The Secret Life of Hackathon Code
Where does it come from and where does it go?

Yash Joshi
2021 11027

Harsh Savaliya
2021 11052

Jahanvi Thakkar
2021 11058

Vaidehi Bhandarkar
2021 11064

Shrey Yagnik
2021 11072

Introduction

The word 'hackathon' is often correlated with coding challenges. However, it is way more than coding. A hackathon is a big innovative competition where engineering students pitch their ideas to solve real-life problem statements. Though hackathons are open to all, few are meant primarily for engineering students. Throughout the past ten years, hackathons have gained a lot of recognition. A hackathon brings more opportunities to you as many recruiters offer internships/full-time opportunities to fresh talented engineering students via these competitions. It is an awesome way of getting exposure to show off your talent and meet and network with the right people to get internships or full-time job offers in your desired company.
Undoubtedly, hackathons are fun but they are grueling too, as they will challenge you to the core. It is all about creating and thinking of undefeatable ideas and pitching to the panel of judges by explaining what you did, why you did, how you came up with this idea, and why it would be great to implement your idea, ultimately showing the reach and value of your solution. This project aimed to research how much of the hackathon code was made before, during, and after the event. Also, we tried to understand how much of the code they have reused from existing code and how much actual code they have developed during this event. This research helps to understand the role of hackathons and the code reuse phenomenon.

Background

The increasing popularity of hackathons leads us to gather more information about them. While studying hackathons, we read several papers. Currently, there is a lot of research going on, and most of the research focuses on the importance of hackathons for college students [1], and improving the skill and knowledge of students, and increasing motivation of students [2].
There is also some research based on code reusability during hackathons and what happens to the code after hackathons [3]. This research paper [3] includes several statistical data about how much code was created before, after, and during hackathons. It also includes the study of how much hackathon code impacts after the hackathon is over. The methodology used in this research paper [3] helps hackathon organizers to find how much code was

created during an event. It also can be a starting point to do more research in this perspective of the hackathon.
In this research paper, we tried to replicate the same process to get the statistics about how much code was created during, after, and before the hackathon. And we came up with nearly the same outcome. To replicate the same research paper [3] we also required a dataset and servers, Which are easily available on the internet. So data collection tasks became easier. We use the same dataset and methodology to do research, But due to some hardware limitations, we reduced our dataset.

Concept

Hackathons became popular in the last few years so the dataset is not organized in such a manner. It's difficult to get all the datasets from one source. We have collected data regarding hackathon id, Github repository URLs, location, hackathon end date, winner of hackathon, etc. from the DEVPOST. But this data alone is not sufficient to answer our question.
To collect the rest of the data regarding projects, author, and code blob collected from World of Code(Woc). WoC contains information about OSS projects, developers, commits, code blob, files name, and more. WoC also provides a different map between all this information which is very helpful to answer the research question.
All the collected data is not in the form we want and there are lots of unwanted data which is inconvenient. So we need to do a cleaning process to remove unwanted data and outliers. After the cleaning process, we are required to find different maps like the project to commit, commit to blob, etc. From this mapping, we can find the first commit of blob and much more information, and then from that information, we can do further analysis.

Implementation

1. *Data Collection and Cleaning*
   We have taken 1,43,489 projects from different hackathons organized on the DEVPOST platform listed on KAGGLE[4]. First, we took only projects which have been posted on github.com, after filtering these projects we ended up with only 54,393 projects which have valid GitHub URLs. As the project ID used in DEVPOST is different from the project names in WoC, so first we need project IDs in the form of GitHubUserName_RepoName. For doing this we created a simple script in the python language as described in fig. 1.
   Now we need to use the WoC server to generate the different mappings. After getting access to this server we establish a connection between our local machine and server using the SSH pipeline. And for transferring the data we used SCP to securely transfer these files as described in fig.2.

```
with open(input_file,newline='') as urls, open(output_file,'w',newline='') as
projects:
    reader = csv.reader(urls,delimiter = ' ',quotechar = '|')
    writer = csv.writer(projects, delimiter=' ',quotechar='|',
quoting=csv.QUOTE_MINIMAL)
    count = 0
    for row in reader:
        s_row = str(row[0])
```

```
        slen = len(s_row)
        s_row = s_row[:slen - 2]
        tokens = s_row.split('/')
        n=len(tokens);
        project_id =tokens[n-2]+"_"+tokens[n-1]
        row_list = [project_id]
        row_list.append(project_id)
        writer.writerow(row_list)
    print(reader.line_num)
```

figure 1. GitHubUserName_RepoName format generator

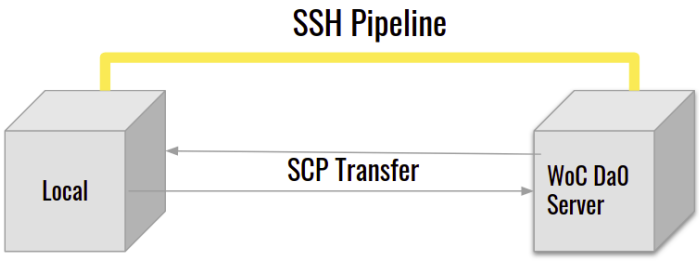

figure 2. SSH connection with WoC Da0 Server

For 54,393 projects, Now we need to get different mappings for our objectives. For 54,393 projects, all these mappings data we found were too huge for our local machines. So we chose relevant 101 projects manually and get these mappings and the count of these is shown in table 1.

Project to Commit maps (p2c)	(Projectname, CommitHashId)	2829
Commit to Blob maps (c2b)	(CommitHashId, BlobHashId)	2,02,719
Blob to Author maps (b2fa)	(BlobHashId,Timestamp,AuthorId,FirstCommitHashId)	1,06,282
Project to author maps (p2a)	(ProjectId, ProjectAuthorEmail)	101

Table 1. Mappings from WoC

For 101 projects we don't have the hackathon's end date associated with the projectID so from hackathon Id we find manually all dates from the DEVPOST.

2. *Analysis and Interpretation*

After finding these all mappings, as our aim in this project is the identification of the reuse of "code", we decided to filter only code blobs

from these all different blobs. So, we used the linguist tool from GITHUB to find out the file types as shown in fig. 3.

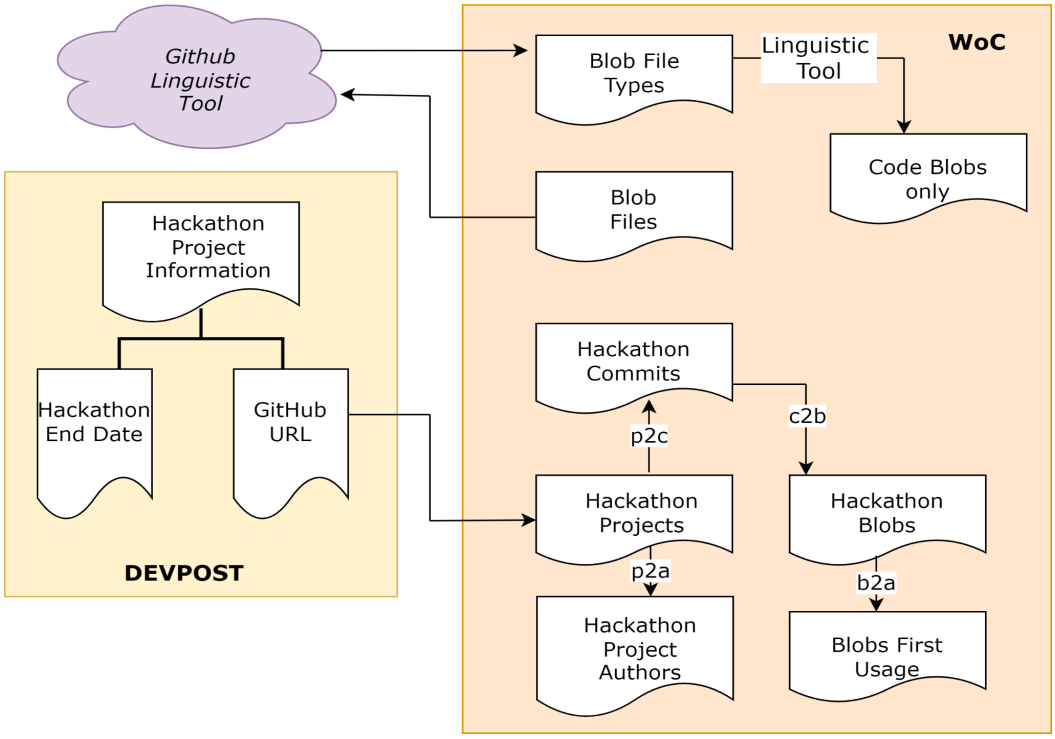

figure 3. Data Extraction Workflow[3]

We had the hackathon's end date so We have the end date for each of the hackathon events from DEVPOST, however, it does not include any information about the start date of the hackathon. We consider the start of a hackathon 72 hours before the end date. This assumption appears reasonable since hackathons are commonly hosted over a period of 48 which are often distributed over three days [5][6].
From all the mappings we perform inner joins on different mappings and we got the first timestamp when the blob was created by the author. By comparing this timestamp with the duration of the hackathon we found the results shown in fig. 4. and fig. 5. respectively.

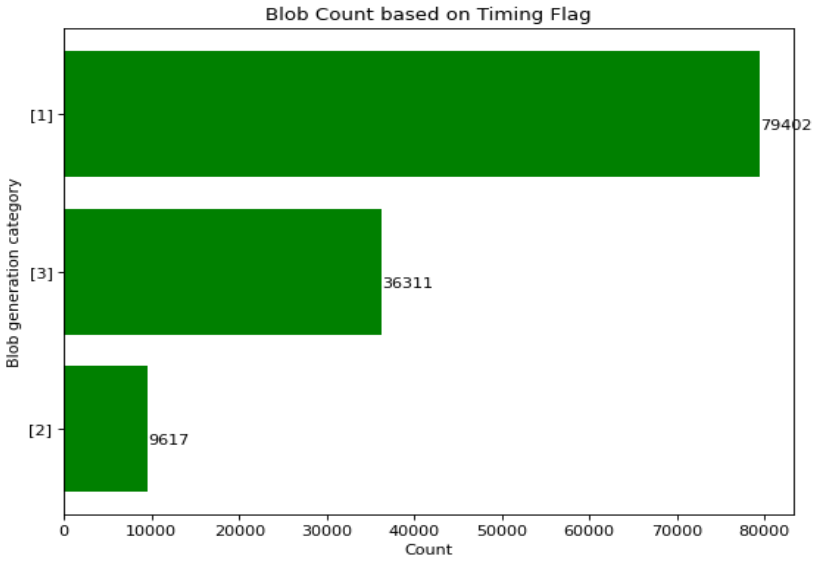

figure 4. Blob count based on the timing flag

In fig 4. And fig.5 timing flag represents '1' as a code generated before the hackathon, '2' as code generated during the hackathon, and '3' as code generated after the hackathon.

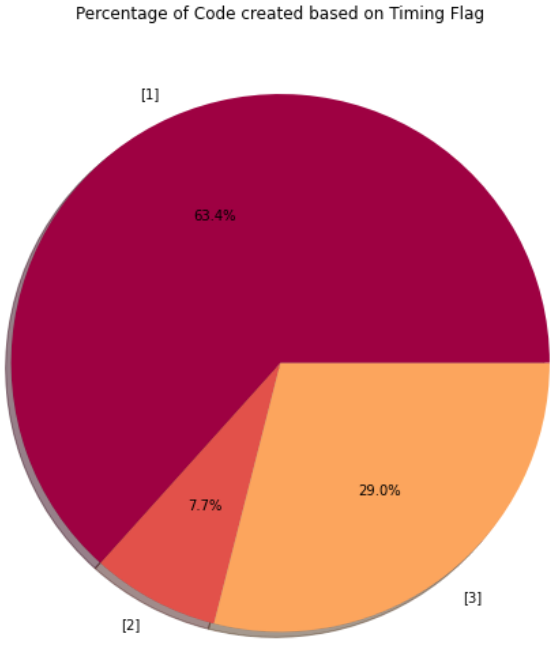

figure 5. Blob count percentage based on the timing flag

Conclusion & Future Work

In this study, we investigated the origins of hackathon code and its reuse Before, During, and After the event, we found that most hackathon projects reuse existing code, and the code created during events also gets reused by other Open Source projects later on. We collected Data from World of Code and DevPost and scraped the data for further analysis using Python Libraries like pandas, NumPy, matplotlib, etc. We tracked the code generation and usage on a blob level - represented in WORLD OF CODE. From that, it can be inferred that very few blobs are committed during Hackathons which is around 7.67%. This indicates that not much new code gets developed during a hackathon, and much of the code used by the teams is reused from existing code. This represents that our perception that hackathons are code-intensive creation events is a myth and most of the code is reused. As b2c(Blob to Commit) file takes more than 300 GB of space for 101 Projects. From this, it is evident that more space is required for numerous projects existing today. For this, if sufficient resources are available, one can find out What happens to Code after the event? When the final commit is performed. c2p (Commit to Project) file depends on b2c file data from which further study of How certain project characteristics can influence hackathon code reuse? can be done. From this, we can discover the popularity of a project as to how much the project implementation is useful even after the event ends.

References

[1] K. Gama, B. Alencar, F. Calegario, A. Neves, and P. Alessio, “A hackathon methodology for undergraduate course projects,” in 2018 IEEE Frontiers in Education Conference (FIE). IEEE, 2018, pp. 1–9.

[2] C. Guerrero, M. del Mar Leza, Y. González, and A. Jaume-i-Capó, "Analysis of the results of a hackathon in the context of service-learning involving students and professionals," 2016 International Symposium on Computers in Education (SIIE), 2016, pp. 1-6, DOI: 10.1109/SIIE.2016.7751857.

[3] Ahmed Imam, Tapajit Dey, Alexander Nolte, Audris Mockus, James D.Herbsleb, “The Secret Life of Hackathon Code Where does it come from and where does it go?”, arXiv:2103.01145v2 [cs.SE]

[4] Dataset Package https://www.kaggle.com/aahuang/devpost-project-data

[5] A. Nolte, E. P. P. Pe-Than, A.-A. O. Affia, C. Chaihirunkarn, A. Filippova, A. Kalyanasundaram, M. A. M. Angarita, E. H. Trainer, and J. D. Herbsleb, “How to organize a hackathon - a planning kit,” ArXiv, vol. abs/2008.08025, 2020.

[6] D. Cobham, K. Jacques, C. Gowan, J. Laurel, S. Ringham et al., “From appfest to entrepreneurs: using a hackathon event to seed a university student-led enterprise,” in 11th annual International Technology, Education and Development Conference, 2017.