Analyzing Academic Institutions' Constributions To The Open Source Community Through GitHub Commits

Evan O'Keefe, Jaskaran Bakshi, Yufeng Han, Huan Xu, Kechen Lu

University of Wisconsin - Madison, Madison, WI

Introduction

Analysis on Github contributions is important: According to the paper, Influence analysis of Github repositories by Hu, Y., GitHub is the culture mosaic for promoting programming languages, new development frameworks, and softwares. It is the front-line of bold technical innovation and the cutting-edge of the technological landscape. When taking a closer look, the influence of GitHub is built commit by commit, repository by repository by the open-source contributors. However, the existing analysis like Open Source Index mainly focused on how big tech cooperations contribute to the community, but gave little to no attention to the performance of the academica. Therefore, to bridge the gap, we analyzed GitHub contribution data of U.S. academic institutions collected from Google's public data warehouse "github_repos". Specifically, we used Node-Link Network diagrams and bar charts to show how individuals from academic institutions contribute to the broader GitHub open-source community. By analyzing individual's contributions and shared history, we can draw important conclusions on how academic institutions interact with the greater open-souce community.

Methods

Goal

Our group strived to observe the impact of academic institutions have on the open-source community through analyzing the number of commits at a user and institution wide level.

Data Processing

Any rows that contained null or NA values were dropped with dyplr's drop_na function. Next the institution names were extracted from email domain from the respective columns column.

Visualizations

Our first visualization created is reactive bar chart displaying the number of commits with options to select grouping the data by users or institutions. Next a static node-link diagram shows the collaborations across U.S. academic institutions. Finally we created a reactive node-link diagram revealing the connections across all committers within the user-selected academic institution.

Results

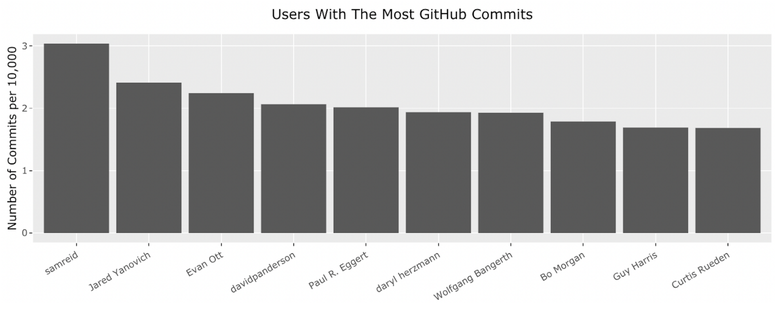

Figure 1: **View of the Reactive Bar Chart.** This is one of the many views from the reactive bar chart currently showing the top 10 users based on their commit numbers. The options to change the number of bars and aggregation to the institutional level are displayed at the top.

To allow any user to rapidly explore the data a reactive bar chart with various different viewing options was created. The bar chart defaults to the top ten users based on their number of commits. Plotly tooltips allow the user to mouse over each bar to see the precise number of commits as well as their academic institution. There is a numeric input box that allows users to change the number of bars to explore more or fewer users/institutions. The y-axis scaling will update dynamically based on the smallest value and the axis title text will update to match. By selecting the "Institutions" radio button the view will switch to the institutions with the most total commits. These views can be further augmented by clicking on an individual bar which will bring up that institution's top users.

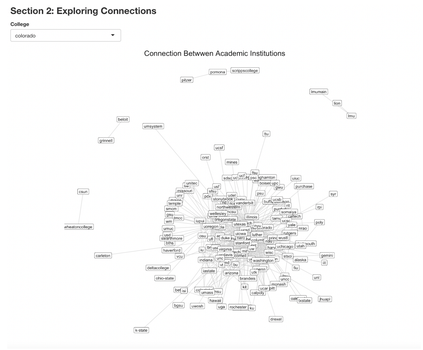

Figure 2: **View of the Static Node-Link Diagram.** This diagram visualizes the connections across institutions based off their member's shared repositories.

This static node-link diagram shows the connections across U.S. academic institutions, where two universities are said to be connected if they both commit to the same repositories. From the visualization, we can identify 5 clusters of academic institutions. At the center, the biggest cluster reveals the broader collaboration landscape for most U.S. universities, nodes that are at the center-most of this cluster (e.g. MIT, UW, NYU) are most widely connected with others and therefore more active and collaborative in the open-source community. On the other hand, institutions at the periphery of this cluster (e.g. UIUC, WM, Haverford) only collaborated with limited, specific institutions, suggesting that they were less collaborative in the open-source community.

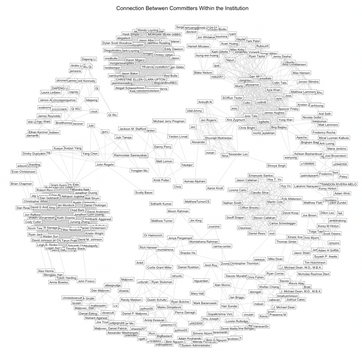

Figure 3: **View of the Static Node-Link Diagram.** Connections across committers within an intitution (in this case, the University of Utah).

This reactive node-link diagram reveals the connections across all committers within the user-selected academic institution. Like the previous graph, two commiters are said to be connected if they both commit to the same repositories. This graph gives us 2 valuable insights: 1) each cluster in the graph can be interpreted as a large-scale collaborative open-source project led by the selected academic institution, where nodes in the cluster (students, staff members, and professors) worked closely to collaborate and contribute. From this graph, we can conclude that, across the history, the University of Utah led 4 such large-scale open-source projects. 2) This graph provides a finer look into the social and performance aspects of individual contributors. For example, we can discover interdisciplinary contributors by finding nodes that connect to multiple clusters. We can also identify the technological "social network" of an academic institution.

Conclusion

Overall, we can see how a vast network of users contributes to various open source repositories through different institutions. We can clearly see that MIT has a vast collection of users contributing to many open source platforms which is a good indication that our analysis is in the right direction because a majority of this platform hosts many of their institutional repositories. With the combination of the bar graph and the network graph we hope to allow users to understand universities creation of repositories by also understanding the users who are pushing the most code in specific universities.

Improvements

Upon reviewing the results from the views of the bar chart an issue was observed where users with the same name would be stacked and display in an awkward manner. In trying to find a solution to the issue it was observed that these were likely the same user under different accounts/emails e.g. @wisc.edu vs @stat.wisc.edu. In order to resolve this issue the hashed email was extracted from the domain and aggregated by this value. To account for individuals who may have changed institutions (Figure 4) within their careers the group by was performed on the hash and institution. The main issue from this process arose with the differing names for these duplicate hashes (Figure 5). A temporary fix for this issue was seleted as summarizing by the maximum name value. This could be a future modification that may bring interesting results if any occur.

19651	929	Doug Torrance	00c8d413d4544a729e5f2cf3d0cb2336c4b3a017@m...	monmouthcollege	00c8d413d4544a729e5f2cf3d0cb2336c4b3a017
4917	239	Doug Torrance	00c8d413d4544a729e5f2cf3d0cb2336c4b3a017@pi...	piedmont	00c8d413d4544a729e5f2cf3d0cb2336c4b3a017

Figure 4: **Example of a Change of Institution.** The above figure shows one user determined by the hashed email address that is present under two different institutions.).

122	Chase 🌪	00017bbd69093c141362359d6a6bed2dc7b7d0a4@u...	uic	00017bbd69093c141362359d6a6bed2dc7b7d0a4
42	Chase Lee	00017bbd69093c141362359d6a6bed2dc7b7d0a4@u...	uic	00017bbd69093c141362359d6a6bed2dc7b7d0a4
1	CDC HMI	00017bbd69093c141362359d6a6bed2dc7b7d0a4@u...	uic	00017bbd69093c141362359d6a6bed2dc7b7d0a4
1	clee231	00017bbd69093c141362359d6a6bed2dc7b7d0a4@u...	uic	00017bbd69093c141362359d6a6bed2dc7b7d0a4

Figure 5: **Example of One User Multiple Names.** The figure above shows a single user that has four rows associated with their unique hash due to the various names. .

References

Hu, Y., Zhang, J., Bai, X. et al. Influence analysis of Github repositories. SpringerPlus 5, 1268 (2016). https://doi.org/10.1186/s40064-016-2897-7

Florian Zandt. How Big Tech Contributes to Open Source. Statistica (2021). URL: https://www.statista.com/chart/25795/active-github-contributors-by-employer/