Understanding the Impact of COVID-19 on Github Developers: A Preliminary Study

Liu Wang[1], Ruiqing Li[2], Jiaxin Zhu[3], Guangdong Bai[2], Weihang Su[1], Haoyu Wang[1]
[1] Beijing University of Posts and Telecommunications, Beijing, China
[2] The University of Queensland, Australia [3] Institute of Software, Chinese Academy of Sciences, China

***Abstract*—The ongoing COVID-19 pandemic has impact almost every aspect of human lives profoundly. This paper investigates the impact of COVID-19 on the activity and contribution of open source software (OSS) developers. Specifically, we make great efforts to harvest the information of all the developers (over 25 million) on GitHub and their contribution activities. With such a large-scale dataset, we perform analysis from four perspectives, including the overall ecosystem level, country level, organization level and developer level, to characterize the impact of COVID-19 on the OSS community. We have revealed a number of interesting observations and trends, which are crucial to understanding the OSS contributors and supporting the collaboration to combat global crisis like COVID-19.**

***Index Terms*—COVID-19, Github, Developer Contribution, Open Source Software**

I. Introduction

The COVID-19 pandemic has changed people's lives in many ways. In this paper, we intend to take a look at the GitHub developers and their contributions during the pandemic, and share our preliminary results.

GitHub developers can make many kinds of contributions. Commits, pull-requests and issue reports are the most commonly studied contributions in literature. In this paper, we consider the comprehensive *contributions* monitored and recorded by GitHub, which has been seldom analyzed before. In particular, 'committing to a repository's default branch or gh-pages branch', 'opening an issue', 'proposing a pull request', and 'submitting a pull request review' are counted as contributions [1]. GitHub presents a *contributions graph* for each user in his/her profile page. It tracks the number of daily contributions and visualizes how active one has been on the site, which we believe is a good indicator to measure the activeness of GitHub developers during COVID-19.

Basically, we study the number of contributions per day and the number of developers who make these daily contributions before and during the pandemic. We first make a huge effort to harvest a comprehensive dataset of developers in Github and their daily contributions. By the time of November 10, 2020, we have collected 25,761,884 developers in total. Leveraging the dataset, we perform a systematical analysis including four perspectives, i.e., the overall ecosystem level, country level, organization level and developer level, which corresponds to four research questions that drive our study. Our first research question is how the overall numbers change (**RQ1**). GitHub developers are from all over the world, and some of them are corporate employees. The severity of the pandemic varies across countries. Different companies also have different strategies to handle the situations. Therefore, we raise the next two research questions, how the numbers change within different countries (**RQ2**) and different companies (**RQ3**). The developers are of different activeness, and we also want to see the trends among them (**RQ4**).

We observe that there is a remarkable increase in developer participation and contribution to GitHub in the early stages of the COVID-19 global explosion. In terms of individual countries, the trends of GitHub developer contributions are tied to the outbreak in the corresponding country. On a company level, the work-from-home setting implemented due to the pandemic seems not to have disrupted software developers' commitment to GitHub. For developers of various activeness levels, their involvement and contribution to GitHub more or less increased during the COVID-19 outbreak, even for those developers who are very inactive. Our observations suggest that COVID-19 did not pose great challenges to GitHub developers and show that GitHub may have played an important role in COVID-19, tapping into the potential of OSS communities like GitHub to respond to public crises.

II. Related Work

Since its outbreak, COVID-19 has attracted great attention from various research communities. A large number of studies were focused on the medical domain [2]–[4]. In the field of mobile app analysis, Wang et al. [5] conducted a systematic analysis of coronavirus-themed mobile malware and revealed huge potential threats beyond the virus. There is also a number of research on COVID-19 in the software engineering community. For example, Wang et al. [6] presented a large-scale empirical study of COVID-19 themed repositories on GitHub and highlighted the practical and potential value of open source technologies and resources in handling such crisis.

In addition, GitHub developers (users) have been studied from many perspectives in literature. Technical roles of the developers and experts of the open source projects are automatically identified via machine-learning based approaches [7], [8]. Contributions of novice developers are characterized [9], and the effect of developer sentiment on fix-inducing changes is analyzed [10]. Collaboration among GitHub developers is investigated to identify the characteristics favoring innovation in the open source community [11]. Researchers also studied

DOI reference number: 10.18293/SEKE2021-132

the motivation behind following others and the influence of popular users on their followers [12]. Besides, some studies have attempted to understand developer productivity at technology companies due to the almost overnight migration of software developers to work from home. For example, Bao et al. [13] presented a case study on Baidu Inc. to investigate the difference of developer productivity between working from home and working onsite. Ford et al. [14] conducted survey studies to understand the benefits as well as challenges of working from home and analyze the factors that have affected developer productivity over time. Ralph et al. [15] performed a questionnaire survey study to investigate the effects of the pandemic on developers' well-being and productivity.

The researches above only focus on a small number of GitHub developers. Chatziasimidis and Stamelos conducted some measurements and association mining with l00K projects and 10K GitHub users/owners of the projects [16]. Different from those studies, in this paper, we target at all the GitHub developers to understand the impact of COVID-19.

III. Data Collection

Aimed at all the GitHub developers, a large-scale dataset is retrieved from GitHub. We employ the GitHub search API [17] to find all the GitHub developers who have at least one public repository. Considering the limitation of 1,000 results per search, we have adopted a segmented approach, i.e., narrowing down the results of a single query using some search qualifiers and performing multiple queries to further consolidate the results to collect a complete list of developers. To do this, we filter the developers based on when they joined GitHub with the *created* qualifier, which takes a date as its parameter with optional time information after the date to search by the hour, minute, and second. *As a result, we collect a total of 25,761,884 developers whose GitHub accounts were created from 2007 to 2020.* Additionally, we plot the distribution of creation dates for all developers in our dataset in Figure 1. It can be observed that the number of daily new developers overall exhibits a continuous upward trend and has a peak in mid-2019, which in a way showcases the growing market of Github over the past decade. The basic information we collect includes *id*, *login*, *name*, *location*, *company*, etc. Of which, the location field reveals that these developers are located in almost every country in the world today, and the distribution map of the number of developers per country is shown in figure 2. The number of developers providing locations is considerably higher in the United States than in other countries. More importantly, for all the developers collected, we capture their daily contributions (e.g., commits, issues and pull requests) from their creation date to November 10, 2020 by crawling the profile pages. The entire data collection involves *over 26 million* HTTP requests, which is a very time and network resource consuming task.

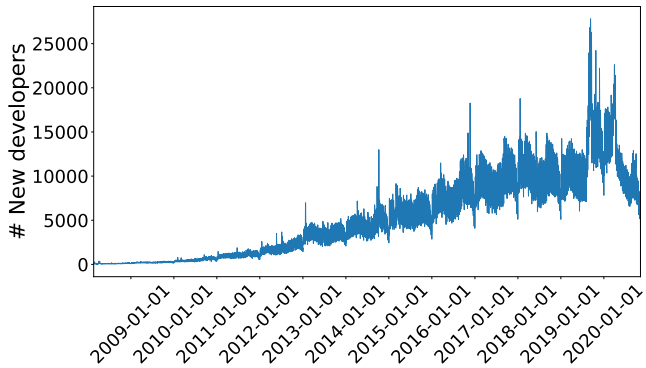


Fig. 1. Distribution of creation dates for all collected developers.

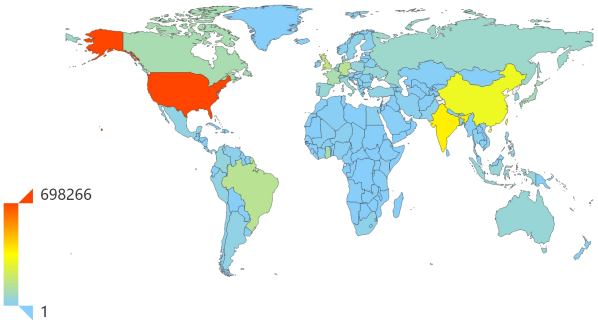


Fig. 2. Distribution of located countries of the collected developers.

IV. Empirical Results

We present the answers to our research questions of the overall ecosystem level, country level, organization level and per-developer level in the following sub-sections.

A. RQ1: Overall Trends

Figure 3 shows the sum of contributions per day in 2019 and 2020 for all developers whose accounts were created before 2019. We can examine the trend of contributions in three stages: (1) *Prior to mid-March*, i.e., before the global pandemic of COVID-19, the total number of daily contributions exhibits a largely stable trend, with only a slight decline in late-2019 and early-2020, which probably due to the Christmas and New Year holidays. (2) *From mid-March to the end of June*, the period when COVID-19 is exploding worldwide, and the number of confirmed cases is rising sharply, meanwhile, the daily contribution of developers shows a clear trend of increase, peaking in April and gradually falls back in May. (3) *After July*, the trend has dropped to flat, but appears to be slightly higher than that in 2019.

We then use Mann-Whitney U test [18], a popular non-parametric test to compare outcomes between two independent groups, to test whether there is a statistically significant difference between developer contributions in 2019 and 2020. Specifically, we generate two samples containing developer contributions in 2019 and 2020, then calculate the test on the samples and print the statistic and p-value (2.43e-35). Typically, if the p-value is below 0.05, the test says there

is enough evidence to reject the null hypothesis and that the samples were likely to have different distributions. Thus this p-value (2.43e-35) strongly suggests that the sample distributions are different, i.e., *the developer contributions in 2020 is statistically significantly different from that of 2019*, as expected. *This observation suggests that the activity of developers is not significantly disrupted by the pandemic. Instead, it has somewhat boosted developers' engagement in GitHub, especially in the early stages of the global pandemic.* In a way, it also reflects that the open source community and collaborative platforms like GitHub may have an important role to play in the face of a public crisis such as COVID-19. Beyond that, it is interesting to see that the trend is cyclical, and the periodicity is one week upon our inspection, which is quite likely related to the weekly work schedule of most employees, even during the pandemic.

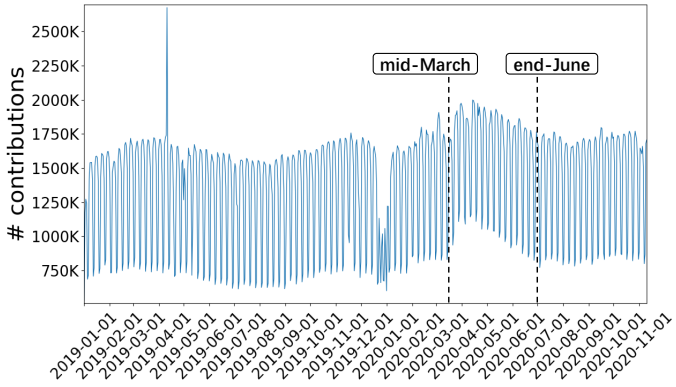
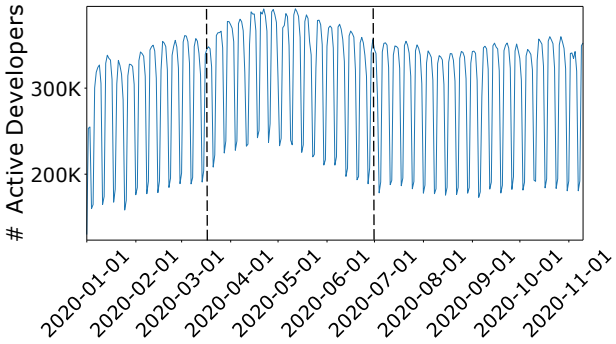
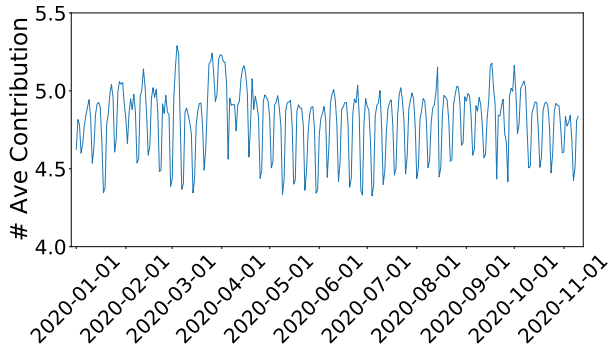


Fig. 3. Distribution of the total number of contributions per day for developers whose accounts were created before 2019.

Further, we attempt to explore the reasons for the increase of the daily developer contributions in the second stage from two aspects. On the one hand, we count the number of *active developers* whose contribution value is greater than 0 for each day in 2020, as shown in Figure 4(a). We can observe that the trend is similar to that in Figure 3, with a significant peak-like trend experienced from mid-March to the end of June. This suggests that the number of developers making contributions per day has increased during the second stage, which leads to the trend of Figure 4(a). On the other hand, we calculate the *average contribution value* of all active developers for each day as displayed in Figure 4(b). It can be seen that the average values are very close, fluctuating from 4.2 to 5.2. Although a slight jump in the average can be seen at the end of March, the increase is too small to cause the significant growth in the second stage. To sum up, *there was a significant increase in the number of active developers in GitHub during the period when COVID-19 began to wreak havoc around the world.* In particular, regarding the outlier that appears on April 11, 2019 in Figure 3, we tried to explain it by manually checking the number of active developers and the average contribution of active developers during this period. We found that the number of active developers on this day was not particularly high but the average contribution showed a similar outlier, suggesting



(a) Active Developers



(b) Average Contribution

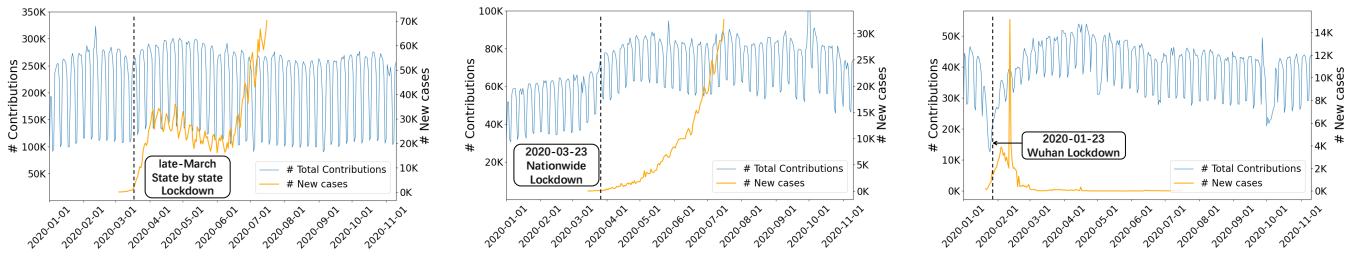
Fig. 4. The trend of active developers and average contribution per day for developers created before 2020.

that some developers may have made a substantial amount of contributions on this day.

B. RQ2: Country-level

As aforementioned, we collect the location information of all developers, while unfortunately 85% of them did not provide their locations. Among the available locations (3,821,170), different formats/representations are used, including city name only, country name only, and both. Thus we collect a list of countries in the world and major cities in each country and use the string matching method to determine the country to which the location belongs. Finally we acquire a total of 3,647,523 (95.5%) valid locations mapping 195 countries. The top three countries in regards to the number of developers are the United States (698,266), India (379,286) and China (300,496). We explore whether the contribution of developers is significantly associated with the corresponding pandemic situation at the national level within the three leading countries. For each country, we collect the daily number of new COVID-19 confirmed cases over time from Google Statistics [19] and place them in charts for comparison, as well as marking the start of lockdown, as shown in Figure 5. It can be seen that all three countries appear to have a rapid rise in daily contributions in the early stage of the outbreak, even though at different rates, e.g., China and India rise significantly, while the US rises slightly.

For the sake of statistical validity, we use the Pearson correlation coefficient to analyze the relationship between total



(a) US (b) India (c) China

Fig. 5. Distribution of total contributions for developers in three countries created before 2020.

contributions and pandemic situation for each country. Since the impact of COVID-19 on developer contributions is mainly observed at the beginning of the outbreak, we calculate the correlation between the number of newly confirmed cases and developer contributions in the one-month period following the lockdown for each country. We find that the developer contributions of the three countries were all positively correlated with their corresponding pandemic growth in varying degrees, with stronger correlations in China (r=0.35, p-*value*=0.06) and India (r=0.29, p-*value*=0.12), and weaker correlations in the United States (r=0.04, p-*value*=0.84). The correlations, though, are not statistically significant (which is speculated due to the volatility of the contributions). Hence, *it seems that the growth of developer activity and contributions to GitHub are correlated with the outbreak in the corresponding country.*

C. RQ3: Organization-level

COVID-19 has forced companies all over the world to adapt to and embrace remote work—at least for the short term. Large tech employers such as Apple, Google, Facebook and Microsoft are among the first to ramp up remote work plans for many or all of their employees around the globe in March. As aforementioned, we also collect the company information of all developers. There are 1,835,336 (7.1%) developers providing their company information. With these available data, we count the number of developers in major companies around the world, and obtain the top three, i.e., Microsoft (21,296), Google (10,041) and IBM (8,101). Figure 6 shows how the activity and contributions of developers working at these companies have changed in 2020. *It can be seen that, as companies move into "work from home" mode in response to COVID-19 at the end of March, developer activity and contributions to GitHub have increased to varying degrees, and although some companies (e.g. Microsoft) have not shown a significant increase, none have decreased.* It suggests that companies are likely to rely more on open source communities and collaboration platforms like GitHub for their work-from-home efforts and also implies the potential and usefulness of these platforms in the face of such public emergencies.

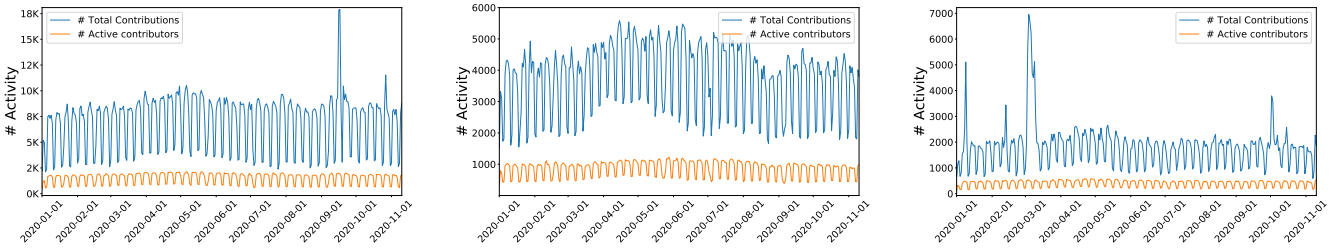
D. RQ4: Developer-level

As reported in §IV-A, the number of active developers on GitHub increased significantly during the rise of COVID-19, thus we would like to learn more about what kinds of developers became active during this period. We seek to categorize all the developers into several types based on how active they are on GitHub. First, for each developer, we calculate the percentage of days they are active on GitHub out of the total number of days they have been on GitHub, and find that it is very small for the vast majority of developers, with less than 1% of developers having more than half of the total number of days active. Then, we group all developers (excluding those who have never contributed) into four levels, i.e., very active, moderately active, inactive and very inactive, as shown in Table I. Figure 7 presents the trend of the number of active developers per day for each type of developers separately. As can be seen, the activity of all types of users increased to a greater or lesser extent at the beginning of the COVID-19 global outbreak, with the less active developers increasing considerably, while the more active users showing relatively minor changes, suggesting that a lot of inactive developers got involved in the use of GitHub during this period. Similarly, Figure 8 shows the trend of the number of daily contributions for each type of developers, which is generally very similar to Figure 7. Both charts show that contributions from any type of developers increased as COVID-19 began to explode, and that even among very inactive groups many more people contributed to GitHub during this period.

TABLE I
FOUR TYPES OF DEVELOPERS BASED ON THE ACTIVITY LEVEL.

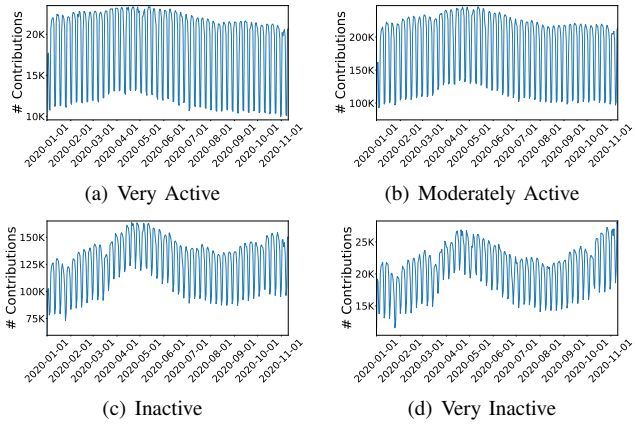
Type	Proportion of Active Days	# Developers
Very Active	[0.5, 1]	28,377
Moderately Active	[0.1, 0.5)	694,689
Inactive	[0.01, 0.1)	4,302,035
Very Inactive	(0, 0.01)	31,838,464

To gain a deeper understanding of developers' activities, we next explore how the developer contributions evolve over time and try to group the evolution into several patterns. Since inactive users make little or no contribution for most of the time, we only focus on the very active developers (the first type in Table I) and examine the trends of their contributions in the year 2020. For each active developer, we retrieve the contribution he/she made on Github per day over the year, take them as feature vectors and perform L2 regularization [20] to apply the K-Means clustering. We use the Elbow Method to find the optimal value of k, which is 3. Thus we group the developer contributions into three clusters. Figure 9 presents



(a) Microsoft (b) Google (c) IBM

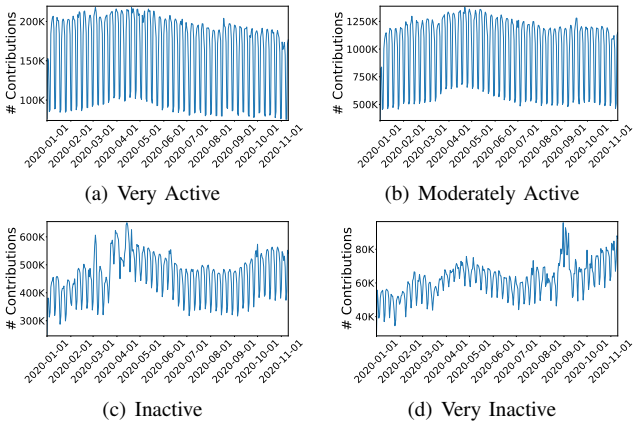
Fig. 6. Distribution of active developers and total contributions per day for developers in three companies created before 2020.



(a) Very Active (b) Moderately Active

(c) Inactive (d) Very Inactive

Fig. 7. Distribution of the number of active developers per day for four types of developers.



(a) Very Active (b) Moderately Active

(c) Inactive (d) Very Inactive

Fig. 8. Distribution of total contributions per day for four types of developers.

the trends of contributions in each cluster. It can be seen that each cluster captures the weekly work cycle. We next briefly describe the characteristics and examples of different patterns.

Cluster 1 (in red) shows an overall downward trend, gradually declining from March until it stabilizes after August. There are 6,366 (22.4%) active developers whose contribution trends belong to this cluster. This pattern reflects the presence of some active developers whose contributions and activeness have decreased after the outbreak. For example, a developer with the account name 'najahiiii' has a contribution trend showing that his daily contribution fluctuates between 7 and 38 in the first three months of 2020 with a daily average

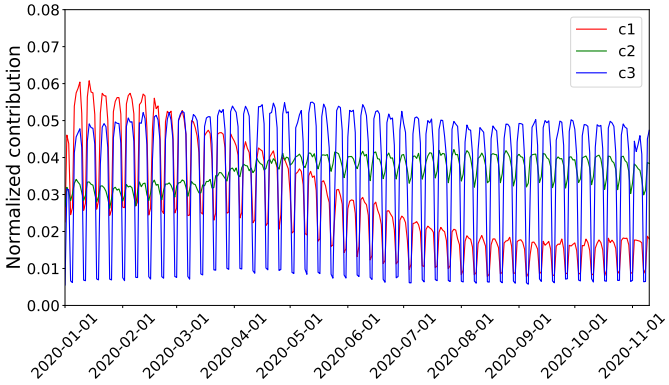


Fig. 9. Three patterns of developer contributions.

contribution value of 12.2, while in the following months it has stayed below 15 with an average daily contribution of 8.

Cluster 2 (in green) indicates an overall upward trend, the opposite of Cluster 1, which started to rise gradually from March to stabilize. 8,394 (29.6%) active developer contribution trends fall into this cluster. This pattern suggests that there are some active developers who have become more active after the outbreak. For example, one developer with the account name 'pavangoyal42' contributed mostly under 10 per day in the first three months of 2020, and increased significantly from April to July, with an average daily contribution of 17.7.

Cluster 3 (in blue) implies a relatively stable trend with little change in the contribution before and after the outbreak. This seems to capture the normal state of work for developers, and they are hardly affected by COVID-19. Roughly a half of the active developers (48%) are in this cluster. For example, a developer with the account name 'lambert-p' has been contributing in a balanced manner in 2020, in a typical weekly work pattern, with more contributions during weekdays and less or no contributions on days off.

V. Discussion

A. Implications

Our investigation addresses the need to provide scholarly evidence concerning how the COVID-19 pandemic affected GitHub developers' contributing activities. On the whole, we register a significant increase in developer activity and contributions on GitHub in the early stages of the COVID-19 global explosion. On an individual country basis, the trend

in GitHub developer contributions is closely related to the outbreak situation in the corresponding country. Our findings also show that working from home as practiced by companies does not affect the commitment of software developers to GitHub. Besides, developers with different levels of activeness have made more contributions during COVID-19, especially inactive developers. To conclude, our research implies that GitHub, as a typical representative of open source communities and sharing platforms, plays an important role in the face of public crisis. It should be noticed that these technologies and resources can be very helpful in other emergencies too. Relevant participants (software engineering practitioners and researchers, etc) should take note of these findings and understand the strengths and usefulness of Github in COVID-19 pandemic, in order to make fuller use of it as a powerful weapon in our response to the crisis.

B. Limitations and Future Work

We recognize that our study carries several limitations and potential threats to validity. First, some of our conclusions rely on case studies, such as country-level and organization-level analyses, and do not include the full range of data. This is mainly because our preliminary study did not concern a lot of workload and made it a focus for future work. Second, this study aims to analyze the impact of COVID-19 on the open source community, however GitHub is not the only platform where people can share their open source projects, which might limit our observations. Third, we only consider the coarse-grained contribution of each developer. Although the overall contribution is representative enough to reflect the activeness of developers, enabling fine-grained analysis of each kind of contribution can offer us more insights, which will be studied in the future. Forth, we acknowledge that there may be some non-developer users in our dataset since Github is being used for purposes other than software development. However, it is impractical for us to identify whether each user is a real developer or not. Nevertheless, all the developers considered in this paper have created at least one public repository (see Section III), thus we believe that the observations in this paper could reflect the general behaviors of Github developers.

VI. Conclusion

This research focuses on GitHub developers' contributions and presents the first large scale empirical study of the impact of COVID-19 on the activity and contributions of GitHub developers. We go to great lengths to collect a dataset of over 25 million GitHub developers and characterize them from four perspectives including the overall ecosystem level, country level, organization level and developer level, to understand the impact of COVID-19 on the open source software community. Our observations suggest that COVID-19 does not present a challenge for GitHub developers, and show the promising direction of applying open source technologies and resources to response to public emergencies.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (grant numbers 62072046 and 61702045). Haoyu Wang is the Corresponding author.

References

[1] "Github documentations," https://docs.github.com/en/github/setting-up-and-managing-your-github-profile/viewing-contributions-on-your-profile, 2020.

[2] Y. Chen and L. Li, "Sars-cov-2: virus dynamics and host response," *The Lancet Infectious Diseases*, vol. 20, no. 5, pp. 515–516, 2020.

[3] D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona, B. S. Graham, and J. S. McLellan, "Cryo-em structure of the 2019-ncov spike in the prefusion conformation," *Science*, vol. 367, no. 6483, pp. 1260–1263, 2020.

[4] V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt *et al.*, "Detection of 2019 novel coronavirus (2019-ncov) by real-time rt-pcr," *Eurosurveillance*, vol. 25, no. 3, p. 2000045, 2020.

[5] L. Wang, R. He, H. Wang, P. Xia, Y. Li, L. Wu, Y. Zhou, X. Luo, Y. Guo, and G. Xu, "Beyond the virus: A first look at coronavirus-themed mobile malware," *arXiv e-prints*, pp. arXiv–2005, 2020.

[6] L. Wang, R. Li, J. Zhu, G. Bai, and H. Wang, "When the open source community meets covid-19: Characterizing covid-19 themed github repositories," *arXiv preprint arXiv:2010.12218*, 2020.

[7] J. E. Montandon, M. T. Valente, and L. L. Silva, "Mining the technical roles of github users," *Inf. Softw. Technol.*, vol. 131, p. 106485, 2021.

[8] J. E. Montandon, L. L. Silva, and M. T. Valente, "Identifying experts in software libraries and frameworks among github users," in *Proceedings of the 16th International Conference on Mining Software Repositories, MSR 2019, 26-27 May 2019, Montreal, Canada*, M. D. Storey, B. Adams, and S. Haiduc, Eds. IEEE / ACM, 2019, pp. 276–287.

[9] I. Rehman, D. Wang, R. G. Kula, T. Ishio, and K. Matsumoto, "Newcomer candidate: Characterizing contributions of a novice developer to github," in *IEEE International Conference on Software Maintenance and Evolution, ICSME 2020, Adelaide, Australia, September 28 - October 2, 2020*. IEEE, 2020, p. 855.

[10] S. F. Huq, A. Z. Sadiq, and K. Sakib, "Understanding the effect of developer sentiment on fix-inducing changes: An exploratory study on github pull requests," in *26th Asia-Pacific Software Engineering Conference, APSEC 2019, Putrajaya, Malaysia, December 2-5, 2019*. IEEE, 2019, pp. 514–521.

[11] D. Celinska, "Coding together in a social network: collaboration among github users," in *Proceedings of the 9th International Conference on Social Media and Society, SMSociety 2018, Copenhagen, Denmark, July 18-20, 2018*. ACM, 2018, pp. 31–40.

[12] K. Blincoe, J. Sheoran, S. P. Goggins, E. Petakovic, and D. E. Damian, "Understanding the popular users: Following, affiliation influence and leadership on github," *Inf. Softw. Technol.*, vol. 70, pp. 30–39, 2016.

[13] L. Bao, T. Li, X. Xia, K. Zhu, H. Li, and X. Yang, "How does working from home affect developer productivity? – a case study of baidu during covid-19 pandemic," 2020.

[14] D. Ford, M.-A. Storey, T. Zimmermann, C. Bird, S. Jaffe, C. Maddila, J. L. Butler, B. Houck, and N. Nagappan, "A tale of two cities: Software developers working from home during the covid-19 pandemic," 2020.

[15] P. Ralph, S. Baltes, G. Adisaputri, R. Torkar, V. Kovalenko, M. Kalinowski, N. Novielli, S. Yoo, X. Devroey, X. Tan *et al.*, "Pandemic programming: how covid-19 affects software developers and how their organizations can help (2020)," *arXiv preprint arXiv:2005.01127*, 2005.

[16] F. Chatziasimidis and I. Stamelos, "Data collection and analysis of github repositories and users," in *6th International Conference on Information, Intelligence, Systems and Applications, IISA 2015, Corfu, Greece, July 6-8, 2015*, N. G. Bourbakis, G. A. Tsihrintzis, and M. Virvou, Eds. IEEE, 2015, pp. 1–6.

[17] "Github search api," https://docs.github.com/en/rest/reference/search, 2020.

[18] P. E. McKnight and J. Najab, "Mann-whitney u test," *The Corsini encyclopedia of psychology*, pp. 1–1, 2010.

[19] "Google statistics," https://news.google.com/covid19/map?hl=en-US&mid=%2Fg%2F11j3990nlq&gl=US&ceid=US%3Aen.

[20] A. Nagpal, "L1 and l2 regularization methods," https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c, 2017.