Google Docs and Microsoft Office Online allow multiple users to type on the same document simultaneously. This is helpful until someone writes directly on top of someone else's work or deletes everything off the page. To solve that collaboration issue Git, and one of its interfaces GitLab, allows users to work on the files on their own and then integrate everything at the end.

Git is a local version control system that stores information about a filesystem by collecting snapshots of the files in a repository. Git remembers the current image with a reference to the repository whenever the project is saved or committed (Git-scm, 2022). Different versions of the code, or branches, can be edited and have their own commits. Once a developer is content with their branch, they can merge it back to the main branch. GitLab is an online interface built on Git (Pipinellis, Read, Sedlak-Jakubowski, Qualls, & Selhorn, 2022). It stores a version of the code online so that anyone who has access can create a clone of the repository on their own machine. Developers write code and can push what they wrote, or pull new information from a colleague from the combined repository. Another benefit of using GitLab is that it is a built-in integration tool (Shipton, 2019). The integration tool allows for a company's website to be tested and built straight from GitLab.

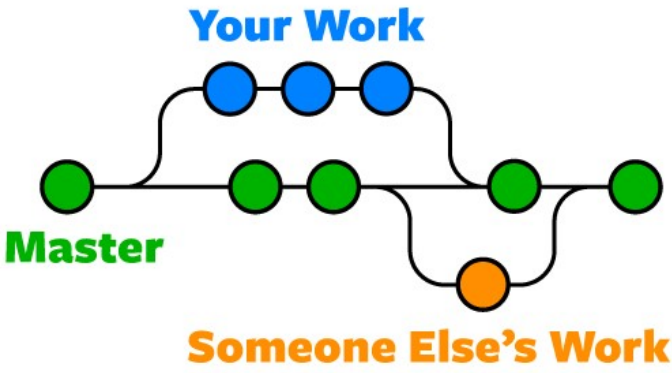

Figure 1. A sample Git repository with branches which are merged together. Copyright 2018 by F. Igor.

General Atomics Commonwealth Computer Research, Inc. (GA-CCRi) use GitLab because of its code version and integration control. GA-CCRi has an agile or fast-paced work environment that needs the continuous integration that GitLab provides (Arefeen & Schiller, 2019). GA-CCRi is an industry leader in geospatial storage, visualization, and analysis (Ralston, 2022). They are a defense contractor that specializes in real-time situational awareness of aerial and nautical vehicles around the globe. They collect data from a variety of sources, including vehicular and weather sensors, with predictive analysis and data science techniques to achieve precise locations for all tracked vehicles with a low time delay. Their clients are various government agencies around the world and commercial companies who have a use for global data or the company's global visualization program. [transition to next section.]

MR SCRAPER

I spent my time at GA-CCRi as a back-end developer gathering data from the company's GitLab application programming interface (API) and displaying it in a digestible manner. An API is a type of software that communicates with the specified website and gives document reports of the information on the page (IBM, 2022). The information is read from those documents with a scraper program. The scraper collects Merge Request (MR) and project data which is then displayed on a Grafana dashboard. Grafana is a web application that provides charts, graphs, and text for viewers to understand the data that has been collected (Grafana, 2022). This helps the company immensely because it allows for easy comprehension of the productivity at GA-CCRi. I will be looking at this through Star's infrastructure lens to understand how my scraper was embedded in the company and how it could lead to an increase in productivity as a whole.

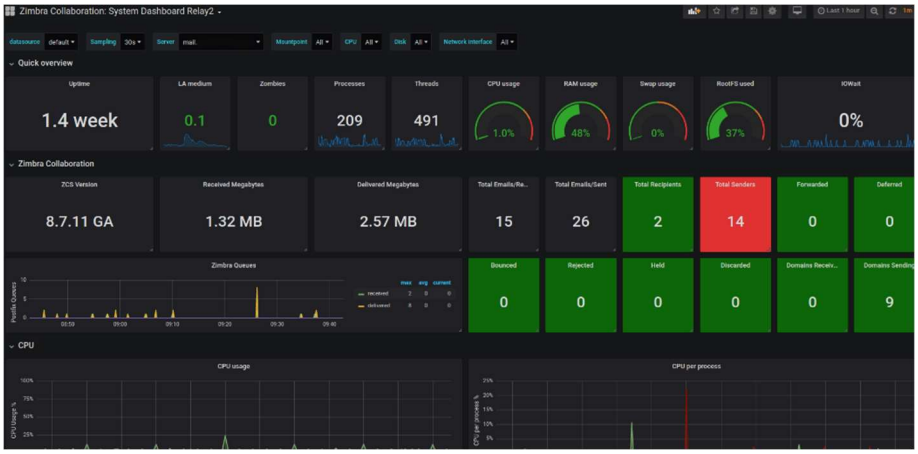

Figure 2. A sample Grafana dashboard. Copyright 2019 by Capterra.

The merge request scraper that I worked on this summer was colloquially known as the MR Scraper even though it took in data for more than just MRs. The company had an imprecise and incomplete implementation of the data scraper before I started on it. I worked to fix most of the found bugs in the code while also adding features to the scraper. I contributed by adding

more information tables like a table for Git commit, changes information, and project statistics. The code was written in Scala and SQL. Scala was used for the API scraping and for the functions that allow it to be run repeatedly. SQL is how the data was gathered and imputed into the database (Koleoso & O'Reilly Online Learning: Academic/Public Library Edition, 2021). Grafana was connected separately and I also made new tables and added additional information to the previous ones. Most notably the ones for the average time that people were working on an MR, and how many were self-merged.

Merge requests are one of the last lines of defense for buggy code to be deployed out to the complete project. Developers draft them in such a way so their colleagues can decipher what they changed on their branch and why it should be changed for everyone. This can be done in a few ways but comments are one of the most helpful because it doesn't clutter the code itself with comments, but can still have a plaintext way to explain someone's reasoning (Fayock, 2019). One of my additions was to add a comment count and display who was writing efficient comments on which MRs.

There is a process that code has to go through before it can be merged to the main repository. It is first written and tested, but it needs to go through the review process before being merged (Cooper, 2019). Directly after an MR is initiated and after each commit to that branch thereafter GitLab will run automated tests to see if the code would compile and build with the rest of the program. This is part of the Continuous Integration and Continuous Development pipeline (Currie, Mataev, & Clemencic, 2020). This process is extremely useful for agile work environments (Donca, Stan, Misaros, Gota, & Miclea, 2022). Developers collaborate with the infrastructure to guarantee their applications can be deployed quickly and reliably. This was

something I learned from CS 3240 Advanced Software Development and was reinforced with the internship.

That class prepared me well for the workforce because we did work with three-week sprints at GA-CCRi, which is similar to how the class project was run. My UVA course load has prepared me for this internship and my future career by teaching me many different computer languages. Outside of my computer science degree, my data science minor was a large help for this internship. My previous experience with data cleaning and representation allowed me to spend time working on what type of information I wanted to display instead of the system process. GA-CCRi previously rarely presented the gathered information from the scraper because there wasn't much information to display. Now they can use the scraper to get the necessary information to make informed decisions about how to be more productive as a company. They can continue to make such decisions in the future because the code I wrote was embedded into GA-CCRi's infrastructure.

THE MR SCRAPER IS INFRASTRUCTURE

Star defined infrastructure with nine main properties. They are embeddedness, transparency, reach or scope, learned as part of membership, links with conventions of practice, embodiment of standards, built on an installed base, becomes visible upon breakdown, and is fixed in modular increments, not all at once or globally (Star, 1999). An infrastructure is transparent when it can be implemented once and then used repetitively without needing to be remade. Reach or scope means that the technology can touch pieces all throughout the physical system or through time and work in the future. Something becomes visible upon breakdown when the usually hidden parts of the system are displayed and no longer run smoothly.

Technology is fixed in modular increments by not revamping the system all at once. The system is large enough that it has many parts that individually may have a problem, but once fixed locally then the system will be restored. Learned as part of membership is important for an infrastructure for when someone new enters a community. They are presented with the new equipment and are naturally introduced to its purpose.

The MR Scraper has transparency because it automatically runs and updates the new information every hour. It scrapes with a reach and because it scrapes all MRs and project information for all the current GA-CCRi repositories and repositories of the past. When an API does not return all the information, it is visibly broken. The Grafana tables will be missing data points which gives an incomplete dataset (Bounegru & Gray, 2021). Usually, this happens when an MR is immediately merged or deleted without any changes. The scraper is fixed in modular increments because I only worked on one problem at a time. There was a scraper implemented before my time there, but I worked on bugs and new functions individually to keep the structure in place and only modify features. That allowed the system to stay in place during development. It is learned as part of membership because at the All Hands, full company meeting (Simhon, 2022), the Grafana graphs will be shown to inform the company about their own work and productivity while enlightening those who are working well.

Presenting supervisors with the scraper's information can help them know who is doing what and let them know how to redistribute tasks if necessary to balance everyone's workload. Showing the information to employees can highlight people who are doing well and incentivize others to join and use their time wisely at work (Patnaik, 2022). Productivity trackers are beneficial for transparency on how the company works however there is a limit to how much companies should track their employees. Various employers are requiring employees to use

productivity monitoring applications and wearable technology to monitor them outside of work (Ajunwa, 2018). This has led to legal cases, like Quon v. Arch Wireless, but the courts have sided with the companies to allow trackers.

TIKTOK CONTENT GENERATION ALGORITHM

The research question that I will be answering next semester is: How does the content selection algorithm from TikTok impact young adults living in the US as they enter society? Young adults are people who have completed their schooling and who are becoming independent. I am interested in this topic because I am one of these young adults who use the smartphone application TikTok frequently, and I'm interested in the machine learning techniques which could impact my peers and me. I will analyze the content generation algorithm through Star's infrastructure framework (Star, 1999). The elements that I will look at in particular are embeddedness, reach and scope, learned as part of membership, built on an installed base, and becomes visible upon breakdown. These are sensible properties because of how available TikTok can be to young adults who use their smart devices daily.

To research this question, I will be conducting interviews with TikTok personnel as well as gathering data from surveys of users. I will contact data science software developers at TikTok and ask questions like "How does the algorithm connect users?". I will be taking surveys from users to get their perspectives. These surveys will be a google form sent to many of my colleagues who can give their own reports on their TikTok activity. My goal for the information gathered from these are about their TikTok activity with questions like how often they frequent the app, approximately how many swipes they make, and how often they interact with the content. I will also ask about their content like if they enjoy the content they're viewing, if they

have learned anything off of the app, if they are content creators, if they understand the data that is being collected, and if they feel a sense of community inside of TikTok.

CONCLUSION

There are six steps when data scraping. They are to identify a source, determine how that data is represented, write a scraper to harvest the data, traverse all the webpages that contain the wanted data, run test cases, execute the algorithm to scrape the data, and build the final dataset (Landers, 2017). My internship taught me this process. It also taught me how important it is to reflect on past projects and MRs for the betterment of the company. The work culture will benefit from viewing the MR Scraper's results.

The expected results of my research paper are that TikTok users have a sense of community for using the app together. I also believe that people have gained household skills just by viewing content. It is important for social media platforms to create a sense of community in order to keep users engaged (Alien, 2011). TikTok's content selection machine learning algorithm should keep users connected in the same way.

REFERENCES

Ajunwa, I. (2018). Algorithms at Work: Productivity Monitoring Applications and Wearable Technology as the New Data-Centric Research Agenda for Employment and Labor Law. *Saint Louis University Law Journal*, *63*(1), 21–54.

Alien, S. E. O. (2011, December 27). Promoting a Sense of Community in Social Media | SEO-Alien. Retrieved October 27, 2022, from https://www.seo-alien.com/tips-and-tricks/promoting-sense-community-social-media/

Arefeen, M. S., & Schiller, M. (2019). Continuous Integration Using Gitlab. *Undergraduate Research in Natural and Clinical Science and Technology Journal*, *3*, 1–6. https://doi.org/10.26685/urncst.152

Bounegru, L. (Eds.), & Gray, J. (Eds,). (2021). *The Data Journalism Handbook: Towards a Critical Data Practice*. Amsterdam: Amsterdam University Press. (Internet materials). Retrieved from http://proxy01.its.virginia.edu/login?url=https://www.degruyter.com/openurl?genre=book&isbn=9789048542079

Cooper, R. (2019, May 23). The Perfect Code Review Process. Retrieved October 26, 2022, from Osedea website: https://medium.com/osedea/the-perfect-code-review-process-845e6ba5c31

Currie, R., Mataev, R., & Clemencic, M. (2020). Evolution of the LHCb Continuous Integration system. *EPJ Web of Conferences*, *245*, 05039. https://doi.org/10.1051/epjconf/202024505039

Donca, I.-C., Stan, O. P., Misaros, M., Gota, D., & Miclea, L. (2022). Method for Continuous Integration and Deployment Using a Pipeline Generator for Agile Software Projects.

Sensors, *22*(12), 4637. https://doi.org/10.3390/s22124637

Fayock, C. (2019, May 2). Why you should write merge requests like you're posting to Instagram. Retrieved October 26, 2022, from FreeCodeCamp.org website: https://www.freecodecamp.org/news/why-you-should-write-merge-requests-like-youre-posting-to-instagram-765e32a3ec9c/

Git-scm. (2022). Git—What is Git? Retrieved October 26, 2022, from https://www.git-scm.com/book/en/v2/Getting-Started-What-is-Git%3F

Grafana. (2022). Grafana: The open observability platform. Retrieved October 26, 2022, from Grafana Labs website: https://grafana.com/

IBM, C. E. (2022, September 28). What is an Application Programming Interface (API)? Retrieved October 26, 2022, from https://www.ibm.com/cloud/learn/api

Igor, F. (2018). Branches.png (550×477). Retrieved December 8, 2022, from Git for Beginners website: https://developerhowto.com/2018/10/12/git-for-beginners/

Koleoso, T. & O'Reilly Online Learning: Academic/Public Library Edition. (2021). *Beginning jOOQ: Learn to Write Efficient and Effective Java-Based SQL Database Operations*. S.l.: Apress. (Internet materials). Retrieved from http://RE5QY4SB7X.search.serialssolutions.com/?V=1.0&L=RE5QY4SB7X&S=JCs&C=TC_046418529&T=marc

Landers, R. N. (2017). Crash Course in I-O Technology: A Crash Course in Web Scraping and APIs. *TIP: The Industrial-Organizational Psychologist*, *55*(2), 5–11.

Pipinellis, A., Read, E., Sedlak-Jakubowski, M., Qualls, A., & Selhorn, S. (2022). Git on the command line | GitLab. Retrieved October 26, 2022, from https://docs.gitlab.com/ee/gitlab-basics/start-using-git.html

Ralston, B. (2022). About Us. Retrieved October 26, 2022, from GA-CCRi website: https://www.ga-ccri.com/about

Screenshot Grafana.png (1808×882). (2019). Retrieved December 9, 2022, from Capterra website: https://wiki.zimbra.com/images/6/6c/Screenshot-grafana-2019.07.07-09-43-46.png

Shipton, L. (2019, October 14). GitHub vs GitLab: Which Platform should I choose? Retrieved October 26, 2022, from Venture Lessons website: https://www.venturelessons.com/github-vs-gitlab/

Simhon, S. (2022). What is an all-hands company meeting? Retrieved October 27, 2022, from Connecteam website: https://connecteam.com/hr-glossary/what-is-an-all-hands-company-meeting/

Star, S. L. (1999). The ethnography of infrastructure. *The American Behavioral Scientist*, *43*(3), 377–391. https://doi.org/10.1177/00027649921955326