Improving Collaboration Efficiency in Fork-based Development

Shurui Zhou

Thesis Committee: Christian Kästner (Chair), James D. Herbsleb, Laura A. Dabbish, Andrzej Wąsowski

Carnegie Mellon University
Collaboration is Everywhere
Globally Distributed Software Development
Help software developers to better collaborate
Help software developers to better collaborate

- Advances in tooling & SE principles
- Insights from other disciplines
- Mix a wide range of research methods
Help software developers to better collaborate

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods
Help software developers to better collaborate

- Advances in tooling & SE principles
- Insights from other disciplines
- Mix a wide range of research methods
Help software developers to better collaborate

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods
Help software developers to better collaborate

- Advances in tooling & SE principles
- Insights from other disciplines
- Mix a wide range of research methods

Problem → Intervention → Evaluation
Help software developers to better collaborate

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods
Traditional Collaboration Model
Traditional Collaboration Model
Traditional Collaboration Model
Fork-Based Development Changed Everything
Fork-based Dev. Changed Everything
Traditional Collaboration Model
Traditional Collaboration Model
Traditional Collaboration Model
Traditional Collaboration Model
Traditional Collaboration Model
Traditional Collaboration Model

Description

Subject: [PATCH] Patch for pre-calculated loops_per_jiffy

Attached is a patch which allows for setting a pre-calculated loops_per_jiffy. This patch was derived from the CONFIG_INSTANT_ON feature in the CELF source tree, which was developed by MontaVista. This feature is already available in the CELF source tree, for the ONAP board.

loops_per_jiffy (LPJ) is the value used internally by the kernel for the delay() function. Normally, LPJ is determined at boot time by the routine calibrate_delay(), in init/main.c. This routine takes approximately 250 ms to complete on my test machine. Note that the routine uses a sequence of programmed waits to determine the correct LPJ value, with each wait taking about 1 Hz (usually 10 ms) period. With a pre-calculated value, this calibration is eliminated.

This patch is currently against a Linux 2.4.20 kernel, for the x86 architecture.

When the patch is applied, a new option appears in the General setup menu of menuconfig: "Fast booting". When this option is enabled, you are asked to set the value of another new option: 'Loops per jiffy'. These set the config variables CONFIG_FASTBOOT and CONFIG_FASTBOOT_LPJ.

diffstat for this patch:

| Documentation/Configure.help | 23 ++++++++++++++++++++ |
| arch/i386/config.in | 6 ++++
| init/main.c | 13 +++++

3 files changed, 42 insertions(+)

To apply the patch, in the root of a kernel tree use:
patch -p1 <fastboot_lpj.patch

Source code

Signed-off-by: Tim Bird <tim.bird@am.sony.com>

diff -u -rNw linux-2.4.20.org/Documentation/Configure.help linux-2.4.20/Documentation/Configure.help
--- linux-2.4.20.org/Documentation/Configure.help Thu Nov 28 15:53:08 2002
+++ linux-2.4.20/Documentation/Configure.help Tue Sep 30 15:32:35 2003
@@ -5274,6 +5274,29 @@
 replacement for kernel.d) Say Y here and read about configuring it in <file:Documentation/kmod.txt>.

+Fast booting support
+CONFIG_FASTBOOT
+ Say Y here to enable faster booting of the Linux kernel. If you say + Y here, you will be asked to provide hardcoded values for some + parameters that the kernel usually probes for or determines at boot + time. This is primarily of interest in embedded devices where + quick boot time is a requirement.
+
+ If unsure, say N.
+
+Fast boot loops-per-jiffy
+CONFIG_FASTBOOT_LPJ
+ This is the number of loops passed to delay() to achieve a single + Hz of delay inside the kernel. It is roughly BogoMips * 5000. + To determine the correct value for your kernel, first turn off + the fast booting option, compile and boot the kernel on your target + hardware, then see what value is printed during the kernel boot. + Use that value here.
+
+ If unsure, don’t use the fast booting option. An incorrect value + will cause delays in the kernel to be incorrect. Although unlikely, + in the extreme case this might damage your hardware.
+
ARP daemon support
Traditional Collaboration Model
Fork-based Development

Upstream

Fork/Branch
Fork-based Development

Upstream

Fork/Branch

Commit
Fork-based Development

Upstream

Fork/Branch

Commit
Fork-based Development

Fork-based / Branch-based / Pull-based Dev.

Pull Request / Merge Request
Fork-based Dev. Lowers Entry Barriers
Fork-based Dev. Lowers Entry Barriers
Fork-based Dev. Lowers Entry Barriers

Upstream Fork/Branch → Pull Request (PR)
Fork-based Development
Fork-based Development
Fork-based Dev. Becomes Popular

<table>
<thead>
<tr>
<th>#Forks</th>
<th>#GitHub Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;50</td>
<td>114,120</td>
</tr>
<tr>
<td>&gt;500</td>
<td>9164</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]
### Fork-based Dev. Becomes Popular

<table>
<thead>
<tr>
<th>#Forks</th>
<th>#GitHub Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;50</td>
<td>114,120</td>
</tr>
<tr>
<td>&gt;500</td>
<td>9164</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]
Fork-based Dev. Becomes Popular

<table>
<thead>
<tr>
<th>#Forks</th>
<th>#GitHub Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;50</td>
<td>114,120</td>
</tr>
<tr>
<td>&gt;500</td>
<td>9164</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]
Fork-based Dev. Becomes Popular

https://github.com/customer-stories?type=enterprise
But
Problem -- Lost Contributions
Problem -- Redundant Development

fooesel commented on Aug 22, 2017

Sorry, but I can't stop laughing right now. I added *exactly* the same kind of functionality yesterday (just with a configurable ambient value and a debug command to also modify it during run time). See fbcbb3f

I can't believe this coincidence XD

Noiredd commented on Nov 3, 2017

Duplicate of #5869 and #5972, partially also #5879.
Problem -- Fragmented Community
Problem -- Fragmented Community

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

by Michael Jones
Apr 4, 2018
Problem

- Lost Contribution
- Redundant Development
- Fragmented Community
Problem

Lost Contribution

Redundant Development

Fragmented Community

Is duplicate always bad?
It is hard for individual teams to know who is doing what, which features exist elsewhere, and what code changes are made in other forks [1,2].


Project 'gitlab-org/gitlab-ce' was moved to 'gitlab-org/gitlab-foss'. Please update any links you still have the old path.

Closed  Opened 4 years ago by Adriano Vieira

I'd like to see all forked projects of one project

You have on the project home page a button which show us a quantity of forks from one project.

I'd like to see all forked projects of one project (even mine).

How could we see all forked projects of any project?
List of Forks

8 forks: 4 public and 3 protected

- aleksandrs-lebovskis / omnibus-gitlab
  Why relevant, or what is fixed or whatever copy goes here, in other words why this forked exists.

- bbodenmiller / omnibus-gitlab
  Why relevant, or what is fixed or whatever copy goes here, in other words why this forked exists.

- chaws / omnibus-gitlab
  Why relevant, or what is fixed or whatever copy goes here, in other words why the project was forked.

- chinnyannieb / omnibus-gitlab
  Why relevant, or what is fixed or whatever copy goes here, in other words why the project was forked.

- chinnytp / omnibus-gitlab
  Why relevant, or what is fixed or whatever copy goes here, in other words why the project was forked.

- donkey / omnibus-gitlab
  Why relevant, or what is fixed or whatever copy goes here, in other words why the project was forked.

2 protected forks you have no access to.
Problem

Network View

- Smoothieware
- Contributors
- Community
- Commits
- Code frequency
- Dependency graph

Owners

<table>
<thead>
<tr>
<th>May</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contributors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commits</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code frequency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependency graph</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoothieware</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>626Pilot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clementleger</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wolffmanjm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BitLabProjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Problem

Network View

Lack of Overview
Problem

Lack of Overview

Lost Contribution

Redundant Development

Fragmented Community
Improving Collaboration Efficiency

Software Dev.
Improving Collaboration Efficiency

Software Dev. → Distributed
Improving Collaboration Efficiency

Software Dev. → Distributed → Fork-Based
Improving Collaboration Efficiency

- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community
Improving Collaboration Efficiency

**Problem**
- Distributed
  - Lack of Overview
  - Lost Contribution
  - Redundant Development
  - Fragmented Community

**Solution**
- Fork-Based
  - Natural Intervention
    - Identifying Best Practices
  - New Intervention
    - Identifying Features
    - Identifying Redundancies

- [FSE’19]
- [ICSE’20]
- [ICSE’18]
- [SANER’19]
Improving Collaboration Efficiency

Software Dev. → Distributed

Problem
- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community

Solution
- Natural Intervention
- Identifying Best Practices
- New Intervention
- Identifying Features
- Identifying Redundancies

Fork-Based

References:
- [FSE’19]
- [ICSE’20]
- [ICSE’18]
- [SANER’19]
Problem:
- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community

Solution:
- Natural Intervention
  - Identifying Best Practices
  - New Intervention
    - Identifying Features
    - Identifying Redundancies

References:
- [FSE’19]
- [ICSE’20]
- [ICSE’18]
- [SANER’19]
Problem

- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community

Solution

- Natural Intervention
- Identifying Best Practices
- New Intervention
- Identifying Feature
- Identifying Redundancies

References:

- [SANER’19]
- [ICSE’18]
- [FSE’19]
- [ICSE’20]
- [SANER’19]
Problem

- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community

Solution

- Natural Intervention
- Identifying Best Practices
- New Intervention
  - Identifying Features
  - Identifying Redundancies

[ICSE’18]
[SANER’19]
[FSE’19]
[ICSE’20]
I study how communities using forks, design measures to quantify inefficiencies in fork-based development. To mitigate the inefficiencies, I propose two strategies: first, I conduct a cross-sectional correlational study to identify best practices and generate evidence-based recommendations that could improve collaboration efficiency; second, I design awareness tools to generate a better overview of code changes in an open source community, and detect redundant development to reduce waste of maintenance & development effort.
Problem

- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community

Solution

- Natural Intervention
- Identifying Best Practices
- New Intervention
- Identifying Features
- Identifying Redundancies

References:
- [FSE’19]
- [ICSE’20]
- [ICSE’18]
- [SANER’19]
Problem:
- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community

Solution:
- Natural Intervention
  - Identifying Best Practices
    - New Intervention
      - Identifying Feature
      - Identifying Redundancies

References:
- [FSE’19]
- [ICSE’20]
- [ICSE’18]
- [SANER’19]
Solution 1 – Identifying Natural Interventions

[FSE 2019]


Shurui Zhou
Carnegie Mellon University, USA

Bogdan Vasilescu
Carnegie Mellon University, USA

Christian Kästner
Carnegie Mellon University, USA
Solution 1 – Identifying Natural Interventions


Shurui Zhou
Carnegie Mellon University, USA

Bogdan Vasilescu
Carnegie Mellon University, USA

Christian Kästner
Carnegie Mellon University, USA
Projects are different

- Project proposal
- Resolve issues on the issue tracker
Projects are different

- Project proposal
- Resolve issues on the issue tracker

- Open for any contribution
Projects are different

- Centralized Mgmt
- Upfront Coordination through Issue Tracker

VS

- De-centralized Mgmt
- No Upfront Coordination
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]
What characteristics and practices of a project associate with efficient forking practices?
Research Method

- Interviewing Stakeholders
- Literature/Theory Search

Deriving Hypotheses
Centralized Management ➔ Larger portion of contributing forks
Test Hypotheses
Test Hypotheses

Cross-sectional Correlational Study

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods

Developmental Psychology
Test Hypotheses

Cross-sectional Correlational Study

• A single point in time.
• No need to manipulating variables
• Considers several characteristics at once
• Analyzes the prevailing characteristic in a given population

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods
Research Method

Interviewing Stakeholders → Deriving Hypotheses

Literature/Theory Search

Sampling

Quant. (Inefficiencies, Practices, Context Factors) → Test Hypotheses

Modeling
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>

Multiple Regression Modeling

Quantifying Inefficiencies
Practices
Context Factors

Sampling
Test Hypotheses

Sampling

Quantifying

Inefficiencies
Practices
Context Factors

<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>

Multiple Regression Modeling
Test Hypotheses

Sampling

Quantifying

Inefficiencies
Practices
Context Factors

Hypo: Centralized Management ➔ Larger portion of contributing forks
Operationalization - Centralized Management

Measure: Number of PRs referring to an Existing Issue

All the PRs

Fix issue #13048 - Documentation regarding p-value bootstrapping #14759
Operationalization - Centralized Management

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

Measure: \( \frac{\text{Number of Forks submitted PR(s)}}{\text{All the Active Forks}} \)

more efficient
Test Hypotheses

Sampling

Quantifying

Inefficiencies

Practices

Context Factors

Multiple Regression Modeling

<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>
Centralized Mgmt $\rightarrow$ More Contributing Forks (R2 = 17%)

Ratio Contributing Forks

+ 

Centralized Management
(18% of deviance explained)

Plus controls for:
- Number of Forks
- Project Age
- Size
Evidence-based Intervention

For practitioners:
- Coordinating planned changes through an issue tracker
Evidence-based Intervention

For practitioners:
- Coordinating planned changes through an issue tracker

Trade-offs?
Evidence-based Intervention

For practitioners:
- Coordinating planned changes through an issue tracker
Hypotheses

Centralized mgmt ➔ Higher likelihood of community fragm.
Old Notion of Forking: Splitting off a Community

A need of a community that was not fulfilled by the original project.
A need of a community that was not fulfilled by the original project.
Forking was a Weighty Decision

A strong norm against forking [Yoo 2016]
Community Fragm. is Expensive
“Some open-source forks have made life difficult for developers. ... that will force developers to pick sides.” – Lauren Orsini
“Some open-source forks have made life difficult for developers. ... that will force developers to pick sides.” –Lauren Orsini

Node.js and io.js are settling their differences, merging back together
Different kinds of Forks

Hard Fork VS (Social) Fork
Operationalization – Community Fragmentation

Hard Fork VS (Social) Fork
Operationalization – Community Fragmentation

Hard Fork  vs  (Social) Fork

How Has Forking Changed in the Last 20 Years?
A Study of Hard Forks on GitHub

Shurui Zhou  
Carnegie Mellon University, USA

Bogdan Vasilescu  
Carnegie Mellon University, USA

Christian Kästner  
Carnegie Mellon University, USA

[ICSE’20]
Detecting Hard Forks

Operationalization – Community Fragmentation
Operationalization – Community Fragmentation

Identifying Evolution Patterns of Hard Forks
Identifying Evolution Patterns of Hard Forks
Identifying Evolution Patterns of Hard Forks

Table 2: Evolution patterns of hard forks

<table>
<thead>
<tr>
<th>Id</th>
<th>Category</th>
<th>Sub-category</th>
<th>Example</th>
<th>Count</th>
<th>Interview(s)</th>
</tr>
</thead>
</table>
Identifying Evolution Patterns of Hard Forks

Table 2: Evolution patterns of hard forks

<table>
<thead>
<tr>
<th>Id</th>
<th>Category</th>
<th>Sub-category</th>
<th>Example</th>
<th>Count</th>
<th>Interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Success</td>
<td>632 Upstream remains inactive</td>
<td>● ● ● ●</td>
<td>576</td>
<td>P12</td>
</tr>
<tr>
<td>2</td>
<td>Revive Dead Project</td>
<td>Upstream active again</td>
<td>● ● ● ●</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Not success</td>
<td>620 Upstream remains inactive</td>
<td>● ● ● ●</td>
<td>419</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Only merge</td>
<td></td>
<td>● ● ● ● ● ●</td>
<td>26</td>
<td>P10</td>
</tr>
<tr>
<td>5</td>
<td>Both Alive</td>
<td>723 Only sync</td>
<td>● ● ● ● ● ●</td>
<td>197</td>
<td>P2, P13, P15</td>
</tr>
<tr>
<td>6</td>
<td>Merge &amp; sync</td>
<td></td>
<td>● ● ● ● ● ●</td>
<td>28</td>
<td>P9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>● ● ● ● ● ●</td>
<td>562</td>
<td>P1, P4, P5, P7, P14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>● ● ● ● ● ●</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>No interaction</td>
<td></td>
<td>● ● ● ● ● ●</td>
<td>6313</td>
<td>P6, P9, P11</td>
</tr>
<tr>
<td>12</td>
<td>Only merge</td>
<td></td>
<td>● ● ● ● ● ●</td>
<td>388</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Fork does not out live upstream</td>
<td>6251 Only sync</td>
<td>● ● ● ● ● ●</td>
<td>762</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Merge &amp; sync</td>
<td></td>
<td>● ● ● ● ● ●</td>
<td>199</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>No interaction</td>
<td></td>
<td>● ● ● ● ● ●</td>
<td>4982</td>
<td></td>
</tr>
</tbody>
</table>
Operationalization – Community Fragmentation

Ratio of Community Fragmentation:

Hard Fork + (Social) Fork
Example -- Fragmented Community

Ultimaker

Ultimaker / Ultimaker2Marlin

Hard Fork

MarlinFirmware / Marlin
Hypoth: Centralized Mgmt $\rightarrow$ Community Fragm.

Community Fragmentation

Plus controls for:
- Number of Forks
- Size

Centralized Management

(12% of variance explained)
Trade-off: Centralized Management

Centralized Management

better

Contributing Forks (Efficient)

worse

Community Fragmentation (Inefficient)
Evidence-based Intervention

For Practitioners:
- Coordinating planned changes through an issue tracker.
- Making deliberate trade-off decision about to what degree:
  - they can remain open to various external contributions
  - they are willing to accept some degree of fragmentation
Evidence-based Intervention

Avoid Cargo Cult Science/thinking
What characteristics practices of a project associate with efficient forking practices?

- Coordination
- Modularity
Evidence-based Intervention

For Researchers & Tool Builders:

- Tooling to navigate and understand changes among fragmented communities/hard forks.
Evidence-based Intervention

For Researchers & Tool Builders:

- Tooling to navigate and understand changes among fragmented communities/hard forks.
- Making Practice Transparent

Centralized Management Index: 98%
Modularity: High
Problem:

- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community

Solution:

- Natural Intervention
- Identifying Best Practices
- New Intervention

Advances in tooling & SE principles
Insights from other disciplines
Mix a wide range of research methods
Designing New Interventions

Lack of Awareness

Organizational Theory
Social Behavior

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods
Problem

Lack of Overview
Lost Contribution
Redundant Development
Fragmented Community

Solution

Natural Intervention
Identifying Best Practices

New Intervention
Awareness Tools

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods
Problem

- Lack of Overview
- Lost Contribution
- Redundant Development

Solution

- Natural Intervention
  - Identifying Best Practices
  - [FSE’19]
  - [ICSE’20]

- New Intervention
  - Identifying Features
    - [ICSE’18]
  - Identifying Redundancies
    - [SANER’19]

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods
Problem

Lack of Overview
Lost Contribution

Solution

Natural Intervention
Identifying Best Practices

New Intervention
Identifying Features

Lost Contribution

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods
Solution 2 – Identifying Features in Forks

[ICSE 2018]

Identifying Features in Forks

Shurui Zhou
Carnegie Mellon University

Ștefan Stănciulescu
IT University of Copenhagen

Olaf Leßenich
University of Passau

Yingfei Xiong
Peking University

Andrzej Wąsowski
IT University of Copenhagen

Christian Kästner
Carnegie Mellon University
Goal: a Better Overview of Forks

Which are the active forks?

What kind of code changes have been made in forks?

What features were implemented in forks?
Summarizing forks that have un-merged commits

Mapping between feature to code changes
Email System

- Signature
- Encryption

<table>
<thead>
<tr>
<th>Feature id</th>
<th>Keyword List</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sig.</td>
<td>Signature, isSigned, ...</td>
<td>23</td>
</tr>
<tr>
<td>Enc.</td>
<td>Encryption, Decryption, isEncrypted, Decrypt, ...</td>
<td>100</td>
</tr>
</tbody>
</table>
Email System

- Signature
- Encryption
- Decryption

<table>
<thead>
<tr>
<th>Feature id</th>
<th>Keyword List</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sig.</td>
<td>Signature, isSigned, ...</td>
<td>23</td>
</tr>
<tr>
<td>Enc.</td>
<td>Encryption, isEncrypted, ...</td>
<td>55</td>
</tr>
<tr>
<td>Dec</td>
<td>Decryption, Decrypt, ...</td>
<td>45</td>
</tr>
</tbody>
</table>
## INFOX

<table>
<thead>
<tr>
<th>Feature</th>
<th>Navigation</th>
<th>Keyword List</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sig.</td>
<td>Prev.</td>
<td>Signature, isSigned, ...</td>
<td>23</td>
</tr>
<tr>
<td>Enc.</td>
<td>Next</td>
<td>Encryption, Decryption, isEncrypted, Decrypt, ...</td>
<td>100</td>
</tr>
</tbody>
</table>

```cpp
+ connect (ui->MaximumGeneratedBlockWeight, SIGNAL ( textChanged ( const QString &)), this , SLOT ( showR
+ connect (ui->MaximumConsensusBlockWeight, SIGNAL ( textChanged ( const QString &)), this , SLOT ( showR

@@ -207,11 +216,16 @@ void OptionsDialog::setMapper()

+ mapper-> addMapping (ui->MaximumGeneratedBlockWeight, OptionsModel::MaximumGeneratedBlockWeight);
+ mapper-> addMapping (ui->MaximumConsensusBlockWeight, OptionsModel::MaximumConsensusBlockWeight);
+ void OptionsDialog::setOkButtonState ()
+ ui->okButton-> setEnabled ( okbutton_blockweight | okbutton_proxy );

@@ -273,6 +287,28 @@ void OptionsDialog::showRestartWarning(bool fPersistent)

+ int mgbw, mcbw;
+ mgbw = ui->MaximumGeneratedBlockWeight-> text (). toInt ();
+ mcbw = ui->MaximumConsensusBlockWeight-> text (). toInt ();
```
# INFOX

<table>
<thead>
<tr>
<th>Feature</th>
<th>Author</th>
<th>Last Commit</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>onre.</td>
<td>fakerequest</td>
<td>May 28</td>
<td>28</td>
</tr>
<tr>
<td>hea.</td>
<td>qnetworkrepli</td>
<td>data, reply, buffers</td>
<td>10</td>
</tr>
<tr>
<td>sett.</td>
<td>settings, a, phantomcfg, not, bug, fix, websecurityen, qwebset, setattribute, qwebsettings</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>dow.</td>
<td>download, com, pull, file, ad, support, ariya, http</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>get</td>
<td>get, qt, are, kepier, el, mouse, require, clicks, hard, absolute, setfocus, button, coordinates, keypress</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>readlin</td>
<td>readlin, asyncreadrequest, asyncread, qobject, qstring, readline, data, data, qvariant, m_file, m_data, file, read</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>uint</td>
<td>uint, tmp_value, value, tmp, octet, qvariant, data, namesize, readrawdata, fromvalue</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>frame</td>
<td>frame, bmconsumeok, bmdeliver, method_id, id_enum, bmgetempty, bmreject, bmrecover</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>cono.</td>
<td>conoack, 0x04, consumeoptions, coexclusive, declar_flag, consumeoption, cononlocal, cononwait, flag</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>DeviaVir/phantomjs</td>
<td>DeviaVir/phantomjs, last commit: Jan 25, 2016</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
INFOX Overview Page

<table>
<thead>
<tr>
<th>Feature</th>
<th>ibradypod/phantomjs, last commit: May 28</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>onere</td>
<td>oneresourcequery, body, downloadmultibuffer, qnetworkrepl, response, qbytearray, data, reply, buffers</td>
<td>28</td>
</tr>
<tr>
<td>hea</td>
<td>header, getcookiestringfromurl, body, cookie, get, qurl, qnetworkrepl, response, url</td>
<td>10</td>
</tr>
<tr>
<td>sett</td>
<td>settings, a, phantomcfg, not, bug, fix, websecurity, qwebset, setattribute, qwebsettings</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>raff/phantomjs, last commit: Mar 5</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>dow</td>
<td>download, com, pull, file, ad, support, arinya, http</td>
<td>39</td>
</tr>
<tr>
<td>get</td>
<td>get, qf, are, kei, enter, el, mouse, require, clicks, hard, absolute, setfocus, button, coordinates, keypress</td>
<td>29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>rickahler/phantomjs, last commit: Feb 2</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>readlin</td>
<td>readlin, asyncreadrequest, asyncread, qobject, qstring, readline, data, qvariant, m_file, m_data, file, read</td>
<td>30</td>
</tr>
<tr>
<td>uint</td>
<td>uint, temp_value, value, tmp, octet, qvariant, data, namesize, readrawdata, fromvalue</td>
<td>80</td>
</tr>
<tr>
<td>frame</td>
<td>frame, bmconsumemk, bmdeliver, method_id, id_enum, bmetempty, bmreject, bmrecover</td>
<td>29</td>
</tr>
<tr>
<td>cono</td>
<td>conoack, 0x04, consumeoptions, coexclusive, declar_flag, consumeoption, conolocal, conowait, flag</td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>DeviaVir/phantomjs, last commit: Jan 25, 2016</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>allow</td>
<td>allow, set, customwebpag, ratio, m_customwebpag, devicepixelratio, webpage, setdevicepixelratio</td>
<td>7</td>
</tr>
</tbody>
</table>

Problem

Network View
INFOX

Dependency graph for code changes (static analysis)

Clustering features (community detection)

Labeling features (NLP)
Dependency Graph

File 1: Email.h

```c
struct email {
    char *subject;
    char *body;
    int isEncrypted;
};
void printMail (struct email *msg);

+ int isEncrypted (struct email *msg);
+ int isSigned (struct email *msg);
```

3 Dependencies

- **DU** – Definition-Usage
- **CF** – Control Flow
- **H** – Hierarchy; **A** - Adjacency

File 2: Email.c

```c
+ void printMail (struct email *msg) {
    printf("SUBJECT:", msg->subject);
    printf("SIGNED:", msg->isSigned);
    if (0 == (isEncrypted(msg)))
        printf("BODY:", msg->body);
    else
        printf("Encrypted msg.");
}

+ int isEncrypted (struct email *msg) {
    return msg->isEncrypted;
}

+ int isSigned (struct email *msg) {
    return msg->isSigned;
}
```
Dependency Graph

File 1: Email.h

```c
struct email {
    char *subject;
    char *body;
    int isEncrypted;
};
void printMail ( struct email *msg);
int isEncrypted ( struct email *msg);
int isSigned ( struct email *msg);
```

3 Dependencies

- **DU** – Definition-Usage
- **CF** – Control Flow
- **H** – Hierarchy; A – Adjacency

---

File 2: Email.c

```c
+ void printMail ( struct email *msg)
{
    printf ("SUBJECT:", msg->subject);
    printf ("SIGNED:", msg->isSigned);
    if (0 == (isEncrypted(msg) ))
        printf ( "BODY:", msg -> body );
    else
        printf ( "Encrypted msg.\n" );
}
+ int isEncrypted ( struct email *msg)
+ {
    return msg->isEncrypted;
    }
+ int isSigned ( struct email *msg)
+ {
    return msg->isSigned;
    }
```
**Dependency Graph**

**File 1: Email.h**

```c
struct email {
    char *subject;
    char *body;
    int isEncrypted;
};

void printMail ( struct email *msg);

+ int isEncrypted ( struct email *msg);
+ int isSigned ( struct email *msg);
```

**File 2: Email.c**

```c
+ void printMail ( struct email *msg) {
    printf ("SUBJECT:", msg -> subject );
    printf ("SIGNED:", msg->isSigned);
    if (0 == (isEncrypted(msg) ) )
        printf ( "BODY:", msg -> body );
    else
        printf ( "Encrypted msg. ");
}

+ int isEncrypted ( struct email *msg)
+ {
+     return msg->isEncrypted;
+ }

+ int isSigned ( struct email *msg)
+ {
+     return msg->isSigned;
+ }
```

3 Dependencies

- **DU – Definition-Usage**
- **CF – Control Flow**
- **H – Hierarchy; A - Adjacency**
Dependency Graph

File 1: Email.h

```c
struct email {
    char *subject;
    char *body;
    + int isEncrypted;
};
void printMail ( struct email *msg);
+ int isEncrypted (struct email *msg);
+ int isSigned (struct email *msg);
```

File 2: Email.c

```c
+ void printMail ( struct email *msg) {
    printf ("SUBJECT:", msg -> subject );
    + printf ("SIGNED:", msg->isSigned);
    + if (0 == (isEncrypted(msg) ) )
        printf ( "BODY:", msg -> body );
    + else
        printf ( "Encrypted msg." );
    }
+ int isEncrypted (struct email *msg)
    + { return msg->isEncrypted;
+ }
+ int isSigned (struct email *msg)
    + { return msg->isSigned;
+ }
```

3 Dependencies

- DU – Definition-Usage
- CF – Control Flow
- H – Hierarchy; A – Adjacency
Dependency Graph

INFOX: Goal: a Better Overview of Forks

Dependency graph

Clustering features

Labeling features

upstream  fork

diff

labeled, changed code
base code
Dependency Graph

Dependency graph

Clustering features

Labeling features

Goal: a Better Overview of Forks
Dependency Graph

Splitting

Dependency graph

Clustering features

Labeling features

INFOX

Goal: a Better Overview of Forks
Dependency Graph

Network Analysis
Girvan–Newman algorithm

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods

Dependency graph
Clustering features
Labeling features
Dependency Graph

Splitting

Dependency graph

Clustering features

Labeling features
Dependency Graph

Splitting

Joining

Dependency graph

Clustering features

Labeling features

INFOX

Goal: a Better Overview of Forks
Dependency Graph

- commit message
- code
- comment

TF-IDF, N-Gram

<table>
<thead>
<tr>
<th>Sig.</th>
<th>Signature, isSigned, ...</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enc.</td>
<td>Encryption, Decryption, isEncrpyed, Decrypt, ...</td>
<td>100</td>
</tr>
</tbody>
</table>

Dependency graph

Clustering features

Labeling features
### Goal: a Better Overview of Forks

<table>
<thead>
<tr>
<th>Feature</th>
<th>ibradypod/phantomjs, last commit: May 28</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>onre.</td>
<td>onresourcerequest, bodi, downloadmultibuffer, qnetworkrepli, respons, qbytearray, data, reply, buffers</td>
<td>28</td>
</tr>
<tr>
<td>hea.</td>
<td>header, getcookiestringfromurl, bodi, cookie, get, qurl, qnetworkrepli, respons, url</td>
<td>10</td>
</tr>
<tr>
<td>sett.</td>
<td>settings, a, phantomcfg, not, bug, fix, websecurityen, qwebset, setattribute, qwebsettings</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>raff/phantomjs, last commit: Mar 5</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>dow.</td>
<td>download, com, pull, file, ad, support, ariya, http</td>
<td>39</td>
</tr>
<tr>
<td>get</td>
<td>get, qt, are, kei_enter, el, mouse, require, clicks, hard, absolute, setfocus, button, coordinates, keypress</td>
<td>29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>ricokahler/phantomjs, last commit: Feb 2</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>readlin</td>
<td>readlin, asyncreadrequest, asyncread, qobject, qstring, readline, data, qvariant, m_file, m_data, file, read</td>
<td>30</td>
</tr>
<tr>
<td>uint</td>
<td>uint, tmp_value_, value, tmp, octet, qvariant, data, namesize_, readrawdata, fromvalue</td>
<td>80</td>
</tr>
<tr>
<td>frame</td>
<td>frame, bmconsumeok, bmdeliver, method_id, id_enum, bmgetempty, bmreject, bmrecover</td>
<td>29</td>
</tr>
<tr>
<td>cono.</td>
<td>conoack, 0x04, consumeoptions, coexclusive, declar_flag, consumeoption, conolocal, conowait, flag</td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>DeviaVir/phantomjs, last commit: Jan 25, 2016</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>allow</td>
<td>allow, set, customwebpag, ratio, m_customwebpag, devicepixelratio, webpage, setdevicepixelratio</td>
<td>7</td>
</tr>
</tbody>
</table>
## INFOX

<table>
<thead>
<tr>
<th>Feature</th>
<th>Navigation</th>
<th>Keyword List</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sig.</td>
<td>Prev.</td>
<td>Signature, isSigned, ...</td>
<td>23</td>
</tr>
<tr>
<td>Enc.</td>
<td>Prev.</td>
<td>Encryption, Decryption, isEncrypted, Decrypt, ...</td>
<td>100</td>
</tr>
</tbody>
</table>

### Code Snippet

```cpp
+ connect (ui->MaximumGeneratedBlockWeight, SIGNAL ( textChanged ( const QString &)), this , SLOT ( showR...)
+ connect (ui->MaximumConsensusBlockWeight, SIGNAL ( textChanged ( const QString &)), this , SLOT ( showR...)
```

```cpp
@ -287,11 +216,16 void OptionsDialog::setMapper()
+ mapper-> addMapping (ui->MaximumGeneratedBlockWeight, OptionsModel::MaximumGeneratedBlockWeight);
+ mapper-> addMapping (ui->MaximumConsensusBlockWeight, OptionsModel::MaximumConsensusBlockWeight);
+ void OptionsDialog::setOkButtonState ()
+ ui->okButton-> setEnabled ( okbutton_blockweight || okbutton_proxy );
```

```cpp
@ -273,6 +287,28 void OptionsDialog::showRestartWarning(bool fPersistent)
+ int mgbw, mcbw;
+ mgbw = ui->MaximumGeneratedBlockWeight-> text (). toInt ();
+ mcbw = ui->MaximumConsensusBlockWeight-> text (). toInt ();
```
INFOX - Evaluation

Effectiveness

Usefulness

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods
RQ1: To what extent do identified clusters correspond to features?
INFOX - Effectiveness

RQ1: To what extent do identified clusters correspond to features?

Quantitative Study
RQ1: To what extent do identified clusters correspond to features?

Quantitative Study

Ground Truth?
INFOX - Effectiveness

- 10 C/C++ projects with #ifdef
- 156 test cases per project

<table>
<thead>
<tr>
<th>Project</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cherokee</td>
<td>328</td>
</tr>
<tr>
<td>clamav</td>
<td>285</td>
</tr>
<tr>
<td>ghostscript</td>
<td>816</td>
</tr>
<tr>
<td>Marlin</td>
<td>280</td>
</tr>
<tr>
<td>MPSolve</td>
<td>17</td>
</tr>
<tr>
<td>openvpn</td>
<td>276</td>
</tr>
<tr>
<td>subversion</td>
<td>409</td>
</tr>
<tr>
<td>tcl</td>
<td>2,481</td>
</tr>
<tr>
<td>xorg-server</td>
<td>1,360</td>
</tr>
<tr>
<td>xterm</td>
<td>453</td>
</tr>
</tbody>
</table>
INFOX - Effectiveness

- 10 C/C++ projects with #ifdef
- 156 test cases per project

INFOX assigned features with 90% accuracy.

<table>
<thead>
<tr>
<th>Project</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cherokee</td>
<td>328</td>
</tr>
<tr>
<td>clamav</td>
<td>285</td>
</tr>
<tr>
<td>ghostscript</td>
<td>816</td>
</tr>
<tr>
<td>Marlin</td>
<td>280</td>
</tr>
<tr>
<td>MPSolve</td>
<td>17</td>
</tr>
<tr>
<td>openvpn</td>
<td>276</td>
</tr>
<tr>
<td>subversion</td>
<td>409</td>
</tr>
<tr>
<td>tcl</td>
<td>2,481</td>
</tr>
<tr>
<td>xorg-server</td>
<td>1,360</td>
</tr>
<tr>
<td>xterm</td>
<td>453</td>
</tr>
</tbody>
</table>
INFOX - Evaluation

<table>
<thead>
<tr>
<th>Effectiveness</th>
<th>Usefulness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative Study</td>
<td>Human-subject Study</td>
</tr>
</tbody>
</table>

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods
Human-subject Study

- 11 developers

<table>
<thead>
<tr>
<th>Project</th>
<th>#Forks</th>
</tr>
</thead>
<tbody>
<tr>
<td>MarlinFirmware/Marlin</td>
<td>4,149</td>
</tr>
<tr>
<td>Smoothieware/Smoothieware</td>
<td>566</td>
</tr>
<tr>
<td>grpc/grpc</td>
<td>2,226</td>
</tr>
<tr>
<td>timscaffidi/ofxVideorecorder</td>
<td>60</td>
</tr>
<tr>
<td>arduino/Arduino</td>
<td>5,592</td>
</tr>
<tr>
<td>bitcoin/bitcoin</td>
<td>9,696</td>
</tr>
<tr>
<td>ariya/phantomjs</td>
<td>4,921</td>
</tr>
</tbody>
</table>
Most of the developers agree with the features that INFOX detected after a few steps of splitting and merging.
Can INFOX help developers to gain a better overview of repository forks?

### Human-subject Study - Usefulness

<table>
<thead>
<tr>
<th>Feature</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ibradypod/phantomjs, last commit: May 28</td>
<td>28</td>
</tr>
<tr>
<td>one resourcesrequest, body, downloadmultibuffer, qnetworkrepli, respons, qbytearray, data, reply, buffer</td>
<td>10</td>
</tr>
<tr>
<td>header, getcookiestringfromurl, body, cookie, get, qurl, qnetworkrepli, respons, url</td>
<td>10</td>
</tr>
<tr>
<td>settings, a, phantomconfig, not, bug, fix, websecurityen, qweset, setattribute, qwebsettings</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ralf/phantomjs, last commit: Mar 5</td>
<td>39</td>
</tr>
<tr>
<td>dow. download, com, pull, file, ad, support, ariya, http</td>
<td>29</td>
</tr>
<tr>
<td>get q, are, kei_, enter, el, mouse, require, clicks, hard, absolute, setfocus, button, coordinates, keypress</td>
<td>29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ricokahler/phantomjs, last commit: Feb 2</td>
<td>30</td>
</tr>
<tr>
<td>readlin readlin, asyncreadrequest, asyncread, qobject, qstring, readline, data, qvariant, m_file, m_data, file, read</td>
<td>80</td>
</tr>
<tr>
<td>uint uint, tmp_value, value, tmp, octet, qvariant, data, namesize, reeadrawdata, fromvalue</td>
<td>29</td>
</tr>
<tr>
<td>frame frame, bmconsumeok, bmdelete, method_id, id_enum, bmgetempty, bmreject, bmrecovery</td>
<td>29</td>
</tr>
<tr>
<td>cono. cono, bmconsume2, consumeoptions, consumeoptions, consumeoptions, consumeoptions, consumeoptions, consumeoptions, consumeoptions, consumeoptions, consumeoptions, consumeoptions, consumeoptions</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DevilaVir/phantomjs, last commit: Jan 25, 2016</td>
<td>7</td>
</tr>
<tr>
<td>allow allow, set, customwebpag, ratio, m_customwebpag, devicexmlratio, webpage, setdevicexmlratio</td>
<td>7</td>
</tr>
</tbody>
</table>

### Problem

Network View

VS
Interesting and Reusable Contribution

P5: “If it is only exists in this fork, then I want to somehow get this fork into my fork.”
Redundant Development

“It does look like somebody did a very simple one-function. I think they should use our code, there is great reason to use it.”
Forks Insight - Forks for Atom

**Language:** JavaScript  
**Forked on GitHub:** 8654  
**Active Forks:** 885  
**Forks containing unmerged code:** 106  
**Updated at:** 2018-02-24 18:43 (UTC)

<table>
<thead>
<tr>
<th>Fork</th>
<th>% Commits</th>
<th>% Changed Files</th>
<th>% Lines of code changed</th>
<th>Representative Keyword</th>
<th>% Last Commit</th>
<th>% Create</th>
<th>Add Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>ToniFerra72/atom</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>hola</td>
<td>2018-02-22</td>
<td>2018-02-22</td>
<td>✔️</td>
</tr>
<tr>
<td>larsdroid/atom</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>developing, documented, api, reference, docs, for</td>
<td>2018-02-14</td>
<td>2018-02-13</td>
<td>✔️</td>
</tr>
<tr>
<td>alexheretic/atom</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>normalizedpath, oldpath, normalized, startswith, starts, old</td>
<td>2018-01-24</td>
<td>2018-01-24</td>
<td>✔️</td>
</tr>
<tr>
<td>rk182/atom</td>
<td>1</td>
<td>74</td>
<td>2785</td>
<td>class, div, href, img, compatible, jeg</td>
<td>2018-01-12</td>
<td>2018-01-12</td>
<td>✔️</td>
</tr>
<tr>
<td>elusora/atom</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>transparent</td>
<td>2018-01-10</td>
<td>2018-01-10</td>
<td>✔️</td>
</tr>
</tbody>
</table>

Showing 1 to 5 of 106 entries
Problem

- Lack of Overview
- Redundant Development

Solution

- Natural Intervention
- Identifying Best Practices
- New Intervention
- Identifying Redundancies

Advances in tooling & SE principles
Insights from other disciplines
Mix a wide range of research methods
Solution 3 – Identifying Redundancies in Forks

[SANER 2019]

Identifying Redundancies in Fork-based Development

Luyao Ren  
Peking University, China

Shurui Zhou, Christian Kästner  
Carnegie Mellon University, USA

Andrzej Wąsowski  
IT University of Copenhagen, Denmark
Problem -- Redundant Development

foosel commented on Aug 22, 2017

Sorry, but I can't stop laughing right now. I added exactly the same kind of functionality yesterday (just with a configurable ambient value and a debug command to also modify it during run time). See fbcbb3f

I can't believe this coincidence XD

Noiredd commented on Nov 3, 2017

Duplicate of #5869 and #5972, partially also #5879.
Cost / Waste

For maintainer:
- **Maintenance effort**
  Before a duplicate PR is identified:
  2.6 reviewers
  5.2 review comments [Li et al. 2018]

For developers:
- **De-motivate developers** [Steinmacher et al. 2018]
Research Method

Manually analyze duplicate PRs

Developing clues as indicators

Operationalization

ML predicting redundancies
Evaluation - Effectiveness

RQ1: How accurate is our approach to help maintainers identify redundant contributions?

RQ2: How much effort could our approach save for developers in terms of commits?
RQ1: helping maintainers to find duplicate PRs

Randomly sample 400 PRs from each project

Precision 55%-82%
Recall 10%-25%
RQ2: helping developers to find duplicate changes early

Recall 46% - 71%
0.07–0.5% false positive rate
Save 1.9 - 3.0 commits per PR
Application Scenario

DuplicatePR-bot
Application Scenario

DuplicatePR-bot commented 19 days ago

Hi there! This pull request looks like it might be a duplicate of #1370, since it has the same issue number, a similar title, and similar commits.

To improve our bot, you can help us out by clicking one of the options below:
- This pull request is a duplicate, so this comment was useful. check
- This pull request is not a duplicate, but this comment was useful nevertheless. check
- This pull request is not a duplicate, so this comment was not useful. check
- I do not need this service, so this comment was not useful. check

This bot is currently in its alpha stage, and we are only sending at most one comment per repository. If you are interested in using our bot in the future, please subscribe. If you would like to learn more, see our web page.

sergeyrolich commented 19 days ago

Duplicate #1370, close

👍 1

sergeyrolich closed this 19 days ago
Duplicate PR-bot commented 19 days ago

Hi there! This pull request looks like it might be a duplicate of #1370, so number, a similar title, and similar commits.

To improve our bot, you can help us out by clicking one of the options:
- This pull request is a duplicate, so this comment was useful. check
- This pull request is not a duplicate, but this comment was useful. check
- This pull request is not a duplicate, so this comment was not useful. check
- I do not need this service, so this comment was not useful. check

This bot is currently in its alpha stage, and we are only sending at most 1 comment per PR. If you are interested in using our bot in the future, please subscribe. If you see our web page.

sergeyrolich commented 19 days ago

Duplicate #1370, close

👍 1

sergeyrolich closed this 19 days ago

Duplicate PR pairs successfully detected:

<table>
<thead>
<tr>
<th>repo</th>
<th>PR1</th>
<th>PR2</th>
</tr>
</thead>
<tbody>
<tr>
<td>gomods/athens</td>
<td>1423</td>
<td>1370</td>
</tr>
<tr>
<td>matomo-org/matomo</td>
<td>14137</td>
<td>14975</td>
</tr>
<tr>
<td>mrdooib/three.js</td>
<td>10597</td>
<td>13706</td>
</tr>
<tr>
<td>mrdooib/three.js</td>
<td>6230</td>
<td>6301</td>
</tr>
<tr>
<td>spring-projects/spring-framework</td>
<td>1695</td>
<td>1617</td>
</tr>
<tr>
<td>spring-projects/spring-framework</td>
<td>1492</td>
<td>1628</td>
</tr>
<tr>
<td>spring-projects/spring-framework</td>
<td>1218</td>
<td>1653</td>
</tr>
<tr>
<td>spring-projects/spring-framework</td>
<td>569</td>
<td>566</td>
</tr>
<tr>
<td>spring-projects/spring-framework</td>
<td>758</td>
<td>1827</td>
</tr>
<tr>
<td>hashicorp/terraform</td>
<td>18186</td>
<td>18296</td>
</tr>
<tr>
<td>cocos2d/cocos2d-x</td>
<td>14883</td>
<td>13687</td>
</tr>
<tr>
<td>cocos2d/cocos2d-x</td>
<td>14794</td>
<td>7565</td>
</tr>
<tr>
<td>facebook/react</td>
<td>12760</td>
<td>13169</td>
</tr>
</tbody>
</table>
Problem

- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community

Solution

- Natural Intervention
  - Identifying Best Practices
  - New Intervention
    - Identifying Features
    - Identifying Redundancies

Advances in tooling & SE principles
Insights from other disciplines
Mix a wide range of research methods
Limitation

Generalizability
Limitation

Generalizability
Construct Validity
Improving Collaboration Efficiency

**Problem**
- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community

**Solution**
- Natural Intervention
  - Identifying Best Practices
  - New Intervention
    - Identifying Features
    - Identifying Redundancies

**Software Dev.**
- Distributed

**Fork-Based**

- Advances in tooling & SE principles
- Insights from other disciplines
- Mix a wide range of research methods

[FSE’19] [ICSE’20] [ICSE’18] [SANER’19]
Improving Collaboration Efficiency

Software Dev. → Distributed → Problem

- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community

Solution → Fork-Based

- Natural Intervention
- Identifying Best Practices
- New Intervention
- Identifying Features
- Identifying Redundancies

Overview of Hard Forks

Centralized Management Index 98%
Modularity High

[FSE’19] [ICSE’20] [ICSE’18] [SANER’19]
Improving Collaboration Efficiency

Software Dev. — Distributed — Fork-Based

**Problem**
- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community
- Insufficient Coordination Capability

**Solution**
- Natural Intervention
- Identifying Best Practices
- New Intervention

- Identifying Features
- Identifying Redundancies
- Identifying Lost Contribution
- Identifying Relevant People
- Facilitating Particular Interactions

---

Centralized Management Index: 98%
Modularity: High

Overview of Hard Forks

---

[FSE’19] [ICSE’20] [SANER’19] [ICSE’18]
Improving Collaboration Efficiency

### Problem

- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community
- Insufficient Coordination Capability

### Solution

- Natural Intervention
- Identifying Best Practices
- New Intervention
- Identifying Feature
- Identifying Redundancies
- Identifying Lost Contribution
- Identifying Relevant People
- Facilitate Particular Interactions

### Software Dev.

- Distributed

### Fork-Based

- [FSE’19]
- [ICSE’20]
- [ICSE’18]
- [SANER’19]

### Additional Notes

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods
Improving Collaboration Efficiency

Software Dev. → Distributed → Interdisciplinary → Fork-Based

Problem:
- Lack of Overview
- Lost Contribution
- Redundant Development
- Fragmented Community
- Insufficient Coordination Capability

Solution:
- Natural Intervention
  - Identifying Best Practices
  - Identifying Feature
  - Identifying Redundancies
  - Identifying Lost Contribution
  - Identifying Relevant People
  - Facilitate Particular Interactions
- New Intervention

Future:
- Advances in tooling & SE principles
- Insights from other disciplines
- Mix a wide range of research methods
Improving Collaboration Efficiency

- **Problem**
  - Lack of Overview
  - Lost Contribution
  - Redundant Development
  - Fragmented Community
  - Insufficient Coordination Capability

- **Solution**
  - Natural Intervention
    - Identifying Best Practices
  - New Intervention
    - Identifying Feature
    - Identifying Redundancies
    - Identifying Lost Contribution
    - Identifying Relevant People
    - Facilitate Particular Interactions

- **Fork-Based**
  - [FSE'19]
  - [ICSE'20]
  - [ICSE'18]
  - [SANER'19]

- **Future**

+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods
Acknowledgement

It’s not yet the end ...