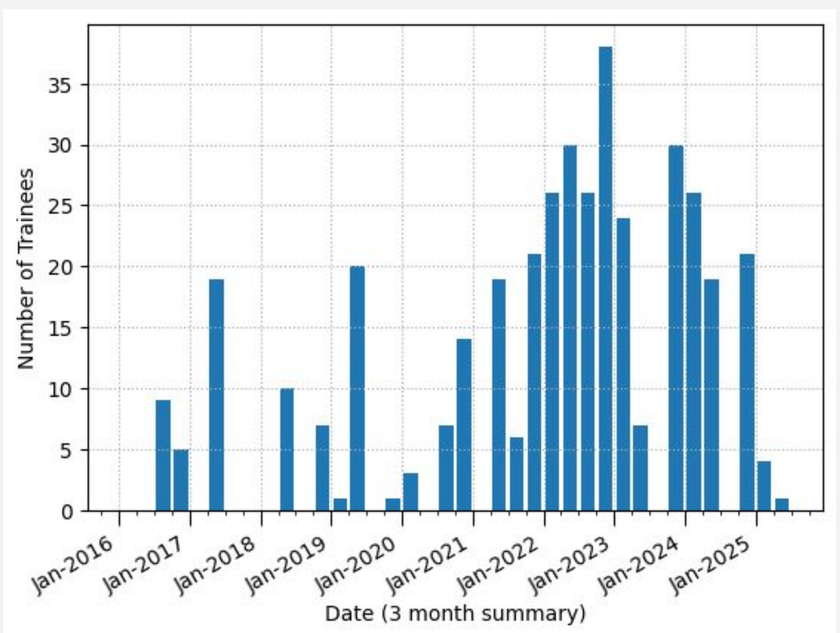
Leveraging GitHub API Data to Evaluate Git and GitHub Training Outcomes

Daniel Brady & Anna Krystalli

daniel.brady@sheffield.ac.uk



Git/Github training run by Sheffield RSE team

- ➔ Started in 2016, and run at least once a year
- ➔ Run over two half-days
- ➔ Teaches the basics of Git and Github using GitKraken
- ➔ Includes exercise where trainees fork an existing repository, add their own changes, and open a pull request

Can we assess how many trainees go on to use Git/Github post-training by looking at their Github activity?

- ➔ How many trainees have limited usage of Git/Github prior to the training?
- ➔ How many trainees show GIthub activity within 1 year post training?
- ➔ Among trainees with prior Github activity, are there changes in this activity 1 year pre- and post- training?

For each of the 527 pull requests:

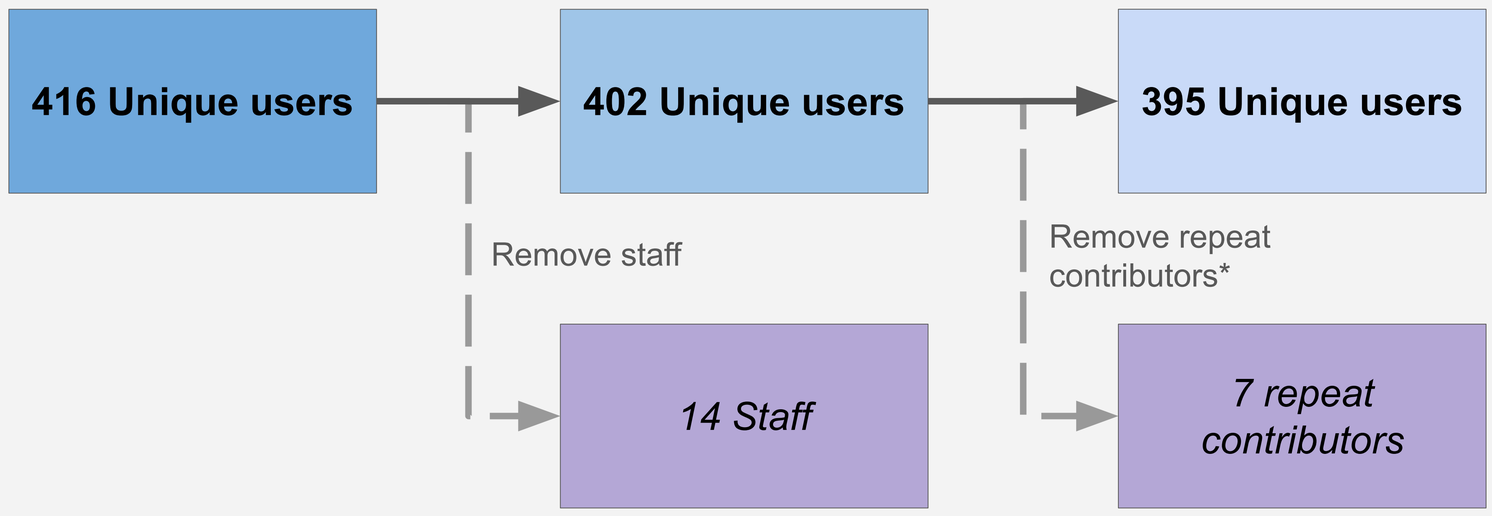
- ➔ Get the user and creation date
- ➔ Filter to get a list of trainees

For each trainees account:

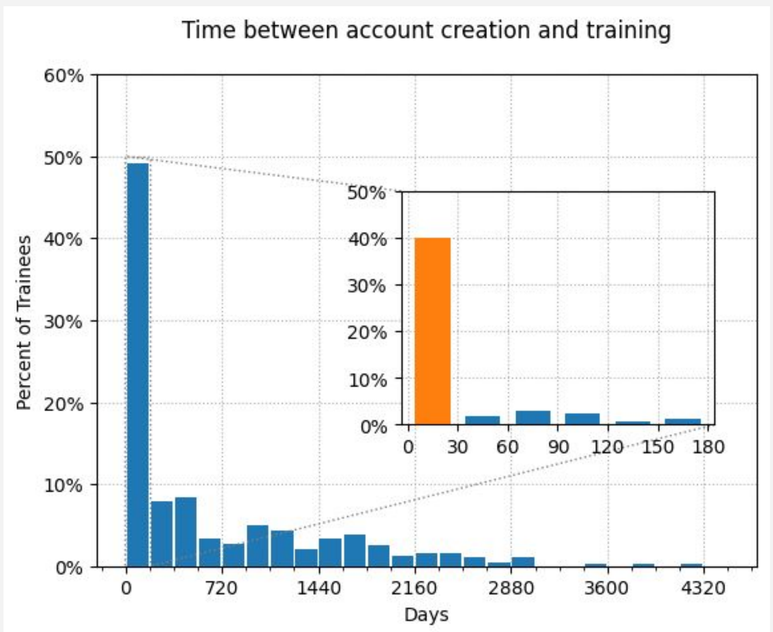
- ➔ Get the account creation date relative to training

Then for each public repository on the account get:

- ➔ First commit of the repository relative to the training
- ➔ Total number of commits for the repository
- ➔ The number of days between the oldest and newest commit



*Users with multiple PRs, on different days



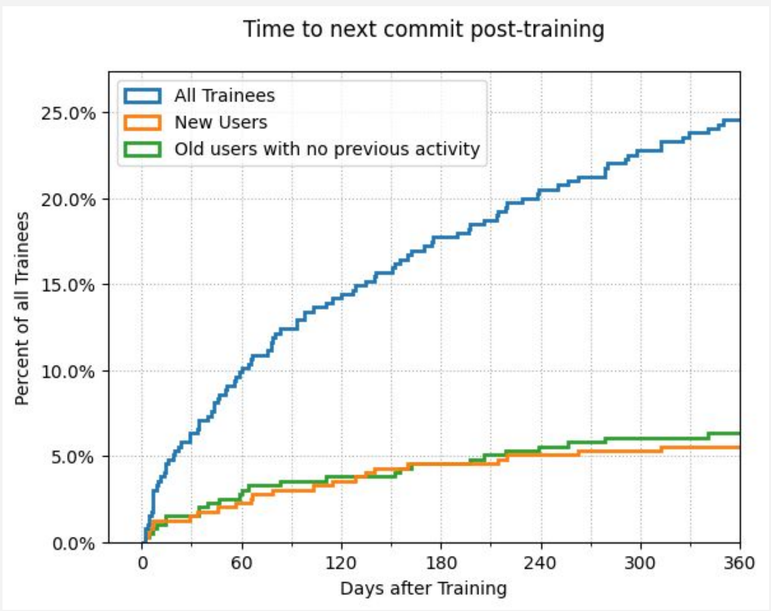
40% (158) of trainees create their Github account within a month of the training (new users). Another ~34% (133) had no activity on their account pre-training.

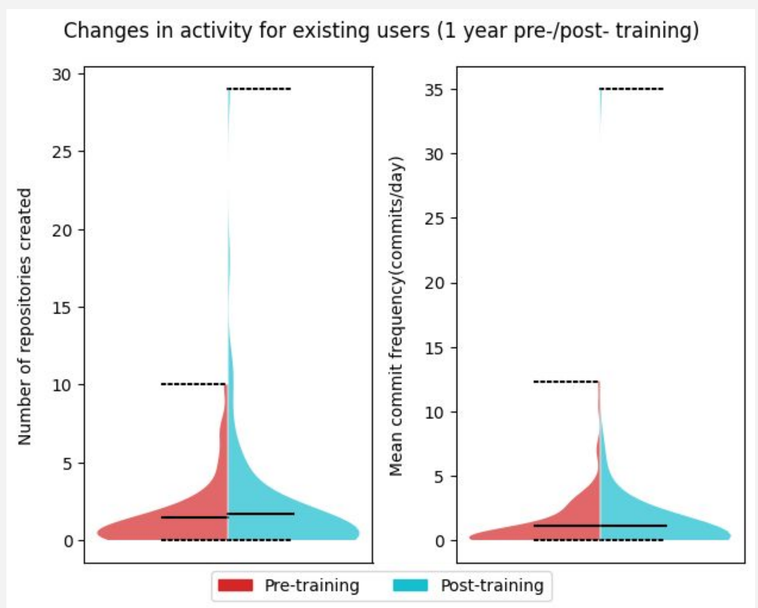
In total ~74% (291) of trainees had limited prior usage of GitHub

Within a year post-training:

- **Approx. 25% of all trainees show further activity**
- **Half of those are users with limited prior usage**

Over all-time (~6 years) this increases to 40% with the same proportion of users with limited prior usage





For the trainees who had previous activity on their Github account, there was very little change in activity pre- and post -training.

The number of new repositories and the average frequency of commits per repository 1 year pre- and 1 year post-training was approximately the same.

This approach only provides an estimate of trainee activity, with several factors that may skew the results for actual activity.

Some of these factors relate to general limitations with this approach, for example:

- It was only possible to survey public repositories
- Activity in repositories on other forges was not surveyed

Other factors relate to specific analytic choices, for example:

- Forked repositories were not included in this analysis
- Other activity (e.g. issue creation, branching, etc) also not analysed

Even if these estimates are broadly correct, this approach alone cannot tell us why 75% of trainees don't show further activity.

Key Findings:

- ➔ Approximately 25% (97) of trainees show activity on Github 1 year post-training.
- ➔ Half (47) are users with limited usage prior to the training, and half (50) are more established users.
- ➔ Of the more established users, there was no change in activity pre- and post-training.

This approach has value as a way of estimating post-training activity without needing direct user responses. However, these results alone lack the contextual information that post-training surveys would provide. This additional contextual information is essential for assessing if and how the training should be redesigned.

Training: https://srse-git-github-zero2hero.netlify.app/

Collaborative exercise: https://github.com/RSE-Sheffield/collaborative_github_exercise

Poster repo: https://github.com/ubdbra001/RSECon25_Poster

r-rse

University of Sheffield