Design Lab

Report

____

Submitted by

Anamitra Mukhopadhyay

20CS30064

Task Description

Multiple authors collaborate in large codebases. We ask the question – is it the same author who introduces a bug, fixes the bug or is it another author? And how long does it take for the bug to be fixed?

Data

Top 100 repositories were picked up from Github-Ranking/Top100 in April 2025. Among them we choose 5 languages (top 5 according to this).

Methodology

Finding Previous Author:

- Get the git blame hunks of a file
- For each hunk, iterate the git log for the file until there is a hit in the hunk line numbers

Labelling Data:

- Keyword based labelling of commit messages as preliminary labelling
- LLM based labelling

Challenges

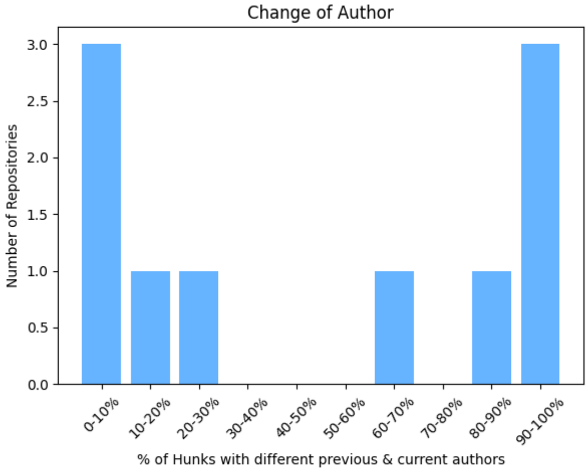
- LLM based labelling is not possible for a large dataset because free API run out of trials

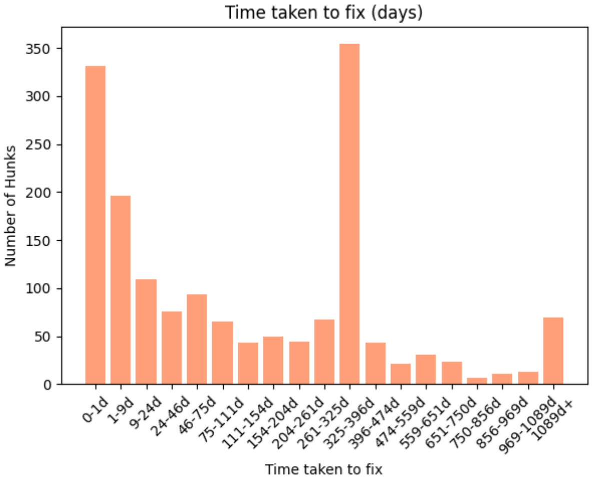
- The previous author finding is a parallelized implementation, even so, it takes significant time to run for large repositories

We are considering small repositories for the analysis as of now.

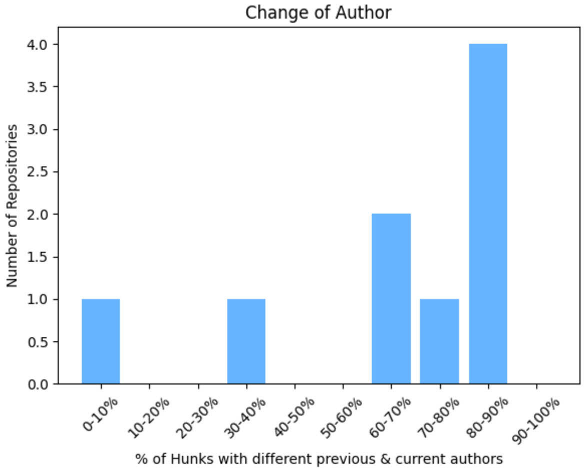
Results

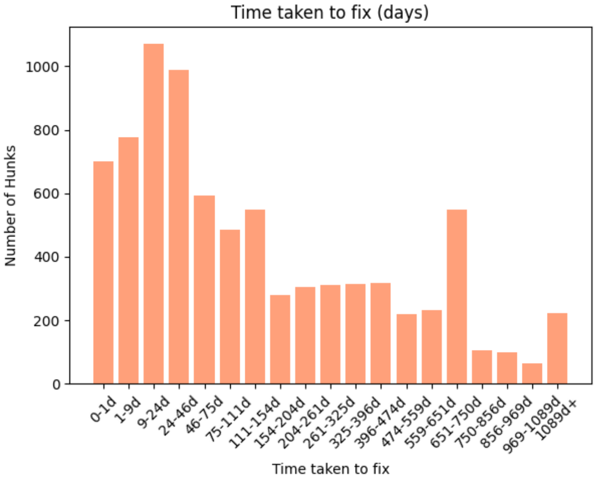
For Python


Change of Author

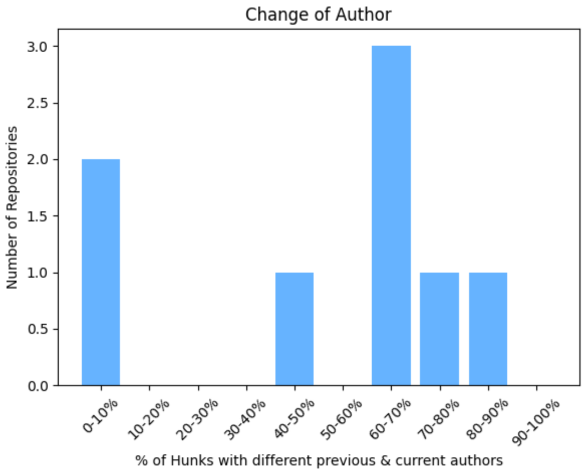

Time taken to fix (days)

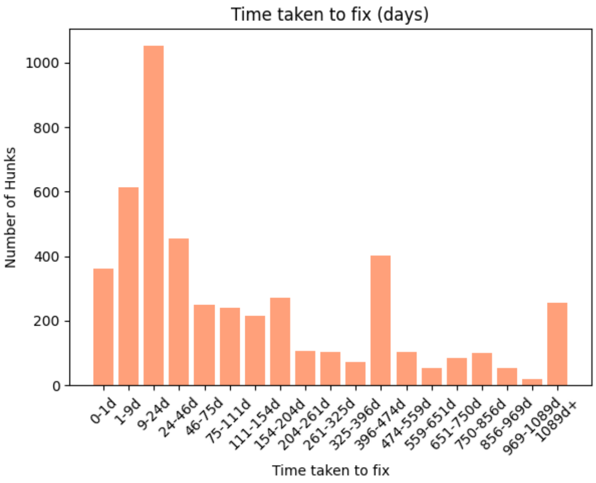
For Go


Change of Author

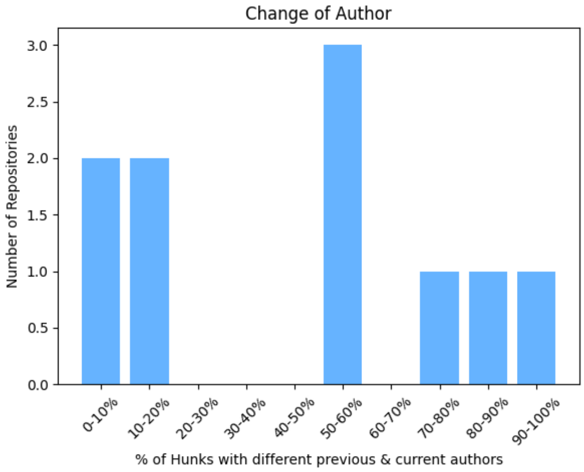

Time taken to fix (days)

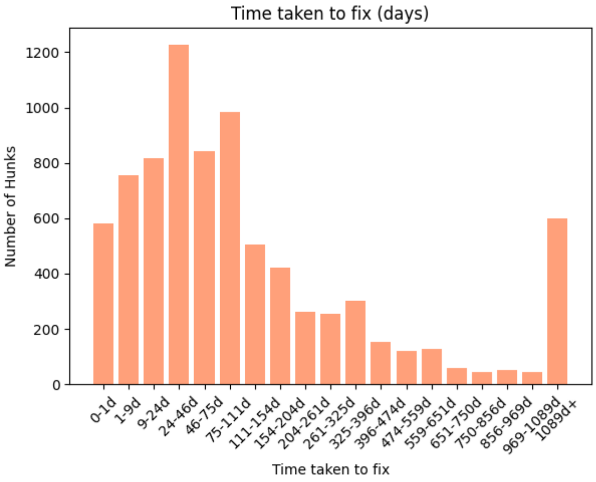
For JavaScript


Change of Author

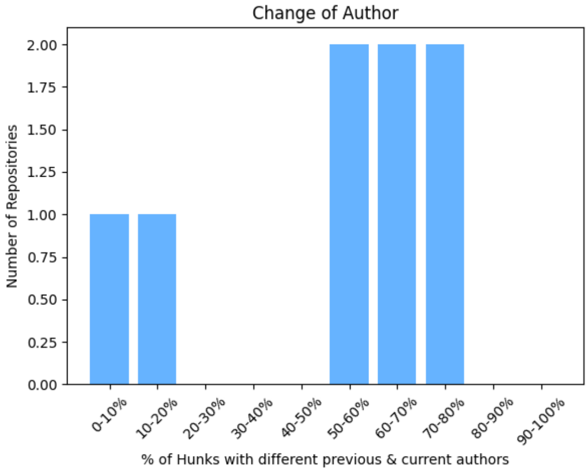

Time taken to fix (days)

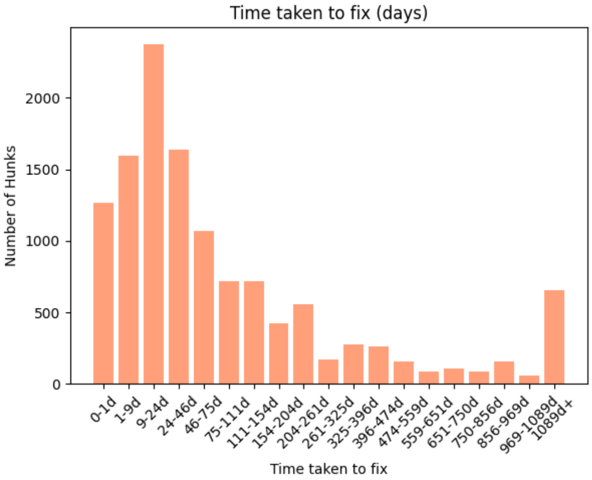
For Java


Change of Author


Time taken to fix (days)

For CPP


Change of Author


Time taken to fix (days)

Note: Due to long processing time, 10 repositories from each language has been used to infer. Also, in a lot of repositories the commit messages are in some language other than English (mostly Chinese). Those repositories have been avoided purposefully.