
Shurui Zhou, Bogdan Vasilescu, Christian Kästner
Fork-based Development
Fork-based Development is Popular

<table>
<thead>
<tr>
<th>#Forks</th>
<th>#GitHub Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;50</td>
<td>61704</td>
</tr>
<tr>
<td>&gt;500</td>
<td>4787</td>
</tr>
<tr>
<td>&gt;1,000</td>
<td>2236</td>
</tr>
<tr>
<td>&gt;5,000</td>
<td>198</td>
</tr>
<tr>
<td>&gt;10,000</td>
<td>72</td>
</tr>
<tr>
<td>&gt;100,000</td>
<td>2</td>
</tr>
</tbody>
</table>

[GHTorrent 2019-06]  [GitHub Network View]
Network View - Lack of an overview
Problems → Inefficiency

- Lost Contribution
- Rejected PRs
- Redundant Development
- Fragmented Community

[Zhou et al. ICSE'18]
Lost Contribution

Only 14% of all forks of nine popular JavaScript projects on GitHub contained changes that were integrated back [Fung et al. 2012]
Problems → Inefficiency

Lack of an overview

- Lost Contribution
- Rejected PRs
- Redundant Development
- Fragmented Community
Rejected Pull Requests

- Demotivating [Steinmacher et al. ICSE'18]
- Misalignment with maintainers’ vision of the project
People Follow Different Processes

bit\text{coin} \hspace{1cm} \text{django} \hspace{1cm} \text{scikit learn} \hspace{1cm} \text{Marlin}
“To a large extent the features are driven by bitcoin improvement proposals, so if I would be looking for a feature, I would go for these proposals”

--Bitcoin developer
People Follow Different Processes

Fix issue #13048 - Documentation regarding p-value bootstrapping #14759

achievermina wants to merge 7 commits into scikit-learn:master from achievermina:p_valueBootstrapping

Conversation 9  Commits 7  Checks 11  Files changed 2

achievermina commented 3 days ago • edited

Issue #13048
People Follow Different Processes

- Project proposal
- Resolve issues on the issue tracker

VS

- Open for any contribution
Rejected Pull Requests

- Demotivating
- Misalignment with maintainers’ vision of the project
Problems → Inefficiency

- Lack of an overview
- Lost Contribution
- Rejected PRs
- Redundant Development
- Fragmented Community
Redundant Development

23% un-merged PRs were rejected due to redundant dev. [Gousios et al. ICSE’14]

Cost of Reviewing [Li et al. MSR’18]

De-motivate developers [Steinmacher et al. ICSE’18]

Detecting duplicate dev. [Zhou et al. SANER’19]
Problems → Inefficiency

Lack of an overview

Lost Contribution

Rejected PRs

Redundant Development

Fragmented Community
Communities Fragmentation (Hard Fork)

Behind the Scenes Bytes

3D Printer Firmware – Which to Choose and How to Change It?

by Michael Jones
Apr 4, 2018

% Hard forks among all forks
RQ: What characteristics and practices of a project associate with efficient forking practices?

- Lost Contribution
- Rejected PRs
- Redundant Development
- Fragmented Community
Research Method

- Interviewing Stakeholders
- Literature Search
- Sampling
- Quant. Practices
  - Inefficiencies
  - Context Factors
- Modeling
- Deriving Hypotheses
- Test Hypotheses
Coordination Mechanism Affects Forking Practices

- Project proposal
- Resolve issues on the issue tracker

VS

- Open for any contribution
Centralization makes it easier to coordinate the divisions’ product types but more difficult to take advantage of the divisions’ private information. [Brandts et al. 2018]
Deriving Hypotheses

Centralized mgmt ➔ Larger portion of merged PRs
Centralized mgmt ➔ Larger portion of contributing forks

(6 more in the paper)
Test Hypotheses

<table>
<thead>
<tr>
<th>Group</th>
<th>#fork</th>
<th>#projects</th>
<th>#projects in sample set</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>[3,000, +]</td>
<td>231</td>
<td>200</td>
</tr>
<tr>
<td>B</td>
<td>[1,000, 3,000)</td>
<td>847</td>
<td>300</td>
</tr>
<tr>
<td>C</td>
<td>[20, 1,000)</td>
<td>116,532</td>
<td>1300</td>
</tr>
</tbody>
</table>

Sampling

Quantifying

Inefficiencies

Practices

Context Factors

Modeling
Operationalization - Centralized Management

Measure: \[
\frac{\text{Number of PRs referring to an Existing Issue}}{\text{All the PRs}}
\]

**Fix issue #13048** - Documentation regarding p-value bootstrapping #14759

achievermina wants to merge 7 commits into scikit-learn:master from achievermina:p_valueBootstrapping

- Conversation 9
- Commits 7
- Checks 11
- Files changed 2

achievermina commented 3 days ago • edited

Issue #13048
Centralized Mgmt → More Merged PRs ($R^2 = 27\%$)

Ratio Merged PRs

Plus controls for:
- SubmitterPriorExpr
- SubmitterSocialConn.
- PR w/ test

Modularity
(6% of deviance explained)

Centralized Mgmt
(4% of deviance explained)
Centralized Mgmt $\rightarrow$ More Contributing Forks ($R^2 = 17\%$)

- Ratio contributing forks
- Modularity (1% of deviance explained)
- Centralized Mgmt (18% of deviance explained)

Plus controls for:
- NumForks
- Size
- ProjectAge
Evidence-based Intervention

For practitioners:
- Coordinating planned changes through an issue tracker

Trade-offs?
- Project proposal
- Resolve issues on the issue tracker
- Open for any contribution
Trade-off: Centralized Mgmt

Community Fragmentation

Centralized Mgmt
(12% of variance explained)

PR Merge Ratio
(35% of variance explained)

Plus controls for:
- NumFork
- Size
RQ: What characteristics and practices of a project associate with efficient forking practices?

- Coordination
- Modularity

Lack of an overview

Lost Contribution
Rejected PRs
Redundant Development
Fragmented Community
Opportunities to Design Further Interventions

- Tooling to navigate and understand changes in forks
- Making practices transparent
- Cost of community fragmentation
A Study of Inefficient and Efficient Forking Practices in Social Coding

- Evidence-based Suggestions
- Further research/tooling directions

Lack of an overview

Lost Contribution
Rejected PRs
Redundant Development
Fragmented Community

Interviewing Stakeholders
Literature Search
Sampling
Quant.
Inefficiencies
Practices
Context Factors
Modeling

Deriving Hypoth.
Test Hypoth.