Analyzing Developer Collaboration on GitHub Before and During COVID-19: Activities, Networks and Sentiment.

Project Report

Jothish Kumar Polaki, Project Member

ADVISOR: Dr. Xuan Lu

INSTRUCTOR: Dr. Greg Chism

Date: 05/14/2025

Contents

1. Abstract

The COVID-19 pandemic changed the way developers worked, with many moving to remote collaboration. This project explores how developer activity on GitHub changed before and during the early months of the pandemic by comparing data from February to April in 2019 and 2020. Using GitHub Archive data with BigQuery and Dask, we analyzed trends in developer activity, collaboration networks, and the tone of communication. Key metrics like average degree, network density, clustering coefficient, etc. were calculated from weekly graphs and VADER was used for sentiment analysis. The analysis also included a focused look at developers active in both years to see any change in behavior. This analysis gives insight into how collaboration patterns shifted during a global event. The findings may help platform maintainers, researchers and engineering teams understand the behavioral impact of sudden global disruptions on developer ecosystems.

2. Introduction

GitHub is the most widely used platform for collaborative software development, where millions of developers work together on open source and private projects. In early 2020, the COVID-19 pandemic forced many teams to shift suddenly to remote work, raising questions about how this change affected developer collaboration.

While some past research looked at productivity during the pandemic, there has not been much large scale analysis using real developer activity data. GitHub Archive offers detailed event logs that make it possible to study how developer behavior and collaboration patterns changed during this time.

This project compares GitHub acitivity from February to April in 2019 (before COVID) and the same period in 2020 (early COVID). It focuses on three main questions:

1. Did developer activity and event types change during the pandemic?
2. Did the structure of collaboration networks shift?
3. Did the tone of developer communication change?

To answer these, I used BigQuery and Dask to collect and process the data, built weekly collaboration graphs and measured metrics like average degree, network density, clustering coefficient, etc. I also used VADER sentiment analysis on commit messages and comments to study changes in communication tone. Finally, a separate analysis was done on developers who were active in both years to see if their behavior changed.

This study helps to understand how collaboration patterns evolved during a global disruption and what that might mean for future remote work settings.

3. Data Collection and Preprocessing

The dataset used in this project was sourced from GHArchive, a public archive that records GitHub activity in near real-time. GHArchive data is hosted on Google BigQuery under the `githubarchive` public dataset, allowing efficient querying and analysis of large scale GitHub event data.

To collect and prepare the required data for this study, a series of SQL scripts were written and executed in BigQuery. All these scripts are organized and stored in the project GitHub repository under the directory:

/src/sql_scripts

1. Accessing and Filtering the Data

For this analysis, data was extracted for the months of February, March and April from both 2019 and 2020. Earlier, the idea was to filter out single contributor repositories. This would require grouping the data by repository and counting the number of unique developers per repo. While the logic is straightforward, this step can be expensive in terms of runtime and BigQuery processing costs, especially when working with large scale GitHub Archive data across multiple months. Instead, to keep the workflow efficient and manageable, I used a filter that removes events where a developer is contributing to their own repository:

actor.login != SPLIT(repo.name, '/')[OFFSET(0)]

This helped reduce self only activity and keep events where different people worked on the same repo. While this does not remove all single-contributor repos and may skip some valid collaborations where the repo owner worked with others, it is a simple and effective way to focus on cross-user collaboration. The scripts to extract the 2019 and 2020 data and to store it to a table is saved in the SQL scripts directory under the names:

- extract_feb_to_apr_2019.sql – To extract the 2019 data along with the filter.
- extract_feb_to_apr_2020.sql – To extract the 2020 data along with the filter.

2. Extracting Common Developer Subset

To explore how behavior changes for consistent contributors, we created a special dataset containing only those developers who were active in both 2019 and 2020. This was done by:

- Extracting distinct developers for each year using the same filtering logic as above.
- Using an `INTERSECT DISTINCT` operation to find developers common to both years.
- Finally, selecting all GitHub events corresponding to those developers across both years.

This logic is saved in the SQL script names 'extract_common_devs_data.sql'

3. Exporting to Google Cloud Storage

After creating the datasets in BigQuery, the ` EXPORT DATA` command was used to export the results as '.parquet' files to a Google Cloud Storage bucket. The export script is also included in the SQL scripts directory for both 2019 and 2020 datasets.

- export_feb_to_apr_2019.sql – To export 2019 data to bucket
- export_feb_to_apr_2020.sql – To export 2020 data to bucket
- export_common_devs_data.sql – To export common developers data from both 2019 and 2020 to bucket.

4. Loading Data into Colab

From the GCS bucket, data was loaded into the Colab environment using the 'gcsfs' python library. This enables seemless integration with Dask fro scalable data processing. The data loading and saving code is available in the notebook:

data_download_and_loading.ipynb

4. **Collaboration Trend Analysis**

To address Research Question 1, which investigates how developer collaboration patterns changes before and during the COVID-19 pandemic, we conducted two key types of analysis using GitHub Archive event data. The code for this analysis is located in /src/collaboration_trends.ipynb. Columns and Event types used –

For this analysis, we relied on the following key columns extracted from BigQuery:

- Created_at : Timestamp to group events weekly.
- type : Event type such as 'PushEvent', 'PullRequestEvent', etc.
- actor : Developer login used to count unique active contributors per week.

We focused on core collaboration event types that reflect contribution and discussion activity:

'PushEvent', 'PullRequestEvent', 'IssueCommentEvent', 'IssuesEvent' and 'PullRequestReviewCommentEvent'.

A. Event Type Frequency Trends

We computed the weekly normalized counts of each event type to track collaboration volume and volatility, and then calculated the week by week percent changes. This helps observe dynamic shifts in developer behavior, particularly comparing 2019 (pre-COVID) and 2020(early COVID).

For each event type, we plotted trends for 2019 vs 2020 and included statistical testing (Mann-Whiteny U test) to validate whether the differences were statistically significant. The These are plots for core event types:

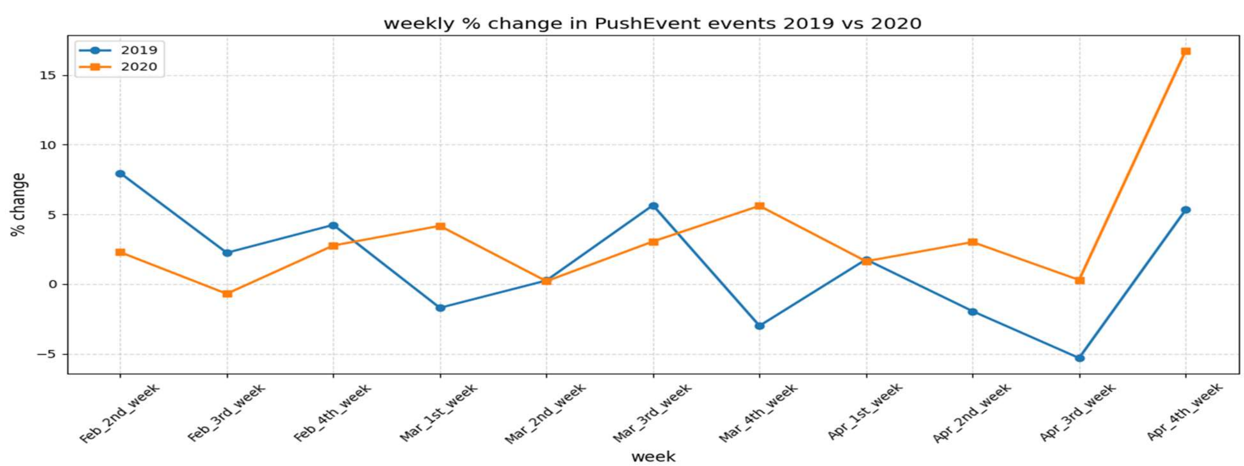

Fig. 1.1 Weekly Percent changes in PushEvent event 2019 vs 2020

In 2019, Push event activity remained relatively stable, with weekly percentage changes mostly within -5% to +5%, indicating consistent development workflows. In contrast, 2020 displayed more variability, moderate increases through February and March culminated in a sharp 16.6% spike during the last week of April. This surge may reflect intensified development efforts as teams adjusted to remote work during the early stages of the COVID-19 pandemic. The pattern suggests that while workflows were disrupted, many developers remained productive and potentially accelerated contributions in response to the new working environment.

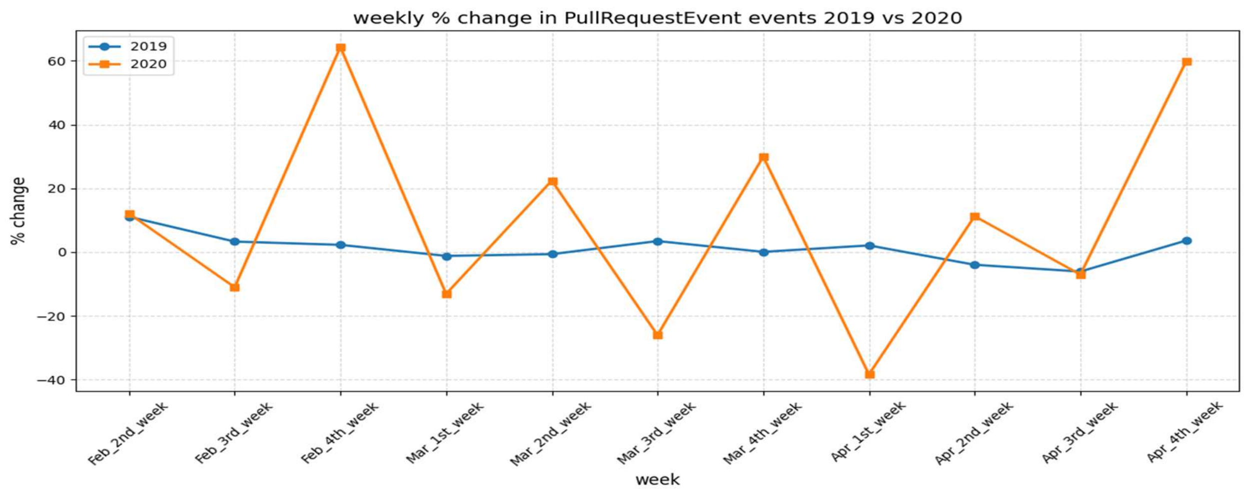

Fig 1.2 Weekly Percent changes in PullRequestEvent event 2019 vs 2020

In 2019, pull request activity stayed mostly stable with week to week changes under -5% to +5%. But in 2020, the pattern was much more erratic, a sharp 64% spike in late February was followed by drops of 15% then another 30% rise and even a steep dip of 39% in early April and also a 13% drop in early March, which may reflect short term disruptions or shifting priorities as teams adapted to COVID workflows. These swings suggest that during COVID, collaborative code review and merging became more unpredictable, possibly due to changing project timelines or team adjustments to remote workflows.

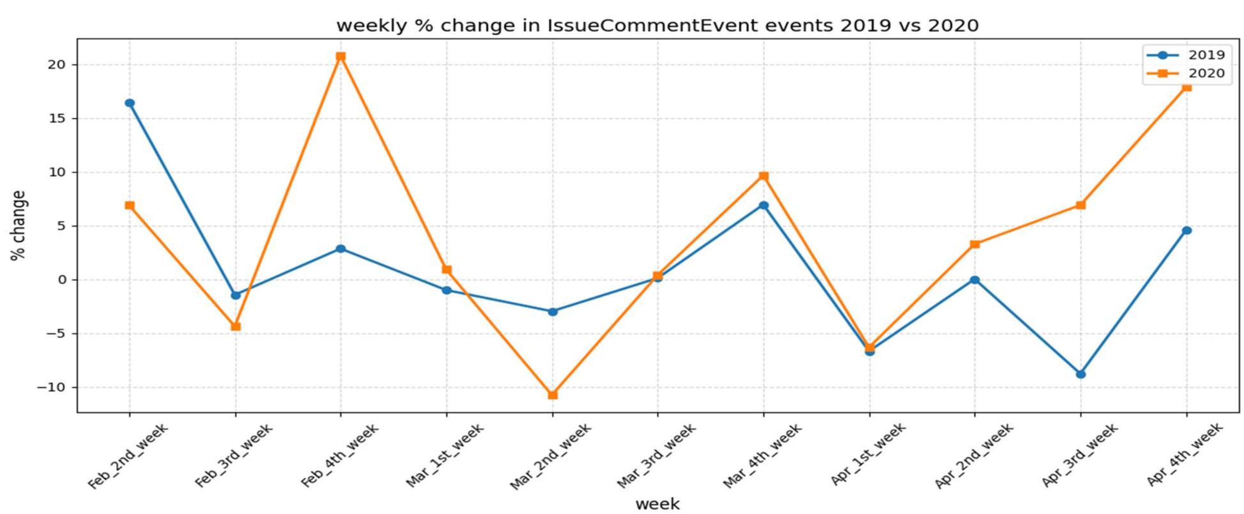

Fig 1.3 Weekly Percent changes in IssueCommentEvent event 2019 vs 2020

In 2019, issue comment activity stayed mostly within +5% to -5%, showing steady discussion patterns. In 2020, the trend was more dynamic, with a 21% spike in late February, a 11% drop in mid March which may align with early pandemic adjustments and a steady rise again through April ending at 18%. This suggests that while developer discussions dipped briefly during the early COVID disruption, they quickly picked back up as teams adapted to remote workflows.

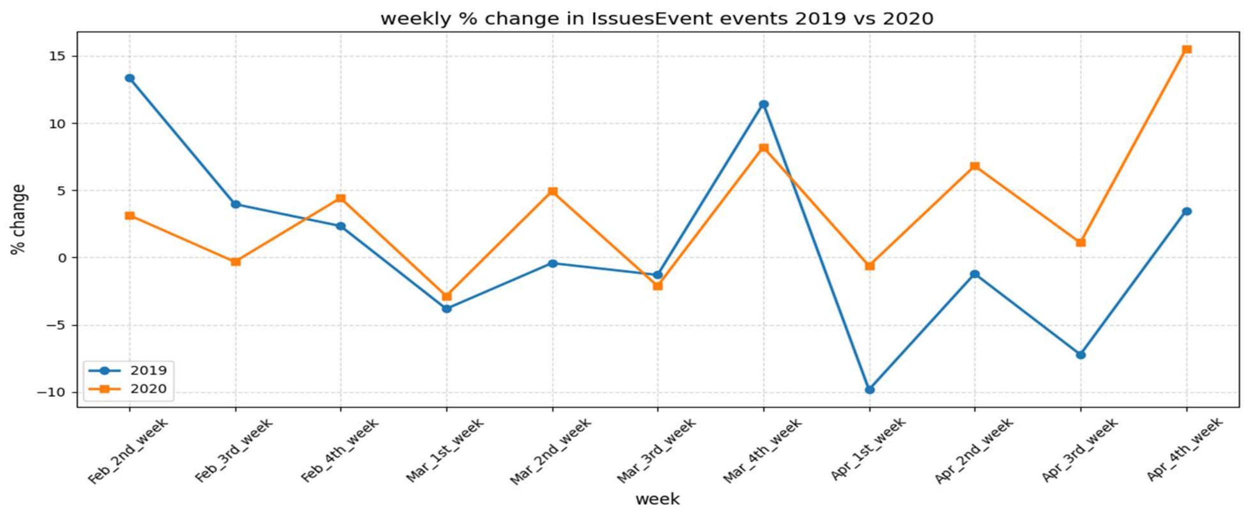

Fig 1.4 Weekly Percent changes in IssuesEvent event 2019 vs 2020

In 2019, issue activity fluctuated mostly within +10 to -10%, with occassional spikes like a 13% spike in late March followed by a sharp dip of 10% in early April, possibly either a sudden change in project focus or possible data noise. In 2020, the changes were more balanced and showed steady growth through April, ending with a 15% increase in the final week. This suggests that while 2020 started off more cautiously, issue activity steadily picked up as teams adapted to remote collaboration.

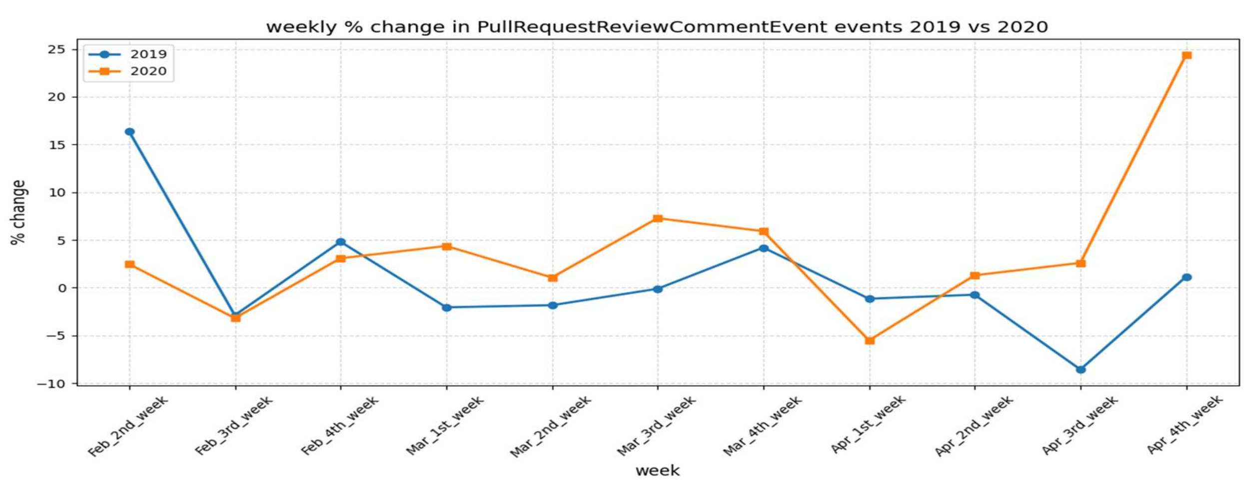

Fig 1.5 Weekly Percent changes in PullRequestReviewCommentEvent event 2019 vs 2020

In 2019, review comment activity stayed mostly between -5% to +5%, showing stable collaboration through code review. In contrast, 2020 showed a sharped upward trend, culminating in a notable 24% spike during the final week of April. This sharp rise could point to increased emphasis on peer review toward the end of the quarter or a final push in collaborative development. It may also reflect adaptation to remote workflows or a shift in project coordination styles during the pandemic response period.

We used the Mann-Whitney U test to determine whether the weekly event count distributions in 2019 and 2020 differ significantly for each event type. Results showed statistically significant differences ($p < 0.05$) for most collaboration events like PushEvent, PullRequestEvent and IssueCommentEvent, confirming that the observed changes were not due to random fluctuations, but meaningful shifts in developer behavior during the pandemic.

B. Active Developer Trends

We calculated the number of unique developers (actors) participating each week to understand participation levels:

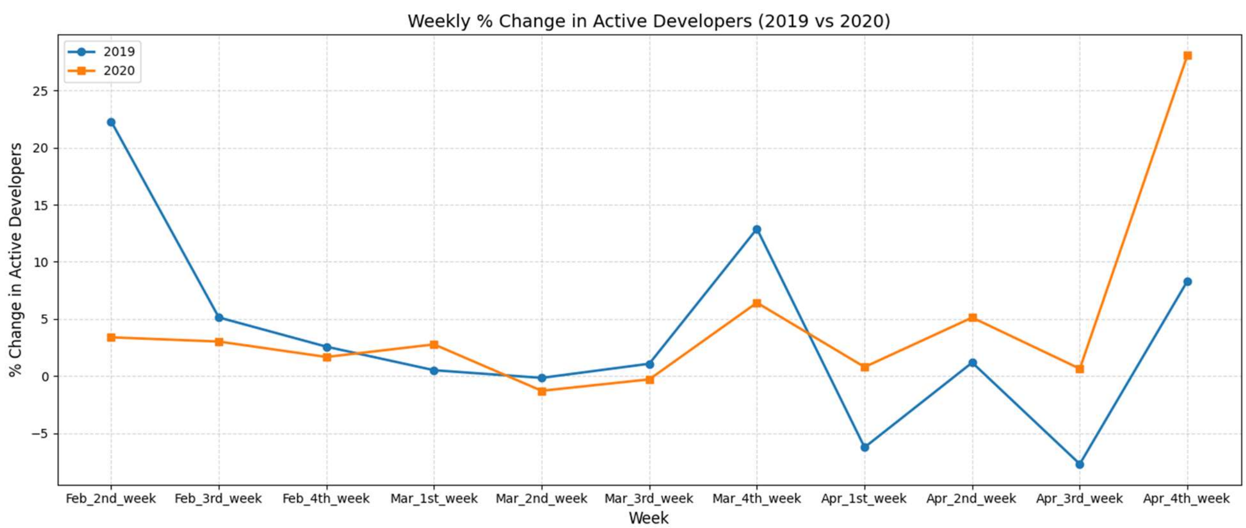

Fig 1.6 Weekly Percent change in Active developer 2019 vs 2020

In 2019, active developer counts showed more fluctuation, with a 22% increase in early February, steady growth through March, followed by sharp drops in April. In 2020, the trend was more stable early on, but ended with a remarkable 28% spike in the last week of April, likely reflecting increased contributions during global lockdowns. Overall, the 2020 trend appears more stable but ramps up later, whereas 2019 features stronger early swings. These patterns may indicate shifting collaboration dynamics pre and post COVID with 2020 reflecting more delayed but intense engagement, possibly driven by remote workflows, virtual sprints or external motivators like open source contributions.

Again, we used the Mann-Whitney U test to compare the distributions of weekly active developer counts. The result – U = 39.0, p = 0.0362, indicated a significant difference between 2019 and 2020 participation trends, providing further evidence that developer engagement shifted due to external factors like COVID-19.

Summary of Collaboration Trend Analysis

To investigate how developer collaboration patterns changed during the early COVID-19 period, we analyzed GitHub event data from February to April for both 2019 (pre-COVID) and 2020 (early COVID). We focused on key collaboration-related event types such as PushEvent, PullRequestEvent, IssueCommentEvent, IssuesEvent, and PullRequestReviewCommentEvent, computing the weekly percentage change for each.

The trends revealed that:

- 2019 displayed relatively stable week-to-week changes across event types, suggesting consistent collaboration behaviors.
- 2020 showed more fluctuation, with several event types exhibiting sharp spikes or drops, especially in late February and April, indicating possible disruption and adaptation in workflows due to the pandemic.

Notably, while there was initial instability in collaboration metrics during early March 2020, activity picked up significantly by April's end in several events (e.g., PushEvents and ReviewComments), possibly reflecting developers settling into remote workflows.

The analysis tables are located in the GitHub repository at /analysis/Analysis_Tables as follows:

event_counts_2019.csv – Contains event type counts per week for 2019 data

event_counts_2020.csv - Contains event type counts per week for 2020 data

weekly_event_pct_changes.csv – Contains weekly event type percent changes in 2019 and 2020.

unique_devs_per_week_pct_changes.csv – Contains weekly active developer count percent changes.

Conclusion

The analysis reveals clear shifts in developer collaboration trends between 2019 and 2020. While 2019 exhibited steadier, more predictable patterns, 2020 showed higher volatility and delayed peaks, particularly in April reflecting the impact of the pandemic. These trends suggest that developers continued to engage actively on GitHub, but with altered rhythms and bursts of activity in 2020 compared to the more stable patterns seen in 2019. These findings support hypothesis of research question 1, confirming that the onset of COVID-19 influenced the volume and timing of collaborative activity, even as overall participation levels remained strong.

5. Network Analysis

To address research question 2 which is 'Did the structure of developer collaboration networks change significantly due to the pandemic?', we constructed weekly collaboration graphs for both 2019 and 2020 using GitHub event data. The code for this analysis can be located at /src/network_analysis.ipynb. Each graph represents developers as nodes and collaborations between them as edges.

Metrics and Methodology

The following key network metrics were computed for each weekly graph to analyze structural differences:

- Number of Nodes: Unique developers participating in collaborations.
- Number of Edges: Total collaborative interactions between developers(nodes).
- Average Degree: Average number of collaborations per developer.
- Network Density: Proportion of actual connections to all possible connections.
- Clustering Coefficient: Likelihood that collaborators of a developer also collaborate with each other.
- Number fo Connected Components: Measure fragmentation.
- Largest Connected Component Size: Number of developers in the biggest collaboration cluster.

Metrics Trends 2019 vs 2020:

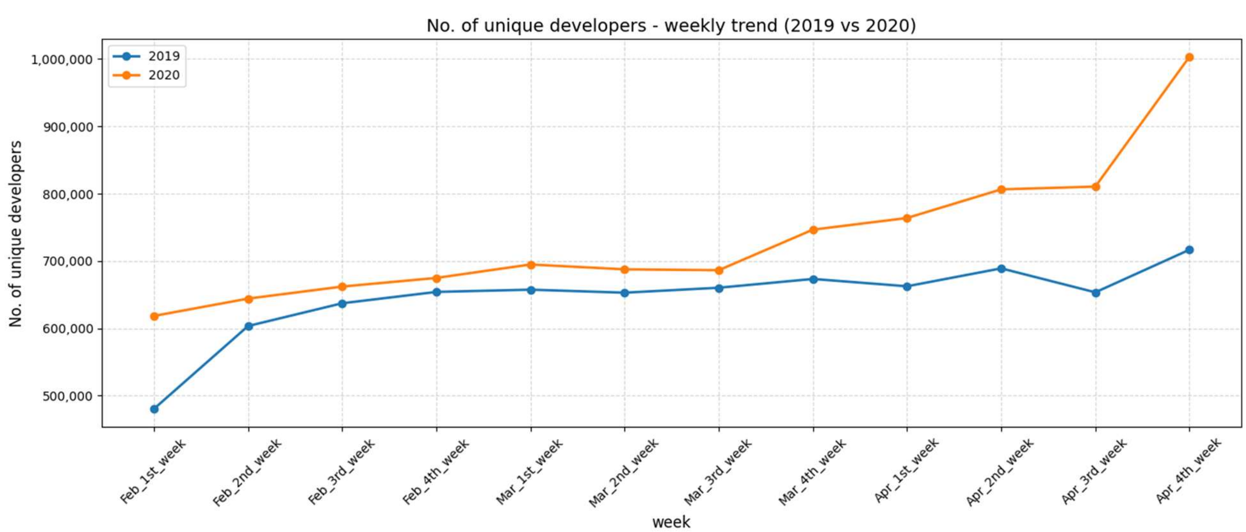

Fig 2.1 Weekly Number of Unique developers (Number of nodes) 2019 vs 2020

Number of nodes was higher in 2020, especially in April, which shows increased participation in 2020 compared to 2019.

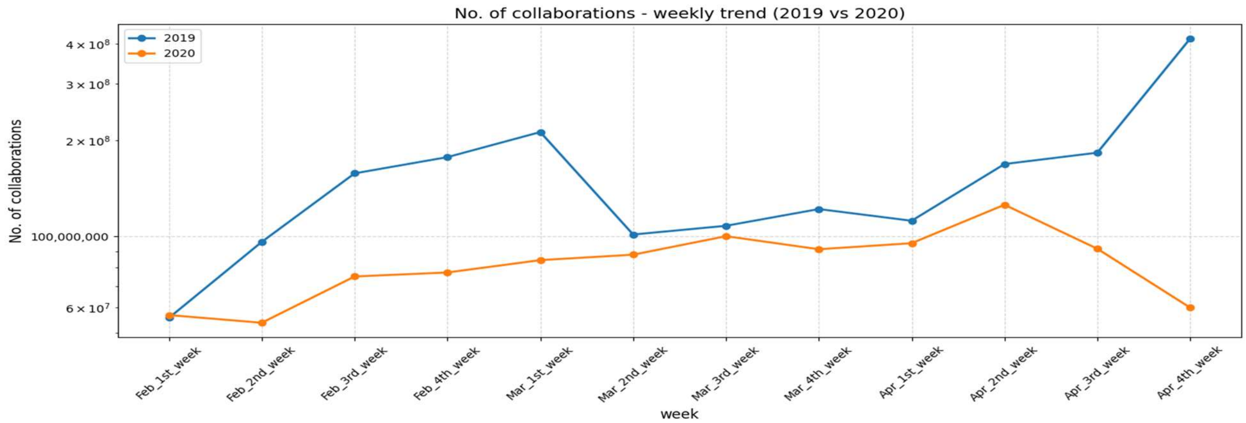

Fig 2.2 Weekly Number of Collaboration(Number of Edges) 2019 vs 2020

Number of edges was consistently higher in 2019 compared to 2020, which suggests that there was more intense or larger scale interactions in 2019, despite of fewer participants to 2020.

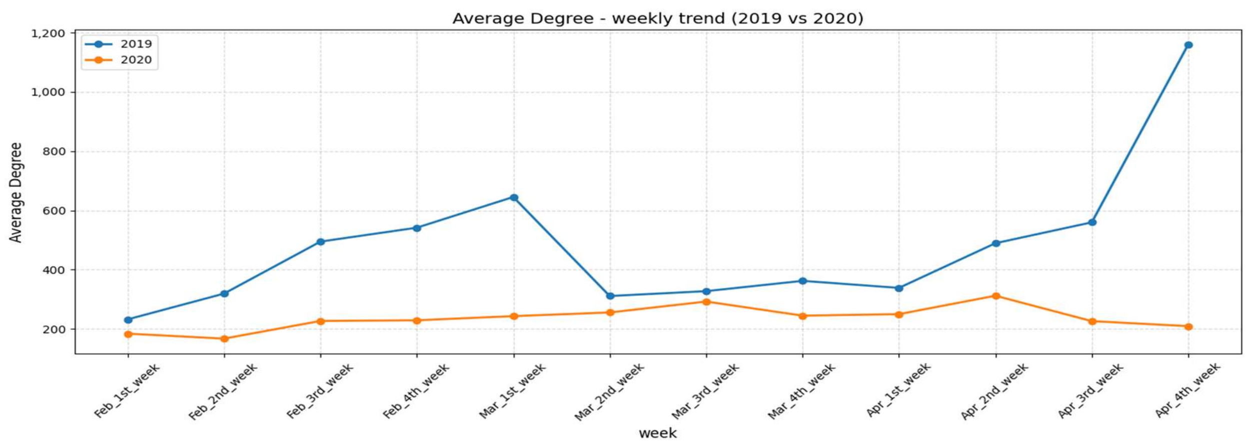

Fig 2.3 Weekly Average Degree 2019 vs 2020

Average Degree was significantly higher in 2019, compared to 2020, indicating that developers were, on average, collaborating with more peers each week.

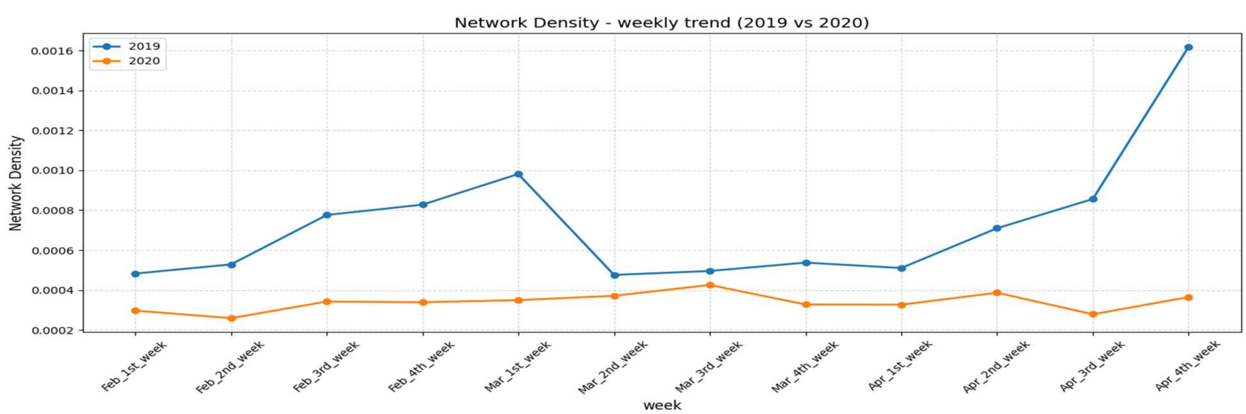

Fig 2.4 Weekly Network Density 2019 vs 2020

Network Density also showed higher values in 2019 and similar trend in comparison to 2020, which suggests denser, more tightly connected collaboration network. In contrast, 2020's density values were consistently lower and showed minimal fluctations, suggesting a more dispersed network structure with fewer connections relative to the number of developers.

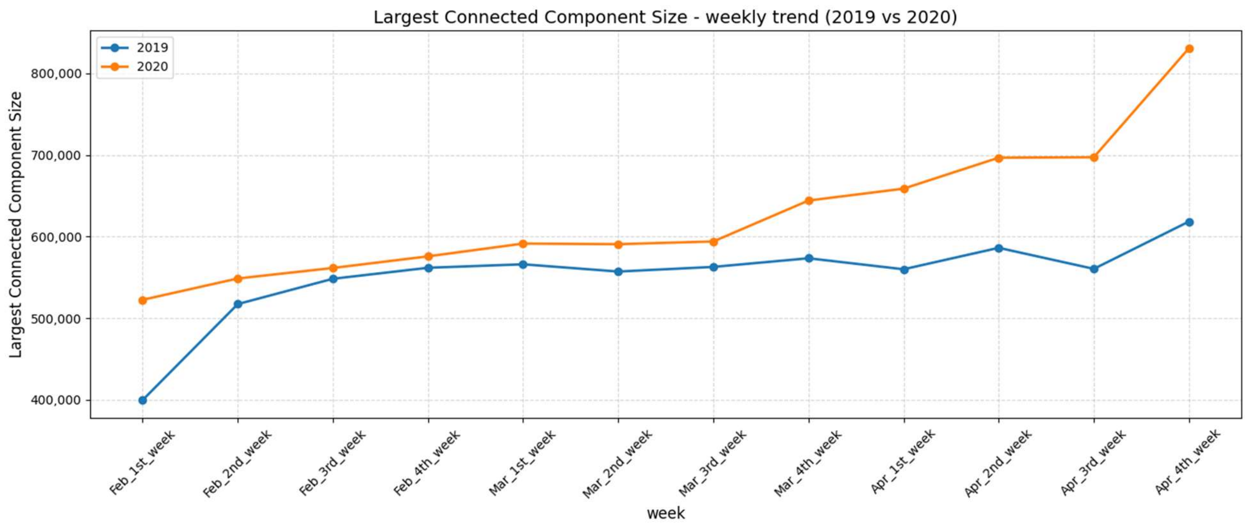

Fig 2.5 Weekly Largest Connected Component Size 2019 vs 2020

LCC size was higher in 2020, showing that more developers were reachable within a single cluster, but it does not imply stronger collaboration, rather that the network had more passive or shallow links.

These trends indicate that while more developers participated in 2020, the structure of collaboration shifted toward smaller or more distributed interactions, possible due to the disruption and transition to remote work.

The network metrics for both 2019 and 2020 can be found at this location on the GitHub repository /analysis/Analysis_Tables/final_network_metrics.csv

Metric Fluctuations

Key fluctuations can be seen in Average Degree and Network Density metrics between February 3rd week to March 2nd week of 2019. By analyzing edge and node counts, it is concluded that the spoke was primarily due to a sharp increase in edges (collaborations), even as node growth remained stable. This caused each developer to have more connections on average. The dip between March 1st week to March 2nd week is also due to the change in the number of edges in that period. This rise in collaborations may have been influenced by short term community events, synchronized feature development cycles, or increased team activity on key repositories during that period.

Statistical Significance

Mann-Whitney U test is performed for the network metrics to compare each metrics between 2019 and 2020. This test confirmed statistically significant differences ($p < 0.05$) across all the metrics, which validate that the differences in network structure are not due to random fluctuations but reflect actual behavioral changes between the 2 years.

Sample Network Visualization

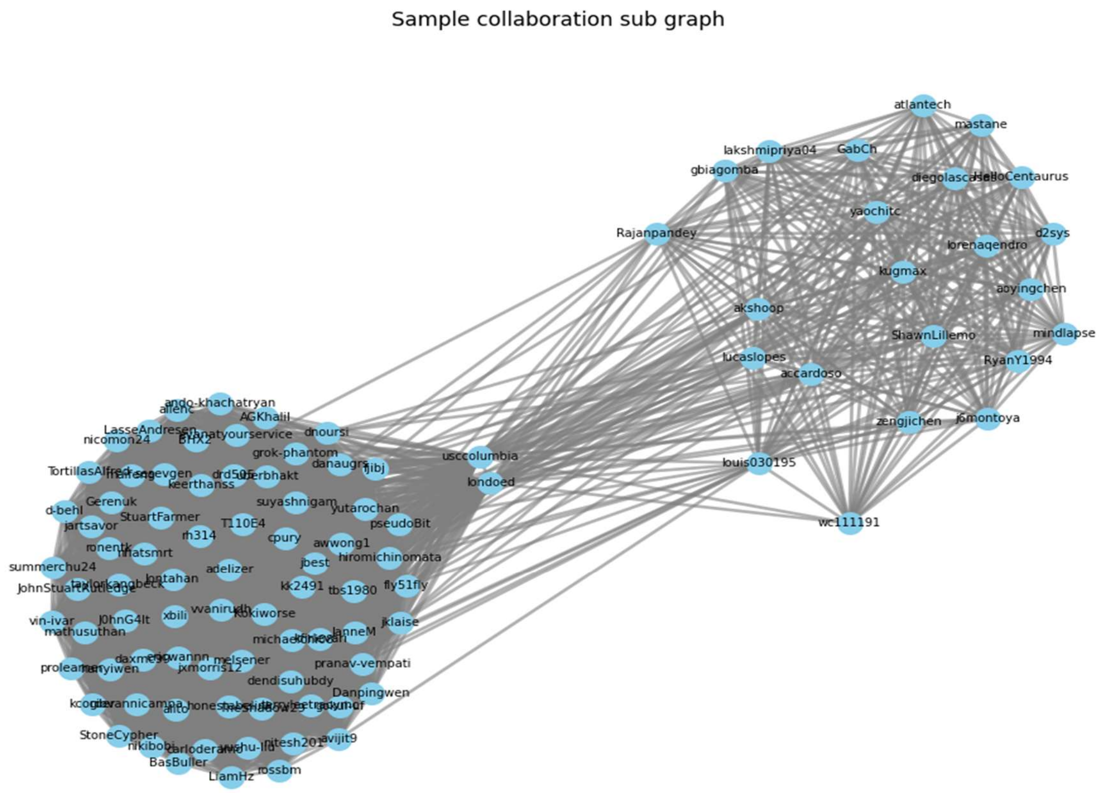

Fig 2.6 Sample Network Visualization of a Sub graph with 100 connected nodes

To complement the quantitative metrics, a sample network visualizations is generated which illustrates the structure of developer collaboration within a specific week. Each node in the graph represents a unique developer, and each edge represents a collaboration (i.e. interaction on a shared repository during that week).

The visualization above reveals how developers cluster into distinct communities, with dense intra-group connections and fewer links between groups. These clusters form because developers in each group contributed to the same repository or set of repositories during the same week. As a result, the structure reflects shared project participation rather than predefined teams or organizations.

These clusters likely represent developers working on different kinds of projects, like frontend vs backend, or different domains like machine learning or web apps. They might also be from different organizations working on the same open source platform. So, these clusters likely reflect groups of developers who share common project characteristics, such as similar types of applications, specific domains or contributions tied to particular organizations or teams.

In this example, two tightly knit clusters can be observed, with a few developers acting as bridges between them. These bridging developers likely contributed to multiple repositories that were touched by members of both groups, effectively connecting otherwise we would have seen separate communities. This kind of structure supports what we see in the metrics like average degree and clustering coefficient, where closely connected groups of developers are a sign of strong collaboration.

This visualization serve to humanize the metrics, offering an intuitive understanding of how collaboration was organized and how centralized or fragmented the network might have been at any given point.

Conclusion

The network analysis revealed that developer collaboration networks were significantly more interconnected and cohesive in 2019 compared to 2020. Metrics such as Average Degree, Density and Clustering Coefficient were consistently higher in 2019, while 2020 showed flatter trends and lower values, indicating more fragmented collaboration patterns during the early COVID period. Despite a higher number of active developers in isolated collaborative groups. These findings support the hypothesis of research question 2, confirming that the structure and intensity of developer networks changes meaningfully during the pandemic, shifting from large, tightly knit communities to more distributed and loosely connected patterns of collaboration.

6. Sentiment Analysis

To explore the potential behavior or emotional shifts among developers during the early COVID period, we performed sentiment analysis on textual data derived from the GitHub events. Specifically, the text was extracted from commit messages from PushEvent and comment texts from IssueCommentEvent and PullRequestReviewCommentEvent across February to April in both 2019 and 2020. This text data was collected from the GitHub Archive and parsed using the Payload column for each event. The code for this analysis can be found at /src/sentiment_analysis.ipynb.

We used the VADER sentiment analysis tool, a lightweight lexicon based model commonly used to social media and short text, to assign sentiment scores. While VADER is efficient and easy to implement, it is not optimized for highly technical language, such as developer commit messages or issue discussions, which often include domain specific jargon and last of emotional expressions.

As a result, the sentiment trends across weeks and years showed only mild variations with no strong or statistically significant patterns. These outcomes highlight the limitations of applying general purpose sentiment tools to technical texts like GitHub texts. Thus, while the analysis aimed to address research question 3, which is ‘Did the sentiment of developer interactions shift during the pademic?’, the results were largely inconclusive and suggest the need for more specialized NLP models or manual annotation for future sentiment studies in developer communication.

7. Analysis on Common Developers from Both 2019 and 2020

To deepen our understanding of developer behavior shifts during the early COVID-19 period, analysis was conducted with a focus on common developer, those who were active in both 2019 and 2020 during the months of February to April. This aimed to isolate changes in collaboration patterns among a consistent group of contributors, thereby reducing the noise introduced by different user populations each year. The code for this analysis can be located at /src/common_devs.ipynb.

The dataset was contructed using a two-step filtering process in BigQuery. First, the distinct developers in 2019 and 2020 were indentified who contributed to repositories they did not own. The, 'INTERSECT DISTINCT' operation was used to extract developers active in both years. This filtered list was then used to collect their corresponding GitHub Archive events for both years, resulting in a unified dataset that exclusively captured events from common contributors.

The same collaboration trend analysis and network analysis were used to this dataset. The goal was to compare whether the trends observed in the full dataset persisted when focusing only on overlapping developers.

Conclusion

The trends identified in the original datasets were reaffirmed in this common developer analysis. Despite having a smaller, consistent developer base, it was observed that there is similar shifts in collaboration volume, network structure and connectivity patterns, most notably, a reduction in average degree and network density in 2020. This consistency further supports the hypothesis that that COVID-19 impacted how developers collaborated, not just who was contributing. The alignment of findings across both full dataset and filtered dataset strengthens the robustness of our conclusion for research questions.

The analysis metrics and tables can be found under /analysis/Analysis_Table folder at GitHub repository of the project as follows:

final_com_network_metrics.csv – Contrains network metrics for the common developers data from 2019 and 2020.

weekly_com_event_changes.csv – Contains weekly event type counts for common developers data from 2019 and 2020.

8. Conclusion

This study explored the evolution of developer collaboration patterns during the early stages of COVID-19 pandemic, using GitHub Archive data from February to April for the years 2019 and 2020. Through both event based collaboration trends and structural network analysis, it was aimed to answer three key research questions related to changes in collaboration behavior, network cohesion and developer sentiment.

For research question 1, we analyzed trends in event types such as PushEvents, PullRequestEvents and Issue related activities. While developer activity remained steady and even increased in 2020, especially in late April, the week to week trends became more erratic. This suggests a shift in how developers collaborated, with more frequent fluctuations likely influenced by external disruptions like remote work transitions. Although the collaboration volume did not show statistically significant changes across all event types, the visual trends support the idea that workflows were affected during the period.

In research question 2, network analysis revealed clear structural differences between the 2 years. Metrics such as average degree, network density and clustering coefficient were consistently higher in 2019, indicating tighter, more interconnected collaboration networks. In contrast, 2020 showed flatter trends and lower values, suggesting a more fragmented and distributed collaboration environment. The Mann-Whitney U tests confirmed that these differences were statistically significant, reinforcing that the shifts in network structure were not random but reflected real behavioral changes during the pandemic.

In research question 3, sentiment analysis was done using VADER to access changes in tone withing developer generated text. While VADER is generally effective for short, social media text, its applicability to technical GitHub content proved to be limited. As a result, the sentiment trends were inconclusive and did not yield strong evidence of emotional or behavioral shifts.

To validate the robustness of our findings, we also performed a focused analysis on common developers, those active in both years. This subset mirrored the trends observed in the full datasets, strengthening our conclusion that the observed changes were not due to varying user base but represented genuine changes in collaboration dynamics.

Overall, the findings confirm that while developer participation remained high during the early COVID-19, the structure and rhythm of collaboration evolved. Developer networks in 2020 were less cohesive and workflows appeared more fragmented, aligning with our hypothesis that the pandemic had a meaningful impact on how open source contributors interacted.

9. Limitations and Future Work

While the project provides valuable insight into how developer collaboration patterns changed during the early stages of COVID-19 pandemic, there are a few limitations to consider:

- Filtering for collaboration: The filtering method used removes self-contributions but does not guarantee the exclusion of all single contributor repositories. A more accurate approach would involve counting unique contributors per repo, which was not implemented due to potential BigQuery cost and runtime concerns.
- Temporal scope: The analysis focused on a 3 month period in both 2019 and 2020. This limited window was chosen due to the large size of the GitHub Archive data and the additional time and resource requirements to process and optimize each step of the workflow. Expanding the timeframe in future work could help identify more gradual or long term changes in collaboration behavior.
- Limited network metrics: Only a core set of network metrics were calculated due to the time it takes to compute, even on samples subgraphs. In future work, more efficient graph processing tools or distributed libraries could be used to compute a wider range of metrics at scale.
- Small scale network graph visualization: Only a small part of the network was visualized due to the large size of full graphs, which were too heavy to render. Scalable tools like Gephi or GPU based rendering can help visualize complete networks or reveal broader patterns in future.
- Sentiment Analysis limitations: VADER is easy to use and works well for short, social media text, but it is not well suited for technical language like commit messages or developer comments. Because of that, the sentiment results may not clearly show real emotional shifts. In future work, more advanced models which work better on technical text could give better results.
- Same username used to find common devs: Common developers were matched using usernames, which may miss users with account changes or multiple identities. Other identifiers like email IDs or commit metadata (if available) could be explored to improve matching across years.

Future Work

- Bots and automation detection: Some GitHub activity may come from bots or automated accounts, which can distort collaboration patterns. Future work could involve detecting and filtering out bot users to focus the analysis on human developer behavior.
- Role based analysis: Developers play different roles in projects, such as maintainers, frequent contributors or newcomers. Identifying these roles can help reveal how collaboration varies across experience levels or responsibilities.

- Explore organizational collaboration patterns: Analyzing collaboration on repositories owned by organizations could help reveal differences between internal team workflows and open source community contributions. In the future, deeper analysis of organization linked repos may uncover how coordination varies across different settings.
- Predicting future collaboration patterns using graph neural networks: In the future, graph based models could be used to predict how developer collaboration networks change over time. Using techniques like link prediction with graph neural networks(GNNs), we can try to forecast new connections between developers. This could help us understand how collaboration patterns evolve after events like the COVID-19 pandemic.

10. References

1. Google BigQuery Documentation

https://cloud.google.com/bigquery/docs

2. GitHub Archive (Data source)

https://www.gharchive.org/

3. Simko, L., Huang, Y., Vasilescu, B., & DeBlasio, D. (2023). *How Did Developers Collaborate on GitHub During the Pandemic?*

https://arxiv.org/pdf/2301.12326

4. SciPy stats (Mann-Whitney U test)
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

5. Dask documentation

https://docs.dask.org/en/stable/

6. NetworkX official documentation
https://networkx.org/documentation/latest/auto_examples/drawing/plot_degree.html#sphx-glr-auto-examples-drawing-plot-degree-py

7. NetworkX GitHub repository

https://github.com/networkx/networkx

8. Intro to Network Analysis (Blog)

https://trenton3983.github.io/posts/intro-network-analysis/

9. VADER sentiment analysis tool - Hutto, C.J. & Gilbert, E.E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.*
https://github.com/cjhutto/vaderSentiment