Final Project Report

COP 701 - Assignment 3

Harish Chandra Thuwal, Shadab Zafar

November, 2017

Contents

List of Tables

List of Figures

1 Introduction

GitHub is a code hosting and collaboration platform for Git repositories (most of which are source code.)

GHTorrent is a mirroring service that monitors the GitHub event stream and stores the data. Which is then provided for download for research related activities.

In this paper we report various analyses that we performed on the GitHub data.

We used d3.js[1] to generate topological maps and bokeh[2] to create other charts and figures.

2 Dataset

2.1 Contents

The GHTorrent dataset[3] we used is a 60 GB compressed tarball containing 20 CSV files. Each file corresponds to a table of a MySQL database, the schema of which is available here. Table 1 lists the CSV files in the dataset.

	File	Size	Data
1	users.csv	1.6 GB	GitHub Users
2	organization_members.csv	16.1 MB	Users that are members of an organization
3	followers.csv	588.7 MB	Users that follow another user
4	watchers.csv	2.9 GB	Users that watch a project
5	projects.csv	11.1 GB	GitHub Projects
6	project_commits.csv	89.8 GB	Commits on projects (including commits on forks)
7	project_languages.csv	4.3 GB	Programming languages used in projects
8	project_members.csv	491.8 MB	Users that are contributors to projects
9	repo_labels.csv	6.6 GB	Labels used in a project
10	commits.csv	68.8 GB	Commits on projects
11	commit_comments.csv	731.6 MB	Comments made on commits
12	commit_parents.csv	13.7 GB	Parent(s) of commits
13	issues.csv	3.0 GB	GitHub Issues made on projects
14	issue_comments.csv	4.1 GB	Comments made on issues
15	issue_events.csv	4.9 GB	Actions taken on issues (closing etc.)
16	issue_labels.csv	262.6 MB	Labels assigned to issues
17	pull_requests.csv	1.2 GB	GitHub Pull-requests made on projects
18	pull_request_comments.csv	2.7 GB	Comments made on pull-requests
19	pull_request_commits.csv	2.3 GB	Commits made on pull-requests
20	pull_request_history.csv	3.5 GB	Actions taken on pull-requests (merging etc.)

Table 1: GHTorrent Dataset Contents
Colored rows indicate files that we performed analysis on.

[1] https://d3js.org/
[2] http://bokeh.pydata.org/
[3] http://ghtorrent-downloads.ewi.tudelft.nl/mysql/mysql-2017-09-01.tar.gz

2.2 Preprocessing

We used xsv[4] , a command line CSV parsing tool, to remove columns that contained information we were not interested in. This helped reduce the size of the files.

From `projects.csv`, we removed the following:

1. The description field which was free form text describing the project.
2. The name field which denoted the name of the project as the name of the project was evident from the url itself.
3. The update at field which denoted the last timestamp when the project was updated.

By removing the above mentioned columns we were able to reudce the size of the file by 40%.

From `project_members.csv`, we removed the ext_ref_id which used to be an old field in the Ghtorrent table but now it is useless.

From `commits.csv`, we removed the sha (160 bits) field which was the hash value associated with each commit. This lead to a significant reduction in the size of the file from 68.8 GB to 37.6 GB (nearly 55%) size of the table.

From `issues.csv`, we removed the pull_request field which denotes whether an issue is a pull request or not (In Github each pull request is also treated as an issue).
We removed this field because another field pull_request_id can be used to deduce whether an issue is a pull request or not. If the pull_request_id is `NULL` then the issue is not a pull request otherwise it is.

File	Fields Removed	Old Size	New Size
users.csv	-	1.6 GB	1.6 GB
followers.csv	-	588.7 MB	588.7 MB
watchers.csv	-	2.9 GB	2.9 GB
projects.csv	Description, URL	11.1 GB	4.4 GB
project_members.csv	ext_ref_id	491.8 MB	321.0 MB
project_languages.csv	-	4.3 GB	4.3 GB
commits.csv	SHA of Commits	68.8 GB	37.6 GB
issues.csv	Pull Request	3.0 GB	2.8 GB
pull_requests.csv	-	1.2 GB	1.2 GB
Total Size		94 GB	55 GB

Table 2: Size before and after preprocessing of files
Total New Size (55 GB) is the amount of data we processed on the cluster

[4] https://github.com/BurntSushi/xsv/

3 Analysis

3.1 Users

The `users.csv` file contains records of 19,925,838 (around 20 million) users.

3.1.1 Exponential Growth

After launching in 2007, GitHub has been growing exponentially, which is made evident from Figure 1.

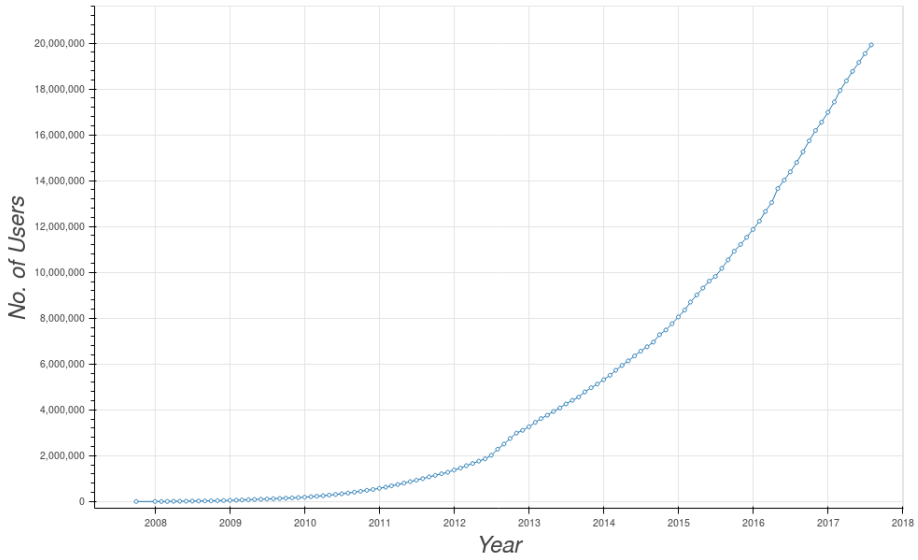

Figure 1: User growth rate

3.1.2 New users per Year-Month

Figure 2 shows that May, 2016 was the month in which most new users were added to the site.

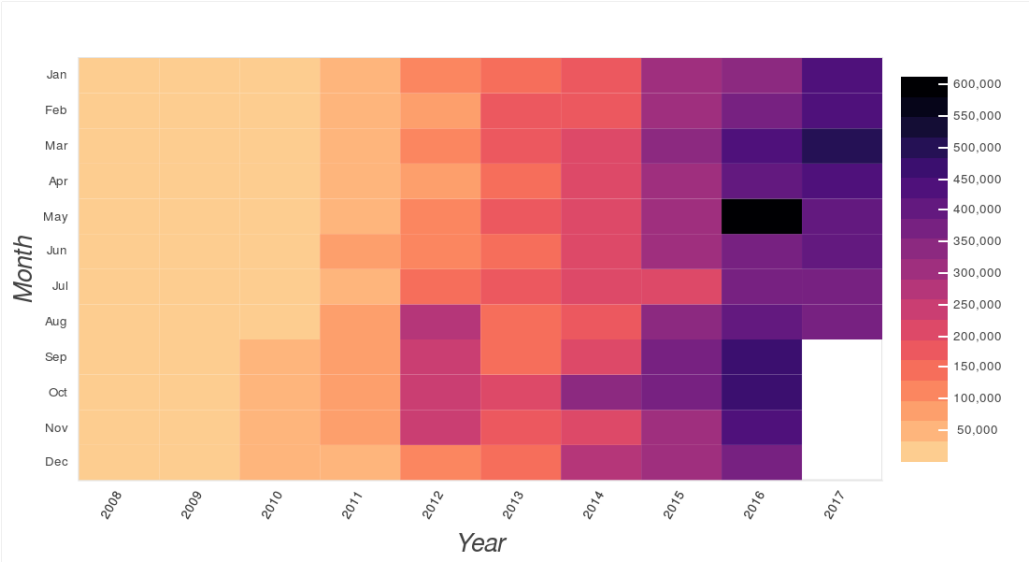

Figure 2: New users added per Year-Month

3.1.3 User distribution across countries

GitHub allows users to enter their location field in a free form text field. Since this user entered data is not validated by GitHub, it can contain anything and does not have to be a valid location. GHTorrent service uses mapping APIs like Bing & Open Street Maps to convert the text data into known locations.

Since not everyone enters their location and or they don't enter it in a valid format, only 7.8% (1,566,019) users have location data that can be used.

Figure 3 shows that majority of such users are from USA, followed closely by India and China.

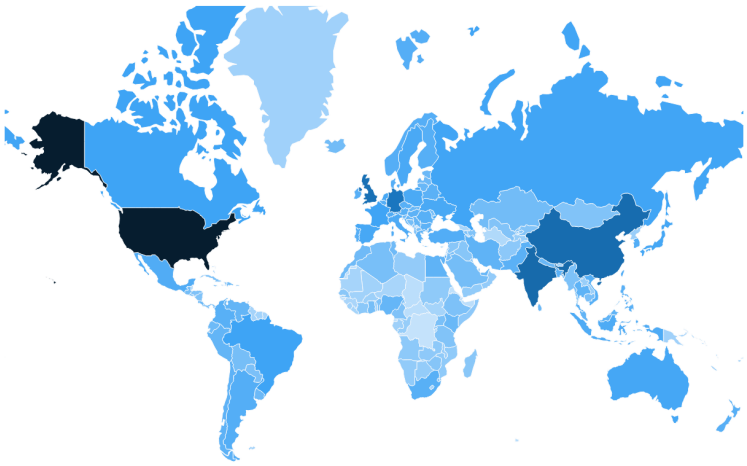

Figure 3: User distribution across countries

3.1.4 User distribution across India

Of the 1.5 million users with mappable location data, only 102,505 are from India. Their state-wise distribution is given in Figure 4.

As can be seen, most GitHub users are from Karnataka and Maharashtra as they are home to IT Hubs like Hyderabad, Bengalore, Pune etc.

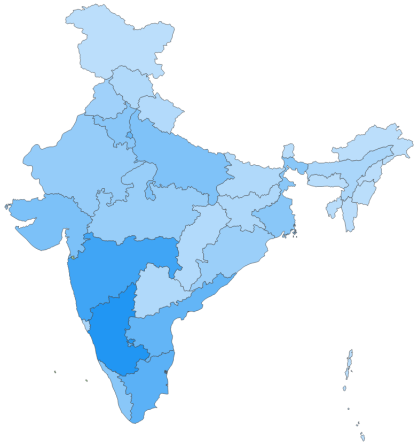

Figure 4: User distribution across India

3.2 Organizations

3.2.1 Top Organizations

On GitHub, users can enter the company the work for in a free form text field. We use that data to find out which organizations have most number of users. Considering only the number of employees of organizations, Microsoft has most number of employees i.e 8148, while IBM only has 2842. (Figure 5)

Another way of finding popular organizations is to see how popular the work done by their users is. This is done by counting their followers and the stars on their repositories. we find that Facebook & Google are at top, perhaps due to a large number of popular projects.

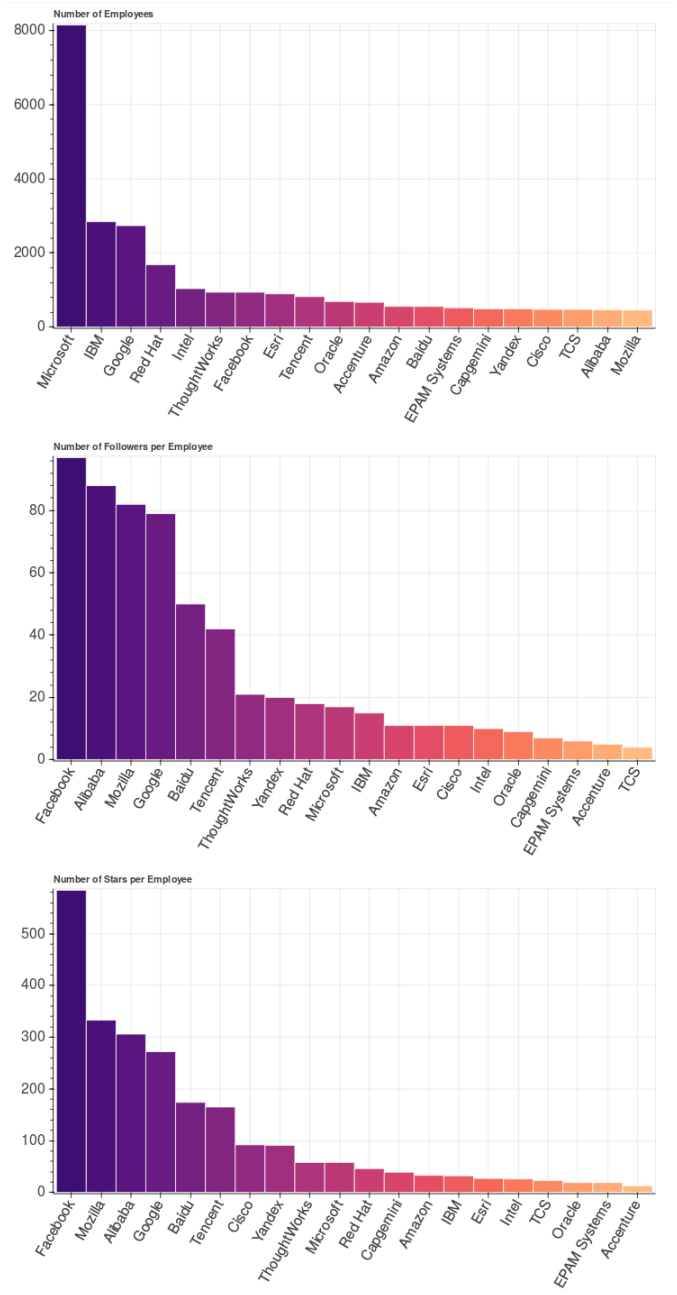

Figure 5: Organisations with most employees (on GitHub)

3.2.2 Top Organizations in India

Extending similar analysis onto organizations whose employees are located in India, we find that TCS has the most employees i.e 434.

Even though Red Hat has very few Indian employees (active on GitHub) their followers per employee & stars per employee value is high indicating that their projects are popular among GitHub users.

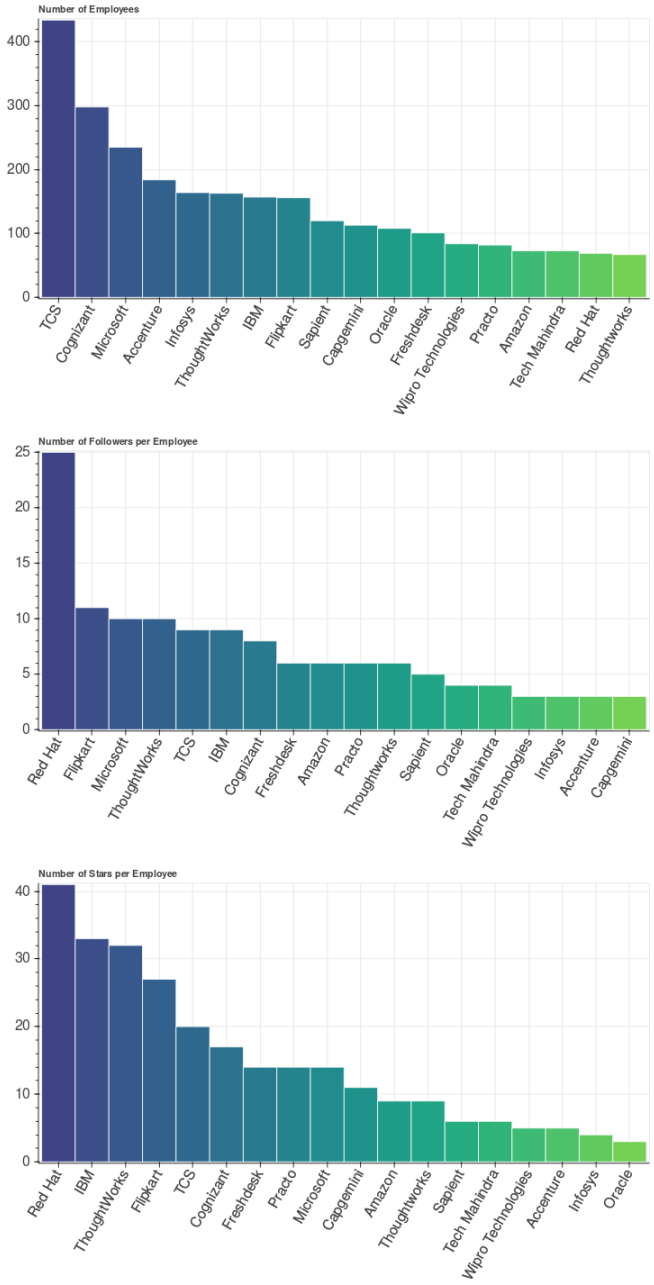

Figure 6: Organisations with most Indian employees

3.2.3 Users at IITs

Since the Company field on GitHub user profiles is a free form text field, students also use it to associate themselves to the Institute the are studying in. By aggregating the data we found, perhaps unsurprisingly, that among all the IITs - IIT Bombay has the most number of users on GitHub, followed by IIT Kharagpur.

IIT Delhi is ranked 5th with 87 users.

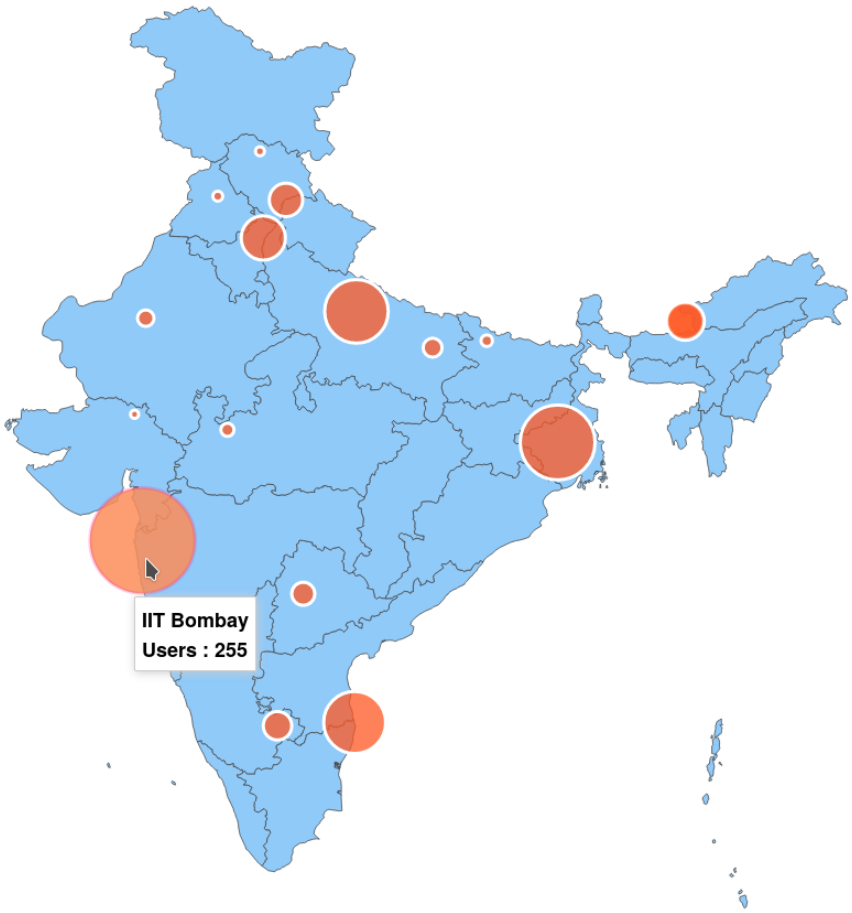

Figure 7: Number of users from various IITs

3.3 Activities

GitHub has a multiple forms of activities that a user can perform on the site - Commits, Issues, Pull Requests. It should be noted that GitHub treats pull request as a special form of Issues.

`commits.csv` contains information of 745,356,807 commits made during 2007 and 2017.

3.3.1 Commit patterns of users

To determine when a user works, we plot a punchcard created using the timestamps of the commits. This is created by considering the users's local timezone (since the GHTorrent data stores it in UTC.)

Comparing Figure 5 & 6, it is clear that both these users have very different work schedules.

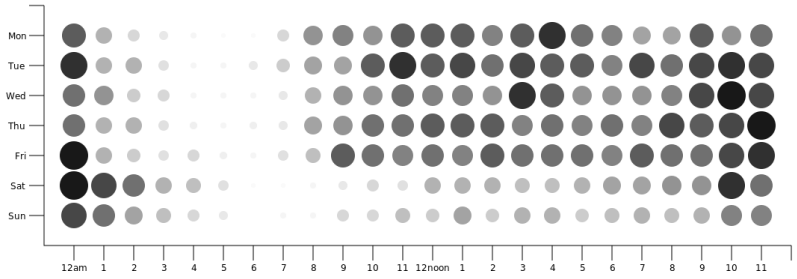

Figure 8: Commit punchcard of user "JakeWharton"

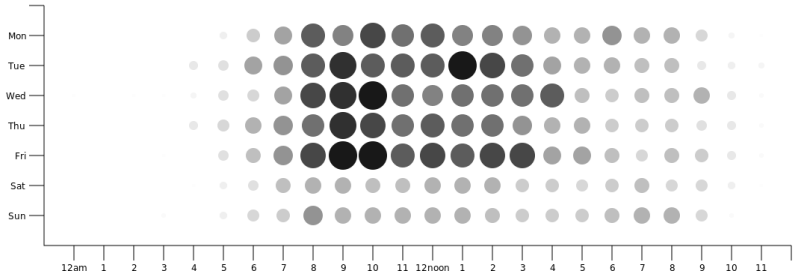

Figure 9: Commit punchcard of user "mbostock"

This can also be extended to entire countries to see when people of a country commit to GitHub.

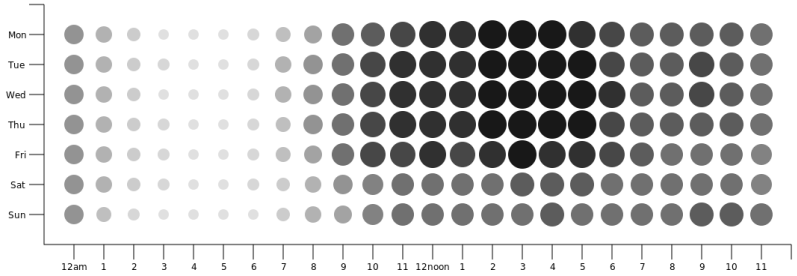

Figure 10: Commit punchcard of all Indian users

3.3.2 Community participation in projects

We defined community participation as the percentage of commits in a repository that were made by someone other than the owner. By analysing the commits data we found that most of the projects on GitHub have 0 participation from others, which means they are done by a single person.

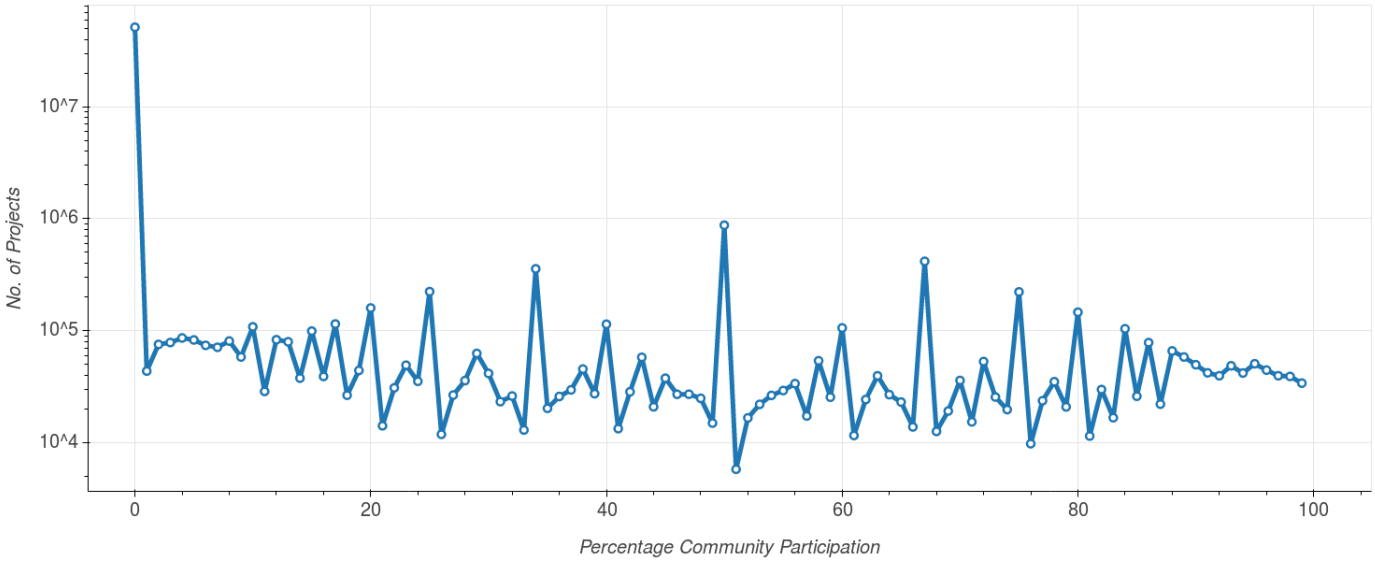

Figure 11: Community participation in projects

3.3.3 Programming Language Popularity

The `project_languages.csv` file contains information on usage of Programming Languages. In Figure 12, we plot the usage of top 10 languages (by their code size in GB).

The data shows that by absolute code size, C was the most used language in 2015 & 2016, while Javascript was the most used in 2017.

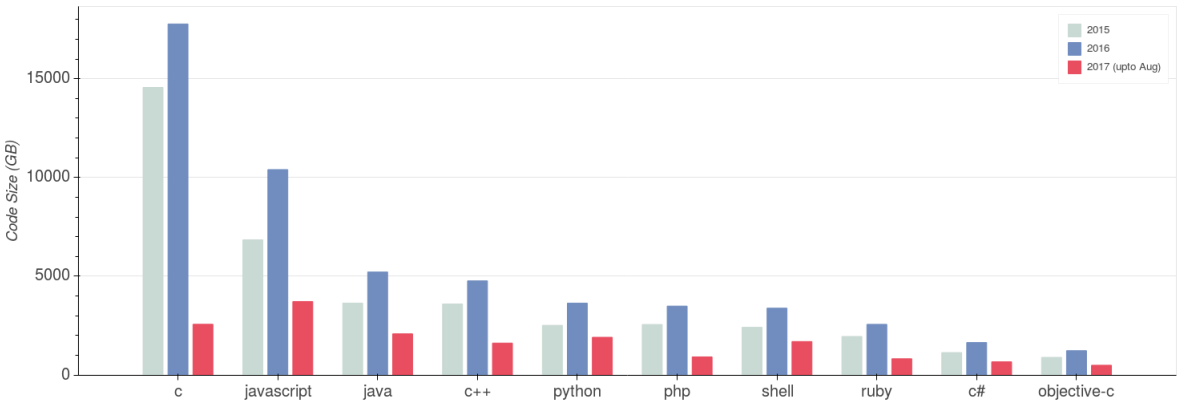

Figure 12: Top 10 programming languages